جامعة أبو بكر بلقايد- تلمسان

Université Abou Bakr Belkaïd de Tlemcen

Faculté de Technologie

Département de Génie Biomédical

Laboratoire de Recherche de Génie Biomédical

MEMOIRE DE PROJET DE FIN D'ETUDES

pour l'obtention du Diplôme de

MASTER en GENIE BIOMEDICAL

Spécialité : **Informatique Biomédicale**

présenté par : SOUIDI Abdelhakim

Indexation contrôlée des textes biomédicaux orientée par l'extraction de connaissances

Soutenu le 26 mai 2016 devant le Jury

Mr.	CHIKH Mohamed Amine	*Prof*	Université de Tlemcen	Président
Mr.	ABDERRAHIM Mohammed El Amine	*MCA*	Université de Tlemcen	Encadreur
Mr.	MOUSSAOUI Djilali	*MAA*	Université de Tlemcen	Examinateur

Année universitaire 2015-2016

Je dédié ce mémoire À :

Mes parents,

A la mémoire de mes grands-parents paternels,

Mes grands-parents maternels,

Mes sœurs et mes frères,

Toute ma famille,

Mes amis,

Qu'ils trouvent ici l'expression de ma sincère gratitude.

Remerciements

D'abord, j'adresse mes vifs remerciements à mon encadreur Mr Abderrahim Mohammed El Amine, maitre de conférences à l'université de Tlemcen, qui s'est montré à l'écoute et disponible tout au long de la réalisation de ce travail malgré ses charges administratives, et surtout pour sa qualité d'encadrement qui était avec beaucoup de pédagogie, de patience et de gentillesse.

Ensuite, je remercie très sincèrement Mr Chikh M.A, Professeur à l'université de Tlemcen, d'avoir accepté de présider ce jury.

Je tiens à remercier Mr Moussaoui Djilali, maitre -assistant à l'université de Tlemcen, qui a participé à l'examen de ce travail.

J'adresse mes remerciements à tous mes enseignants du département Génie Biomédical le long de mes années de formation.

Je remercie également toute l'équipe CREDOM d'avoir m'accueilli avec bienveillance au sein du laboratoire.

Résumé

MEDLINE est la base bibliographique de référence dans le domaine biomédical. Cette dernière connait une croissance explosive dans les dernières années. L'indexation de cette ample base avec le thésaurus MeSH devient de plus en plus une tâche difficile pour un staff qualifié de la US-NLM. La catégorisation de texte (CT) à base des algorithmes d'apprentissage, étudiée dans le contexte d'indexation des références de MEDLINE, est une façon efficace afin d'aider ce groupe d'expert.

Les algorithmes d'apprentissage supervisé, comme les SVM et la classification Naïve Bayes appliqués sur la représentation standard (sac de mots, ou en anglais : bag-of-words) affinée par des techniques de prétraitement montre des performances compétitives : une F-Mesure de 59.6% pour le classifieur NB, et F-Mesure de 58% pour les SVM avec des paramètres standards. Cependant, la représentation d'un texte peut engendrer un espace de dimension impressionnant entravant les classifieurs.

La sélection de descripteurs est une tâche populaire qui consiste à retrouver les mots représentatifs et éliminent ceux qui ne le sont pas. Nous avons appliqué la méthode de sélection chi-2 (chi-squared) avec les techniques SVM et NB. Cette méthode accomplis des bonnes performances, une F-Mesure de : 62.7% pour les SVM et 65.9% pour le classifieur NB.

Mots-clés: MEDLINE, chi-squared, SVM, Naïve Bayes, MeSH thesaurus, catégorisation de texte, Algorithmes d'apprentissage supervisé, prétraitement, sélection de descripteurs, sac de mots, Indexation.

Abstract

MEDLINE is the well-known database of bibliographic citations in biomedicine. It knows an explosive growth in recent years. The indexing of this large database using the MeSH thesaurus is performed by a relatively small group of highly qualified indexing staff at the US-NLM. However, their task is becoming difficult due to the ever increasing size of MEDLINE. Text Categorization based Machine learning – which is studied in the context of indexing MEDLINE citations – is an efficient way that could help this group of experts.

Supervised Machine Learning Algorithms, such as: SVM and the Naïve Bayes classifier applied on a refined bag-of-words representation with preprocessing methods shows competitive performances: an F-Measure of 59.6% for a NB classifier and an F-Measure of 58% for SVM with standards parameters. However, text representation may produce a large number of features, which can hinder the performance of learning algorithms.

Features Selection is a popular task to find representative words and remove unimportant ones. We applied one of those approaches called chi-squared with the SVM and the NB. We achieved better performance: an F-Measure of 62.7% for SVM and 65.9% for NB.

Keywords: MEDLINE, MeSH thesaurus, chi-squared, SVM, Naïve Bayes, Text Categorization, Machine Learning, Features Selection, preprocessing, bag-of-words, Indexing.

ملخص

تعتبر MEDLINE قاعدة البيانات المرجعية للنصوص المتعلقة بالمجال البيوطبي. هاته القاعدة – التي هي بمثابة مكتبة إلكترونية يتم الوصول إليها عن طريق شبكة الأنترنت – تقوم بتخزين النصوص على وسائط الإعلام الإلكترونية وتعرف مؤخرا إرتفاعا هائلا في عدد النصوص المضافة لها. يتكفل مجموعة من الخبراء بتصنيف النصوص فيها عن طريق استخدام تجميعية تحتوي عددا من التعبيرات القياسية المتفق عليها تعرف بالـ MeSH. أصبحت عملية التصنيف للخبراء صعبة نظرا لعدد النصوص المرتفعة، فكان اللجوء إلى تكنولجيا الذكاء الإصطناعي أمر ضروري لتسهيل التصنيف و تقديم الدعم لهؤلاء الخبراء.

خوارزميات الذكاء الإصطناعي مثل: SVM و Naïve Bayes يمكن استخدامهما لتصنيف النصوص بعد تغيير هيئة النص إلى مجموعة من الكلمات الأكثر تمثيلا له و هذا بعد تصفيته. قدمت هاته الطريقة نتائج لا بأس بها. إنّ من المشاكل التي تعترض تمثيل النص، هو العدد المرتفع للخاصيات الممثلة للنص. لذا أصبح من الضروري القيام بعملية اختيار للخاصيات التمثيلية.

اختيار الخاصيات التمثيلية هي مهمة ايجاد الخاصيات الأكثر تمثيلا للنص و الاحتفاظ بها. استخدمنا طريقة من الطرق في اختيار الخاصيات تعرف بـ Chi-2 و قد ساعدت في تحسين النتائج المحصل عليها بعد التقييم.

الكلمات المفتاحية: التصنيف، MEDLINE، Naïve Bayes، اختيار الخاصيات التمثيلية، الذكاء الإصطناعي، SVM، MeSH.

Table de matières

Tableaux et Figures

Liste des tableaux

Liste des figures

Listes d’abréviations

ARFF : Attribute Relation File Format

CT : Catégorisation automatique de Texte

DLN : Deep Learning Networks

HuGE : Human Genome Epidemiology

ICD-9 : International Classification Disease 9

ID : Ingénierie de Descripteurs

IDF : Inverted Document Frequency

JSON : JavaScript Object Notation

K-ppv: K-plus proche voisin

K-NN : K- Near Neighbor

MEDLINE: Medical Literature Analysis and Retrieval System Online

MeSH : Medical Subject Headings

MTI: Medical Text Indexer

NB: Naïve Bayes

NLM: National Library of Medicine

PMC: Perceptron Multi-Couches

PRC: PubMed Related Citation

RI : Recherche d’Informations

RNA: Réseau Neuronal Artificiel

SVM: Séparateurs à Vastes Marges

TALN: Traitement Automatique du Langage Naturel

TF: Term Frequency

UMLS : Unified Medical Langage System

Introduction générale

La quantité d'informations sous format électronique ne cesse d'augmenter que ce soit sur l'Internet ou sur l'Intranet des entreprises. Le domaine biomédical n'est pas l'exception, le nombre de publications scientifiques accroit d'une manière exponentielle. Par exemple, la base bibliographique MEDLINE®[1] (Medical Literature Analysis and Retrieval System Online) développé et gérée par l'U.S National Library of Medicine (NLM) contient plus de 23 millions articles avec 800 000 nouveaux ajoutés chaque année. Ce qui exige aux dirigeants de ces entreprises de trouver des moyens et des mécanismes afin d'organiser cette masse d'informations et la rendre facilement accessible pour pouvoir satisfaire les besoins des utilisateurs utilisant leurs systèmes.

Une discipline nommée la Recherche d'Informations (RI) s'intéresse au développement des Systèmes de Recherche d'Informations (SRI) permettant de faciliter l'accès et retrouver une information pertinente qui répond au besoin d'un utilisateur à partir d'une collection volumineuse de documents. Les SRI se basent sur un processus d'indexation dont le but de choisir les termes d'indexation qui représentent au mieux le contenu de document. Cependant, la sélection des termes d'indexation est une tâche difficile qui pose des problèmes majeures dans la RI. (Ruiz E. M., Srinivasan P., 1998) ont divisé les approches de sélection des termes d'indexation en deux grandes approches : ceux qui peuvent être choisi à partir d'un vocabulaire prédéfini (dit aussi : vocabulaire contrôlé) ou à partir des séquences de mots quelconques dans le texte (non contrôlé ou indexation libre).

La catégorisation de texte, définie comme étant la tâche d'assigner automatiquement des étiquettes (catégorie ou classe) à un texte libre est une forme d'indexation. Dans le domaine biomédical, cette tâche est étudiée avec la tâche d'indexation des références de MEDLINE en utilisant comme termes d'indexation, ceux qui sont trouvés dans le vocabulaire contrôlé. Les articles de MEDLINE® sont indexés en utilisant les descripteurs du vocabulaire contrôlé Medical Subject Headings (MeSH) développé par le NLM qui contient plus de 24 000 descripteurs. Auparavant, cette indexation est assurée manuellement par un groupe d'experts qualifiés de la NLM. Cependant, leur tâche est devenue coûteuse et longue à cause de l'enrichissement que connaisse le MEDLINE® (environ 800 000 articles ajoutés chaque année) ce qui rend la mise à jour de cette base extrêmement difficile et fastidieux. Par conséquent, l'automatisation de ce processus est devenue un enjeu pour la communauté scientifique en faisant appel aux techniques issues du traitement automatique du langage naturel (TALN) et les techniques d'Apprentissage Artificiel (en anglais : Machine Learning).

Notre travail de mémoire se situe dans le cadre de catégorisation de textes biomédicaux en utilisant les techniques d'apprentissage automatique.

[1] Accessible via PubMed à l'URL, http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed

Ce mémoire est organisé en 4 chapitres, chacun d'entre eux comporte une introduction et une conclusion. Un premier chapitre intitulé : « L'indexation de documents », présente le contexte de notre travail. Les deux chapitres suivants chapitres II et III concernent principalement l'état de l'art des approches de représentation du texte et les méthodes d'apprentissage automatique appliquées dans la catégorisation automatique de texte. Le dernier chapitre est consacré à l'évaluation expérimentale. Le détail de cette organisation est donné comme suit :

- Le chapitre I, Indexation des documents, présente la notion d'indexation dans un contexte générale de la RI (*Section 2*) ainsi que ses deux types (voir langage) (*Section 2.1*). Nous présentons en particulier dans (*Section 3*) l'indexation des références de MEDLINE en illustrant le lien entre la catégorisation automatique de texte et l'indexation.
- Le chapitre II, Catégorisation automatique de textes, donne un aperçu sur la catégorisation automatique de texte. Nous commençons d'abord par définir la catégorisation automatique de texte (*Section 2*). Ensuite, nous décrivons dans la (*Section 3*) le processus de catégorisation de texte. Puis, nous abordons une des phases cruciale du processus de la CT qui est l'ingénierie des descripteurs (*Section 4*) dans laquelle nous présentons l'état de l'art des différentes approches utilisés dans la représentation de texte pour la classification de documents biomédicaux (*Section 4.2*).
- Le chapitre III, Algorithmes d'Apprentissage pour la catégorisation de texte, présente les algorithmes d'apprentissages automatique appliqués dans la catégorisation de textes (*Section 2*) et les métriques d'évaluation (*Section 2.3*), nous nous focalisons sur les différentes méthodes d'apprentissage supervisé, ainsi que leur utilisation dans la catégorisation des textes biomédicaux (*Section 2.2*).
- La chapitre IV, évaluations expérimentales, présente notre contribution dans le cadre de la catégorisation de textes biomédicaux issus de la bibliothèque MEDLINE. Nous présentons d'abord la collection utilisée dans la compagne d'évaluation BioASQ. Ensuite, nous décrivons le plan des expérimentations, ainsi que l'environnement logiciel et matériel utilisé. Puis, les méthodologies permettant de passer d'un texte à une représentation vectorielle traitable par les méthodes d'apprentissage. Les résultats expérimentaux obtenus sont présentés et discutés dans (*Section 4*) en utilisant un sous-ensemble de références de la collection utilisée dans le BioASQ.

Chapitre 1 : Indexation de documents

1 Introduction

La recherche d'information (RI) est un domaine historiquement lié aux sciences de l'information qui ont pour objectif d'établir les représentations des documents ainsi que des requêtes de l'utilisateur dans le but d'en récupérer des informations (son, images, textes, etc…), à travers la construction d'index (Ba-Duy D. 2012). Le processus général de la RI comporte deux processus : Indexation des documents et l'appariement requête-document. On s'intéresse au premier processus.

L'indexation est une composante cruciale de tout système de recherche d'informations. De sa qualité dépend en partie la qualité des réponses du système (AZZOUG W. 2013). Cette dernière pourra être réalisée d'une manière manuelle ou automatique dans deux langages (voire types) différents.

Dans ce chapitre, nous donnons une définition à la notion d'indexation dans un contexte général de la RI (*Section 2*). Ensuite, nous montrons ses deux types (langages d'indexation) (*Section 2.1*), puis notre intérêt dans la (*Section 3*) s'est porté sur la démarche d'indexation des références biomédicales de la MEDLINE.

2 Indexation de documents

Avec la croissance constante de documents en termes de formats et de contenus, les méthodes pour traiter ces documents deviennent également divers et sophistiqués. Et alors, la notion d'indexation devient omniprésente, elle a une existence dans différents domaines (Névéol A. *et al.*, 2009).

On ne s'intéresse dans ce mémoire qu'aux documents textuels. Où l'information disséminée dans ces documents n'est pas structurée et donc difficilement accessible, la raison pour laquelle, il est nécessaire d'effectuer une indexation pour transférer l'information contenue dans le texte vers un autre espace de représentation traitable par un système informatique (Roussey C. 2001).

Une définition donnée par AZZOUG (AZZOUG W. 2013) : « *L'indexation consiste à analyser les documents et les requêtes dans le but d'en définir un ensemble de descripteurs (termes d'index) permettant d'exploiter plus facilement leur contenu lors du processus de la recherche*. »

D'une manière générale, l'indexation peut être considérée comme un processus de représentation de texte en associant des descripteurs à un document, qui a pour but d'identifier l'information contenue dans tout texte et de la représenter au moyen d'un ensemble d'entités appelé index (Roussey C. 2001).

2.1 Types d'indexation (langages d'indexation)

(Ruiz E. M., Srinivasan P., 1998), (Ba-Duy .D., 2012), (Névéol A. Claveau V., 2008), (Roussey C., 2001) ont distingué deux types d'indexation : indexation libre vs. Indexation contrôlée.

1. **Indexation Libre (langage libre):**

L'indexeur (manuel ou automatique) extrait les mots clés d'un document ou les choisit librement. Le vocabulaire de l'indexation libre n'est pas limité par un contrôle ni régi par une syntaxe car aucune contrainte n'est pas imposée *a priori*, il est donc composé de tous les descripteurs considérés représentatifs du contenu de document. Par exemple, l'indexation plein texte est une indexation libre car tous les mots du document sont extraits automatiquement pour constituer l'index du document (Roussey C., 2001). Cependant, le vocabulaire évolue rapidement et peut contenir des termes synonymes, polysémiques, etc. ce qui entraine des incohérences et diminue les performances du système de recherche d'informations (Roussey C. 2001).

2. **Indexation contrôlé (langage contrôlé) :**

Dans l'indexation contrôlée, le vocabulaire ou le langage d'indexation est définie *a priori* et son utilisation exclusive s'impose à l'indexeur. Ce vocabulaire est un vocabulaire normalisé, c'est-à-dire que pour éviter les problèmes de polysémie et de synonymie du langage libre, une liste de termes d'indexation est définie. Cette liste appelé aussi la liste d'autorité est une liste structurée de descripteurs relié entre eux par des relations de hiérarchie, d'association ou d'équivalence (Ba-Duy .D., 2012).

3 Indexation des documents MEDLINE

3.1. Base bibliographique MEDLINE

MEDLINE est la base bibliographique de référence dans le domaine biomédical, elle contient plus de 23 millions articles qui concernent principalement la littérature biomédicale écrite majoritairement en Anglais (des références à des revues scientifiques et des comptes rendus des conférences du milieu biomédical) (Jimeno-Yepes A., et al., 2014). Cette base a été créée dans les années 1960 et gérée par la NLM aux Etats Unis depuis 1966 jusqu'à aujourd'hui. Elle couvre tous les domaines biomédicaux : biochimie, biologie, médecine clinique, pharmacologie, les soins infirmiers, les sciences paramédicales, etc…

Le portail PubMed dispose d'un moteur de recherche permettant l'accès à l'information dans MEDLINE, à des résumés d'articles et des articles en texte intégral sur les

sciences de la vie et biomédicales (Ba-Duy .D., 2012). La figure 1 montre un exemple d'une citation provenant de la base bibliographique MEDLINE :

Display Settings: Abstract Send to: Ann Intern Med FULL TEXT

Ann Intern Med. 2013 Jul 2;159(1):1-12. doi: 10.7326/0003-4819-159-1-201307020-00003.

Salicylate (salsalate) in patients with type 2 diabetes: a randomized trial.

Goldfine AB[1], Fonseca V, Jablonski KA, Chen YD, Tipton L, Staten MA, Shoelson SE; Targeting Inflammation Using Salsalate in Type 2 Diabetes Study Team.

Collaborators (27)

Shoelson SE, Goldfine AB, Fonseca V, Jablonski K, Staten M, Jablonski K, Pyle L, Aroda V, Barzilay J, Buse J, Crandall J, Desouza C, Donovan D, Dulin M, Fonseca V, Goldfine AB, Henry R, Hershon K, Lorber D, Mather K, Ovalle F, Piziak V, Pop-Busui R, Raskin P, Rudo A, Umpierrez G, Warren W.

Author information

[1]Joslin Diabetes Center and Harvard Medical School, Boston, Massachusetts 02215, USA. allison.goldfine@joslin.harvard.edu

Abstract

BACKGROUND: Short-duration studies show that salsalate improves glycemia in type 2 diabetes mellitus (T2DM).

OBJECTIVE: To assess 1-year efficacy and safety of salsalate in T2DM.

DESIGN: Placebo-controlled, parallel trial; computerized randomization and centralized allocation, with patients, providers, and researchers blinded to assignment. (ClinicalTrials.gov: NCT00799643).

SETTING: 3 private practices and 18 academic centers in the United States.

PATIENTS: Persons aged 18 to 75 years with fasting glucose levels of 12.5 mmol/L or less (≤225 mg/dL) and hemoglobin A1c (HbA1c) levels of 7.0% to 9.5% who were treated for diabetes.

INTERVENTION: 286 participants were randomly assigned (between January 2009 and July 2011) to 48 weeks of placebo (n = 140) or salsalate, 3.5 g/d (n = 146), in addition to current therapies, and 283 participants were analyzed (placebo, n = 137; salsalate, n = 146).

MEASUREMENTS: Change in hemoglobin A1c level (primary outcome) and safety and efficacy measures.

Figure 1. Exemple d'une référence MEDLINE.

MEDLINE reçoit environ 800 000 nouvelles citations chaque année et un groupe d'experts de la NLM se charge de les annoter manuellement ou semi automatiquement par les termes du vocabulaire contrôlé MeSH (Medical Subject Headings). Dans la section suivante, nous présentons la démarche utilisée par ce staff pour entretenir la mise à jour de cette base.

3.2. Indexation dirigée par un vocabulaire contrôlé

3.2.1. Le thésaurus MeSH

Pour aider et faciliter l'indexation, la recherche et le classement des documents biomédicaux, la NLM a introduit en 1954 son thésaurus MeSH (Medical Subject Headings) qui est un vocabulaire contrôlé avec 27 883 descripteurs recensés en 2016[2]. Un vocabulaire contrôlé est un lexique dont le but est de rendre possible l'organisation des connaissances afin d'optimiser la recherche d'information. Le MeSH est organisé d'une façon hiérarchique comprenant essentiellement des termes qui désignent les concepts biomédicaux, des

[2] https://www.nlm.nih.gov/pubs/factsheets/mesh.html

descripteurs, des relations et des qualificatifs. Dans un niveau d'hiérarchie général on trouve les termes génériques tel que : « Anatomie », les termes très spécifiques se trouvent dans un niveau d'hiérarchie plus bas tel que : « index », « cheville ».

Dans la NLM, chaque référence de la MEDLINE est annotée manuellement par un nombre pertinent de descripteurs MeSH énumérant les thématiques abordés dans le document. Cependant, cette tâche est devenue coûteuse et alors il devient nécessaire de faire appel aux outils semi-automatique et automatique.

3.2.2. L'outil Medical Text Indexer (MTI)

L'outil Medical Text Indexer (MTI) (Aronson A.R., *et al.,* 2004) est utilisé depuis 2002 pour assister les experts de la NLM à indexer les articles de la MEDLINE en leur donnant des recommandations des descripteurs MeSH pour chaque citation. Le programme MTI a deux composantes principales : MetaMap et l'algorithme PubMed Related Citation (PRC). MetaMap effectue une analyse de la référence et lui associe des concepts issus de l'Unified Medical Language System (UMLS). Puis, le mapping à partir l'UMLS vers le MeSH suit l'approche Restrict-to-MeSH basée sur les relations sémantiques entre les concepts de l'UMLS. L'algorithme PRC est un algorithme K-ppv (K plus proche voisin, en anglais : K-NN) adapté qui repose sur la similarité entre les documents pour assigner les descripteurs MeSH. Donc le MTI est un outil qui recommande et suggère des termes MeSH à un document (Jimeno-Yepes A., et al., 2011). Il recommande 25 termes en moyenne à chaque document et constitue une aide semi-automatique pour les indexeurs de la NLM.

3.3. La catégorisation de textes dans le contexte d'indexation

(Névéol A. *et al.,* 2009) ont défini l'indexation: « *as the task of assigning to a document a limited number of terms denoting concepts that are substantively discussed in the document* ». L'indexation est l'activité d'assigner un nombre limité de termes désignant des concepts à un document et qui sont éventuellement abordé dans le document.

Comme nous avons mentionné dans le paragraphe (§ : 3.2.1), l'ajout quotidien des articles à la MEDLINE (environ 800 000 articles ajoutés chaque année) a rendu difficile sa mise à jour. La catégorisation automatique de textes désigne l'activité d'assigner des étiquettes prédéfinies à un texte. Cette tâche est appliquée dans plusieurs contextes parmi lesquelles l'indexation contrôlée (Sebastiano F. 2002). En outre, l'indexation contrôlée est souvent vue comme un problème de catégorisation car l'indexeur doit décider si un concept est éventuellement discuté dans le document ou pas.

4 Conclusion

Nous nous sommes focalisé dans ce chapitre sur une des composantes importantes dans la RI, qui est : le processus d'indexation. Où, on a mis en évidence le rôle central de ses deux types, en décrivant brièvement les limites de l'indexation libre et l'intérêt de l'indexation contrôlée dans ce cadre. Notre but c'est l'indexation de la littérature du domaine biomédical, plus exactement, la base bibliographique de référence dans les sciences de vie et de la biomédecine MEDLINE, en expliquant la démarche avec laquelle la NLM assure la mise à jour de cette dernière avec l'utilisation d'un vocabulaire contrôlé (thésaurus MeSH), puis, nous avons montré que cette indexation n'est qu'un problème de catégorisation. Nous souhaitons qu'on ait mis en exergue la relation entre l'indexation contrôlée et la catégorisation automatique de textes.

Chapitre II : Catégorisation automatique de textes

1 Introduction

Les bibliothèques en ligne et la toile (World Wide Web) sont parmi les facteurs qui produisent quotidiennement d'une manière extrêmement rapide une masse énorme d'information. L'utilisateur submergé par cette masse d'informations, reste incapable de trouver l'information dont il a besoin, parmi toutes celle qui est accessible. Dans ce contexte, la catégorisation de texte devient la méthode clé pour organiser l'information afin d'assister les tâches de traitement de textes à base de contenu (Content-based Text Processing).

La catégorisation de texte est un domaine de recherche important et actif qui se situe à l'intersection des deux domaines : la recherche d'information et l'apprentissage automatique. Ces deux domaines ont permis vers les années 1990 à cette activité d'être très imploré, car elle offre une bonne voie pour étudier la représentation de texte. De plus, elle a un rôle considérable dans de nombreuses applications : comme le filtrage (filtrage des e-mails aux spams ou non-spams), l'organisation, le processus d'indexation dans la RI, ainsi, comme une composante dans les systèmes de compréhension de texte : ces systèmes s'intéressent à un accès plus complexe au contenu du document tel que: le peuplement des bases de données en se basant sur l'extraction d'information, les systèmes Questions / Réponses, etc…

De toute façon, il y a un investissement financier substantiel consacré par les grandes organisations, comme la National Library of Medicine pour indexer manuellement avec un vocabulaire contrôlé. Ces organisations veulent assurer la continuité de leurs bases bibliographique en utilisant les vocabulaires contrôlés pour indexer, et donc, le développement des outils automatiques devient nécessaire pour offrir une aide ou remplacer ce processus manuel intense.

Dans ce chapitre, nous allons dans un prime abord définir la catégorisation automatique de texte (*Section 2*). Ensuite, nous décrivons dans la (*Section 3*) la démarche de catégorisation de texte. Puis, on fait la lumière sur une des phases cruciale du processus de la CT qui est l'ingénierie des descripteurs (*Section 4*) dans laquelle nous présentons l'état de l'art des différentes approches utilisées dans la représentation de texte pour la classification de documents (*Section 4.2*). Une conclusion est en fin du chapitre.

2 Définition de la catégorisation de texte

2.1. Classification ou Catégorisation ?

L'objectif de ce paragraphe est de clarifier notre positionnement par rapport à la tâche visée dans ce mémoire.

La classification dans un sens computationnel est définie par (Ingersoll G. S. et al. 2013) comme : l'activité qui cherche à assigner des étiquettes à un texte. Plus précisément, ces étiquettes à assigner ne sont pas forcément prédéfinies. (Lipka N. 2013) voit la classification de textes comme étant une forme d'extraction de connaissances à partir d'un texte utilisé pour la recherche d'informations dans des collections. Sebastiani (Sebastiani F. 2002) a noté que le terme : « la classification automatique de texte » est utilisé dans la littérature pour désigner plusieurs significations : i) une identification automatique d'un ensemble de catégories, ou aussi ii) une identification automatique d'un ensemble de catégories et grouper des documents sous ces catégories identifiées, cette tâche est appelée aussi : Text Clustering, ou d'une manière générale iii) toute activité qui consiste à placer des textes items dans des partitions, quand l'ensemble des catégories à assigner sont *prédéfinies*, l'assignement de ces dernières s'appelle : iv) « La catégorisation automatique de texte ». Et alors, la catégorisation est une spécialisation de la classification, et c'est cette instance particulière de la classification qui nous intéresse dans ce travail.

Dans la prochaine section, nous allons se focaliser sur la tâche de catégorisation automatique de texte.

2.2. Catégorisation automatique de texte

La catégorisation de textes date des années 1960, où beaucoup de chercheurs se sont intéressés à assigner des classes à un texte d'une manière automatique (Borko H., et al., 1963), (Borko H., et al., 1964). Mais, c'est vers les années 1990 que ce champ a devenu un sous-domaine majeur dans la discipline systèmes d'informations grâce à la croissance des intérêts applicatives dans lesquelles la catégorisation de textes prend une place.

Lewis (Lewis D D. 1992) a défini la catégorisation de texte comme étant : « l'assignement de documents à une ou plusieurs ensemble de catégories préexistées ». JALAM (JALAM R., 2003) formule cette activité comme étant : l'apprentissage d'un modèle afin de rechercher une liaison fonctionnelle entre l'ensemble de textes et l'ensemble de catégories (étiquettes, classes). (Sebastiani F. 2002) ajoute la notion de classes cibles prédéfinies. (NAKACHE D. 2007) en rappelant l'historique du terme « catégorisation », il l'a défini comme étant : « *l'action d'affecter des éléments, qui possèdent des caractéristiques communes, à des catégories préétablies, sans relation d'ordre* ». Sa définition se rapproche avec celle d'Aristote : « *les espèces les plus générales de ce qui est signifié par un mot simple*», de par l'absence d'ordre ou hiérarchie.

La modélisation d'un système de catégorisation de texte peut être résumée dans la **Figure (2).**

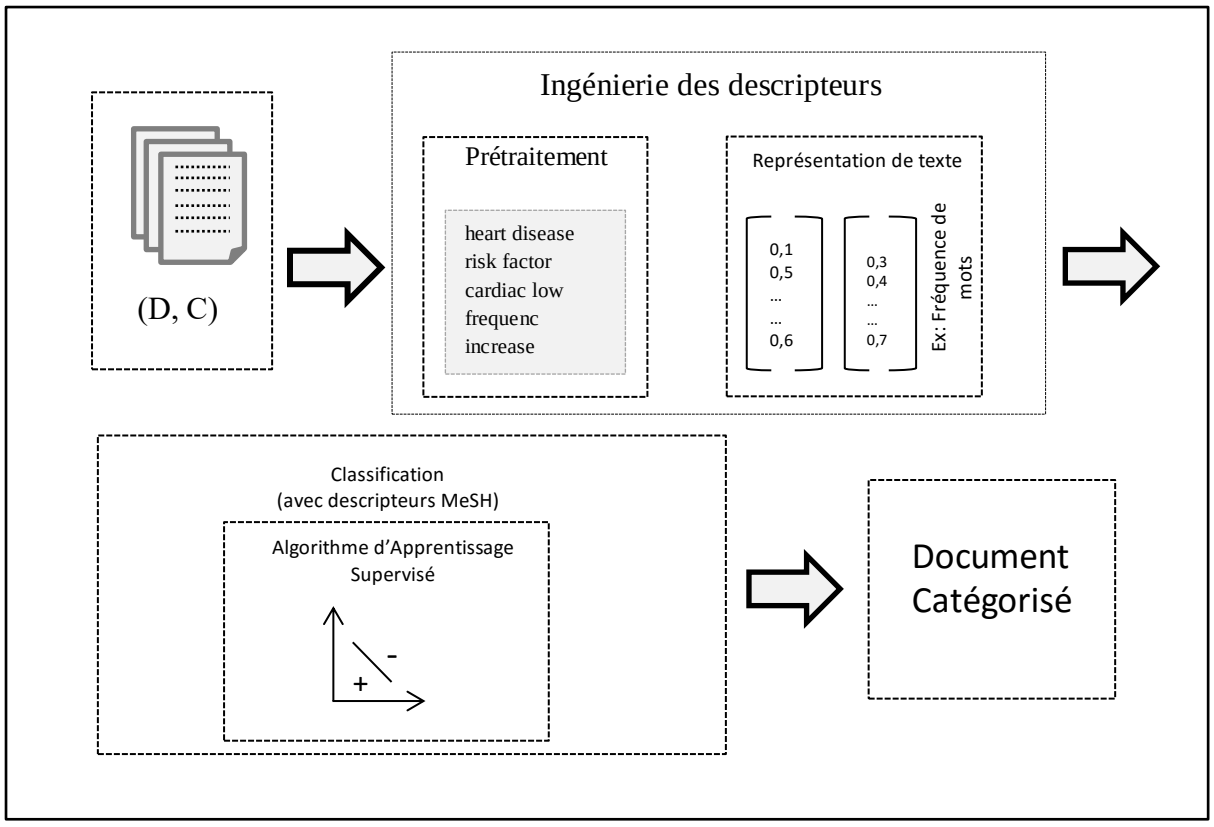


Figure 2. Blocs de développement d'un système de catégorisation automatique des textes

Nous consacrons d'autres sections du chapitre pour détailler ces phases.

2.3. Une définition mathématique à la catégorisation de texte

Nous pouvons représenter la catégorisation de texte sous un angle purement mathématique (Hirotoshi T. 2002), (Sebastiani F. 2002), (NAKACHE D. 2007), (JALAM R. 2013).

Généralement, dans la CT, un document est exprimé par un vecteur de plusieurs dimensions,

$$X = (x_1, x_2, \dots, x_n).$$

Chaque caractéristique du vecteur correspond à deux valeurs : le mot apparaissant dans le document et une valeur réelle indiquant sa fréquence, par exemple, la formule TF-IDF, ou une valeur binaire indiquant sa présence / son absence.

La tâche consiste à associer une valeur booléenne à chaque paire $(d_j, c_i) \in D \times C$ avec :

D : L'ensemble de textes ;

C : L'ensemble de catégories prédéfinies.

La valeur V (Vrai) est associée au couple (d_j, c_i) si le texte d_j appartient à la classe c_i, la valeur F (Faux) lui sera associée dans le cas contraire.

Prenant par exemple, deux documents :

«An emergent cardiovascular hazard in patients selected for low risk of heart disease. » (Document 1)

«Physical activity (PA) predicts cardiovascular mortality in the population at large. » (Document 2)

Ils sont exprimés par: x_1, x_2 dans le tableau (Tab 1). En utilisant trois mots comme des descripteurs: "cardiovascular", "patients", "Physical activity"

	Cardiovascular	Patients	Physical activity
Document1 (x_1)	1	1	0
Document2 (x_2)	1	0	1

Tab 1. Représentation vectorielle des deux documents.

D'une autre part, l'étiquette «y » est donnée, qui indique dans quelle classe le document appartient-il. On peut formaliser ce problème comme étant le calcul d'une fonction $f(x)$ de telle manière la différence entre la classe prédite et la classe réelle soit minimisée, en utilisant un ensemble d'apprentissage S. Avec : S est l'ensemble des données pour apprentissage :

$$S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

3 Processus de la catégorisation de textes

La synthèse des différents documents (par exemple, (Basili R. 2005), (NAKACHE D. 2007), (Khan A. et al. 2010), (JALAM R. 2013), (Lipka N. 2013)) permet de distinguer les phases utilisées dans le processus (voir méthodologie) de catégorisation automatique de textes. Nous distinguons trois phases essentielles dans ce processus où chacune des phases comporte un ensemble d'étapes habituellement suivies et qui sont reconnues universellement par la communauté des chercheurs scientifiques, la figure 3 illustre la démarche de catégorisation automatique de textes.

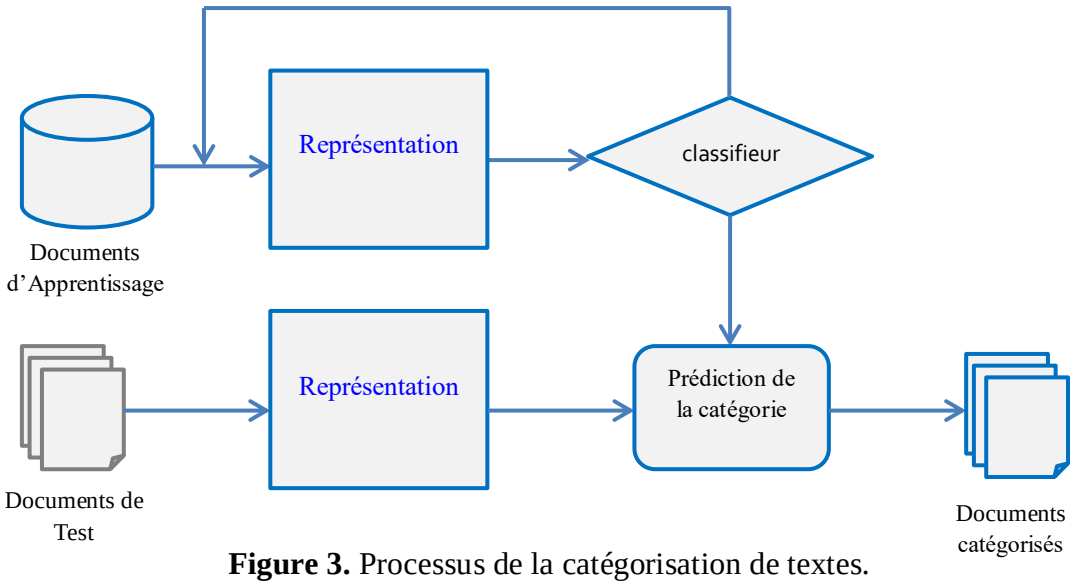


Figure 3. Processus de la catégorisation de textes.

On distingue les trois phases comme suit :

1- Phase d'Ingénierie de descripteurs

Pour assurer une catégorisation de textes basée sur les algorithmes d'apprentissage, il fallait procéder une ingénierie de descripteurs. Cette dernière peut être inclue dans la phase d'apprentissage (par exemple (NAKACHE D. 2007), (JALAM R. 2013)).

Elle comporte deux grandes étapes :

a- Prétraitement : le prétraitement des documents est appliqué afin d'extraire l'information pertinente et de simplifier le document. Parmi les techniques utilisées dans cette étape :

- *Elimination des mots vides (Stop Words Remove);*
- *Normalisation.*

b- La représentation du document: cette étape concerne la façon avec laquelle les exemples sont représentés.

Nous allons expliquer en détail cette phase cruciale, ainsi que ses étapes dans la (*Section 4*).

2- Phase d'Apprentissage

Après avoir bien préparé les données, on rentre dans la phase d'apprentissage. Cette dernière se base sur :

a- la manière dont les exemples sont choisis ;

b- le choix de l'algorithme d'apprentissage : cette étape consiste à identifier l'algorithme d'apprentissage qui donnera les meilleures performances pour la classification d'un document.

L'algorithme d'apprentissage traite chaque exemple étiqueté préalablement et identifiera comment les descripteurs se rapportent à chaque étiquette.

3- Phase de Test

Après la vectorisation des documents, le corpus sera divisé en deux échantillons. Le premier sert à identifier le meilleur classifieur ainsi que les paramètres à lui associés. Mais l'évaluation portera sur le deuxième échantillon qui n'est pas utilisé dans l'apprentissage.

4 L'ingénierie des descripteurs

L'Ingénierie des Descripteurs ID (en anglais, Features Engineering) est une phase primordiale dans une opération de catégorisation de textes à base des Algorithmes d'Apprentissage Artificiel. Cette phase a comme finalité la construction des descripteurs d'une manière à les rendre parfaitement exploités par les techniques d'apprentissage automatique. Les descripteurs représentent des mots clés significatifs décrivant le contenu informatif des documents (AZZOUG W. 2013). Pour ce faire, des étapes telles : le prétraitement, la représentation de documents et la réduction des dimensions sont indispensables.

Dans ce qui suit, nous allons exposer l'ensemble de ces stratégies.

4.1. Prétraitement

Le prétraitement de texte est une étape incontournable dans un processus de catégorisation de textes. Il est appliqué afin d'éliminer tout bruit existant dans la donnée textuelle et garder que l'information pertinente, ce qui va améliorer la précision d'un système de RI (Bassil Y. 2012). Cette étape comporte une analyse lexicale (Tokenisation/Segmentation) (*Section 4.1.1*). Puis, une élimination des mots fonctionnels qui ne sont pas nécessaires (*Section 4.1.2*) et une dernière étape de normalisation des mots restants (*Section 4.2.3*).

4.1.1. Analyse Lexicale (Tokenisation/Segmentation)

D'une manière générale, le texte est constitué de plusieurs paragraphes. Chaque paragraphe est un ensemble d'une ou plusieurs phrases. Chaque phrase est une succession de tokens ou mots. Pour que des traitements ultérieurs comme la recherche d'information ou la fouille de texte (Text Mining) puissent s'appliquer, il faut tronquer ce texte en plusieurs unités lexicales (mots simples ou composés).

4.1.2. Elimination des mots vides

Au cours de cette étape, les mots non porteurs du sens, ou mots vides (pronoms personnels, prépositions, conjonctions, etc…) sont éliminés et que les mots importants et nécessaires sont gardés. L'élimination des mots vides peut se faire de deux manières : i) Soit, en utilisant une liste prédéfinies (dite *stoplist* ou *anti-dictionnaire*). Voici un exemple d'une liste par défaut de mots vides de la langue anglaise proposée par (Reese M. R. 2015) :

"i", "a", "about", "an", "are", "as", "at", "be", "by", "com", "for", "from", "how", "in", "is", "it", "of", "on", "or", "that", "the", "this", "to", "was", "what", "when", where", "who", "will", "with"

ii) ou en excluant les mots trop fréquents ou trop rares dans la collection.

4.1.3. Normalisation

La normalisation est le processus de regrouper les différentes formes morphologiques d'un mot sous une forme canonique. L'objectif étant de ramener les variantes lexicales à une forme de base unique ce qui permet de réduire l'espace de dimension de descripteurs, prenant par exemple les mots : « Hodgkin Disease », « HODGKIN DISEASE », « Hodgkin's Disease », « Hodgkin's disease », « Disease, Hodgkin » sont considérés comme des descripteurs différents, alors que leurs forme normalisée est: « *disease hodgkin* »

Deux processus :

- *La racinisation* (*désuffixation*, ou *stemming* en anglais) : c'est le processus qui vise à réduire un mot à sa racine (en anglais, stem). Ils existent plusieurs algorithmes de désuffixation, par exemple, Porter (Porter M F. 1980), Lovins (Lovins J. 1968) et Paice (Paice C. 1996).
- *La lemmatisation* : est une analyse lexicale permettant de regrouper les mots de la même catégorie grammaticale et les transformer à leur forme canonique appelée *lemme*.

4.2. Représentation du texte

La représentation du texte est un autre aspect important dans la classification des documents comprenant la transformation d'un document vers une forme qui le résume et préserve la globalité de son contenu. Cette partie a un effet direct et majeur sur les résultats de la classification, d'ailleurs, beaucoup de travaux se sont intéressés à l'évaluation des différentes approches de représentation de textes pour la classification des documents : par exemple, hors le domaine biomédical (Lewis D D. 1992), (Scott S. Matwin., 1999), dans le domaine biomédical (Yetisgen-Yildiz M., Pratt W. 2005), (Jimeno-Yepes et al. 2015).

4.2.1. Choix de descripteurs

4.2.1.1. Représentation sac de mots (bag-of-words)

Le modèle de représentation de caractéristiques standard, le plus connu et le plus utilisé dans les travaux de classification de textes est : la représentation sac de mots (en anglais : *bag-of-words*) (Scott S. Matwin., 1999), il s'agit de transformer un texte en un vecteur caractéristique où chaque position dans ce dernier correspond à un mot et accumule une valeur binaire indiquant sa présence ou une valeur numérique indiquant sa fréquence ou même son TF-IDF (Jimeno-Yepes et al. 2015), (JALAM R. 2003).

Dans le contexte de catégorisation des références MEDLINE, beaucoup de travaux ont utilisé la représentation sac de mots, par exemple, (Ruiz M. E., Srinivasan P., 1998), (Yetisgen-Yildiz M., Pratt W. 2005), (Jimeno Yepes et al., 2014), (Jimeno Yepes et al., 2015).

Dans lesquels on constate que cette technique donne des performances compétitives par rapport aux autres représentations quand il s'agit d'un corpus de taille importante. Néanmoins, cette représentation pose quelques problèmes, car elle exclut toute information relative à la grammaire et la syntaxe : l'ordre et la position des mots sont perturbés, la structure syntaxique est brisée.

4.2.1.2.Représentation de textes par phrases

L'objectif d'utiliser cette approche est de préserver l'information relative à la position et l'ordre du mot délaissée dans la représentation sac de mots et ceci en utilisant les phrases comme descripteurs au lieu de mots. Cette technique apparaît prometteuse, mais les expérimentations déroulées ont démontré que cette approche ne donne pas des améliorations par rapport à la représentation sac de mots (Scott S. Matwin S., 1999).

Lewis (Lewis D D. 1992) a consacré une partie importante de sa thèse Ph.D pour expliquer pourquoi la représentation de textes par phrases n'a pas montré de bonnes performances sur la classification de documents dans la littérature, en décrivant comment la représentation par phrases est formée: i) d'une part, cette approche est constituée par un nombre important de termes (s'il y a n mots, il y a potentiellement n^k séquences de longueur k.) ce qui engendre une augmentation de descripteurs et l'espace de dimension, ii) contient des descripteurs redondants et un niveau élevé du bruit, iii) . (Scott S. Matwin S., 1999) ont suggéré comme solution une sélection de descripteurs pour améliorer cette représentation, cette stratégie a pour objectif de réduire l'espace de dimension géant et laisser que les séquences de mots utiles. Ils ont obtenu avec les syntagmes nominaux un taux de micro-average 0.827 sur le corpus Reuters vis-à-vis 0.821 avec sac de mots.

4.2.1.3.Représentation hybrides

Nous avons expliqué précédemment que dans la représentation par phrases les qualités statistiques sont largement dégradées, et dans la représentation sac de mots les qualités sémantiques sont négligées. Donc, pour avoir les avantages de la sémantique et de la statistique, une représentation hybride entre la représentation sac de mots et la représentation par phrases parait prometteuse. Cette approche est utilisée par (Yetisgen-Yildiz M., Pratt W. 2005) dans la classification des documents MEDLINE provenant du corpus OHSUMED, elles ont eu une F-Mesure de : 0.60 vis-à-vis une F-Mesure de sac de mots et par phrases de : 0.58 et 0.57 respectivement. (Jimeno Yepes et al. 2015) a conclu que l'amélioration de la représentation sac de mots passe par combiner les deux représentations précédentes.

4.2.1.4.Représentation par des racines lexicales et lemmes

Il est possible d'utiliser comme descripteurs la forme normalisé d'un mot. Cette substitution de mots par leurs racines ou leurs lemmes réduit l'espace de descripteurs et permet de représenter par un même descripteur des mots qui ont le même sens.

4.2.2. Sélection de descripteurs

L'utilisation d'un nombre impressionnant de descripteurs peut produire un sur-apprentissage et prend un temps de calcul massif. La raison pour laquelle, plusieurs méthodes de sélection ont été proposées afin d'éviter ces problèmes et déterminer les descripteurs les plus appropriés (qui peuvent discriminer entre les documents). La sélection de descripteurs consiste à choisir un sous ensemble de descripteurs d à partir de l'ensemble original des descripteurs D, avec $d < D$. Cependant, la sélection peut entraîner une perte d'information, il faut trouver un bon compromis entre, d'une part, réduire l'espace des descripteurs, et d'autre part, garder suffisamment d'information (Dahmani H. 2012).

Il existe principalement deux types de méthodes pour la sélection des descripteurs :

La première consiste à calculer un score pour chaque descripteur, indépendamment des autres, en s'appuyant sur des mesures statistiques de l'apparition ou l'absence du descripteur en fonction de la classe à laquelle appartiennent les textes. Ensuite, les descripteurs seront classés selon ce score, les descripteurs les plus discriminants seront en tête de liste. L'information mutuelle et la méthode de chi-2 sont parmi les méthodes qui reposent sur ce principe. L'idée derrière l'information mutuelle est mesurer la quantité d'information apportée par la présence ou l'absence d'un descripteur dans un document (Taira H. 2002), tandis que la statistique du χ^2 (chi-2) mesure l'indépendance entre un descripteur t et un thème T.

La deuxième approche est constructive : elle construit itérativement un modèle, en partant d'un ensemble vide et en ajoutant successivement de nouveaux descripteurs en tenant compte des descripteurs déjà sélectionnés. Cette construction est faite en utilisant l'algorithme d'orthogonalisation de Gram-Schmidt. Cet algorithme est issu des méthodes utilisées pour trouver la solution des moindres carrés d'un problème linéaire par rapport à ses paramètres. On fait référence à (Arfken G. 1985) qui donne une description détaillée de l'algorithme Gram-Schmidt.

4.2.3. Méthodes de pondération

Habituellement, les descripteurs assument différents rôles dans différents documents. Ils peuvent être mieux représentatifs dans l'un et peu représentatif dans un autre. La valorisation des descripteurs consiste à mettre une stratégie afin de les pondérer.

La pondération des termes selon (AZZOUG W. 2013) consiste à:« *mesurer l'importance d'un terme t_j dans un document d_i en lui affectant un poids w_{ij} qui exprime son degré de représentativité*.»

Il existe plusieurs formules de pondération, exposées ci-après :

4.2.3.1. Term Frequency (TF)

Intuitivement, les mots ou les termes qui occurrent le plus dans un document particulier ont une chance d'indiquer que le mot est important dans ce document.

La fréquence du terme (ou, Term Frequency en anglais) est le nombre d'occurrences de ce terme dans le document considéré.

$$TF(t_i, d_j) = \frac{freq_{i\ j}}{max_K\ freq_{Kj}}$$

Avec: $freq_{i\ j}$: Nombre d'occurrences du terme t_i dans le document d_j ;

$max_K\ freq_{Kj}$: Fréquence maximale des termes dans d_j .

4.2.3.2. Inverse Document Frequency (IDF)

IDF est une mesure de l'importance du terme dans l'ensemble du corpus. On peut connaitre si le terme est discriminant(ou non uniformément distribué). La formule pour calculer l'IDF est donnée comme suit :

$$IDF_i = log\ \frac{d}{df_i}$$

Avec,

d : Nombre de document de la collection;

df_i : Nombre de document contenant le terme t_i.

4.2.3.3. TF – IDF

La combinaison entre le « Term Frequency (TF) » et « Inverted Document Frequency (IDF) » est communément connu sous la notation : $TF - IDF$, proposée par (Salton G. and Buckley C. 1988)

$$TF_{i,j}IDF_i = TF_{i,j} \times IDF_i$$

5 Conclusion

Ce chapitre a permis d'introduire la catégorisation de textes comme une activité particulière de la classification de textes. Des collections larges d'apprentissage et de test sont disponibles, et la nature de l'activité constitue un avantage pour étudier les techniques d'apprentissage et la représentation de texte.

Ainsi, nous avons présenté l'état de l'art des différentes approches de représentation de texte, particulièrement, leurs avantages et leurs inconvénients.

D'une manière générale, la représentation de textes produit un nombre impressionnant de descripteurs qui engendre un problème pour les algorithmes d'apprentissage, les méthodes de sélection sont nécessaires pour pallier les problèmes de dimension et déterminer les descripteurs les plus appropriés et les plus pertinents pour accomplir cette tâche. Vers la fin nous avons donné un aperçu sur les méthodes de pondérations qui mesurent l'importance des termes dans les documents.

Chapitre III : Algorithmes d'Apprentissage pour la catégorisation de textes

1 Introduction

Dans les années 1980 l'approche la plus populaire pour la conception des systèmes de classification automatique des textes est celle à base de connaissances, où des chercheurs tentent à modéliser des règles utilisées par un spécialiste humain et les incorporent dans ces systèmes, par exemple, (Apte C., et al. 1994), (Hamill K. Zamora A. 1980), (Humphrey S. Miller N. 1987).

Récemment, l'approche la plus dominante au sein de la communauté scientifique est celle qui se base sur les algorithmes d'apprentissage artificiel. L'avantage de cette dernière par rapport à la première approche réside dans son efficacité, aussi son adaptation facile aux différents domaines.

Dans ce chapitre notre intérêt s'oriente vers les approches de classification de textes à base des algorithmes d'apprentissage automatique. Dans la (*Section 2*), on définit l'apprentissage automatique, ainsi que ses types. Ensuite, nous exposons les différentes méthodes d'apprentissage supervisé (*Section 2.2*). Puis, nous montrons comment évaluer la qualité des classifieurs suivant des métriques d'évaluations (*Section 2.3*).

2 Apprentissage automatique

Une définition générale à l'apprentissage automatique donnée par Mitchell (Mitchell T. 1997) : "*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*"

Un programme est dit en Apprentissage à partir d'une expérience E dans une tâche T, si sa performance mesurée par P s'améliore automatiquement avec l'expérience E. Ce programme s'améliore automatiquement en apprenant les propriétés des catégories à partir d'un ensemble d'apprentissage.

Les algorithmes d'apprentissage peuvent se catégoriser selon le type d'apprentissage qu'ils emploient :

2.1. Types d'apprentissage

- *Apprentissage supervisé :* l'objectif de l'apprentissage supervisé est principalement de définir des règles permettant de classer des objets dans des classes en employant les caractéristiques de l'objet. On dispose au départ d'un échantillon dit d'apprentissage dont **le classement est connu**. Cet échantillon est

utilisé pour l'apprentissage des règles de classement. Ensuite, pour étudier la fiabilité des règles, on utilise souvent un deuxième échantillon indépendant, dit de validation ou de test.

- *Apprentissage non supervisé :* à partir de données $(X_i)_i$ les algorithmes d'apprentissage non supervisé tentent à trouver des structures (des partitions) dans ces données (par exemple : des classes), des densités, etc…
- *Apprentissage par renforcement :* L'Apprentissage par Renforcement (AR) consiste à apprendre quoi faire, comment associer des actions à des situations, afin de maximiser quantitativement une récompense. On ne dit pas à l'apprenant quelle action faire, mais au lieu de cela, il doit découvrir quelles actions donnent le plus de récompense en les essayant. Dans le cas le plus intéressant, des actions peuvent affecter non seulement les récompenses immédiates mais aussi la situation suivante, et par là, les récompenses à plus long terme. Ces deux propriétés – *recherche par essai-erreur* et *récompense à long terme* – sont les deux caractéristiques les plus importantes de l'apprentissage par renforcement.

Nous nous focalisons seulement sur les techniques d'apprentissage supervisé.

2.2. Algorithmes d'apprentissage supervisé

2.2.1. Les séparateurs à vaste marge (SVM)[3] :

2.2.1.1. Principe :

Les séparateurs à vastes marges ou, en anglais : « Support Vector Machine » est une technique de classification développée originellement par Vapnik dans les années (1990 – 1995).

L'objectif des SVM est de résoudre les problèmes de discrimination à deux classes. On appelle un problème de discrimination à deux classes, un problème dans lequel on tente de déterminer la classe à laquelle appartient un individu. Pour ce faire, on utilise les caractéristiques connues de cet individu. Ces n caractéristiques sont représentées par un vecteur $X \in \mathbb{R}^n$. La classe u_i à laquelle appartient l'individu est représentée par $u_i \in \{1, -1\}$ constituant l'ensemble d'apprentissage S, où : $S = \{(x_1, u_1), \dots, (x_n, u_n)\}$. Cet ensemble d'apprentissage est utilisé afin de construire une règle qui permet d'effectuer une bonne classification, et alors, la règle trouvée doit être la plus générale possible pour qu'elle soit appliquée sur de nouvelles données qui n'étaient pas dans l'ensemble d'apprentissage. En réalité, plusieurs méthodes ont été suggérées pour étendre l'application des SVM aux problèmes de discrimination à plus de deux classes.

Nous présentons dans ce qui suit comment les SVM font pour trouver cette règle.

[3] Nous préférons utiliser ce terme proposé par (Cornuéjols A. et al., 2003) par rapport à la traduction littérale du terme Support Vector Machine : « Machines à vastes marges».

a- Données linéairement séparables

Supposons donnée l'échantillon d'apprentissage $S = \{(x_1, u_1), \dots, (x_n, u_n)\}$, avec $x_i \in \mathbb{R}^n$, $u_i \in \{1, -1\}$ dont nous voulons se servir pour trouver une règle permettant séparer les données. Supposons aussi que ces données sont linéairement séparables, c'est-à-dire qu'il existe un hyperplan permettant de distinguer les exemples étiquetés par : {VRAI ou 1} des exemples étiquetés par : {FAUX ou -1}. La recherche d'une telle séparatrice revient à chercher une fonction hypothèse $h(x) = w.x + w_0$ telle que:

$$w.x + w_0 \{ {\geq 0 \atop < 0} \Rightarrow u_i = \{ {+1 \atop -1}$$

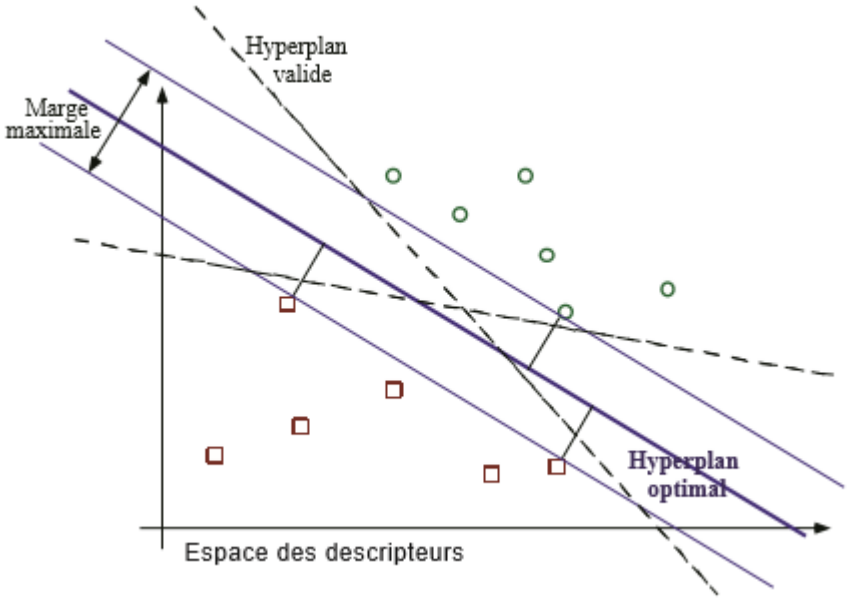


Figure 4. Séparation de deux ensembles de points par des séparateurs linéaires. (Cornuéjols A. *et al*. 2003).

Lorsqu'il existe un séparateur linéaire entre les points d'apprentissage, il en existe en général une infinité. On peut alors chercher parmi ces séparateurs celui qui donnera la règle qui se généralisera le mieux à d'autres données que celles de l'ensemble d'apprentissage. D'où vient la notion de la marge.

La marge d'un hyperplan est définie comme étant la distance entre l'hyperplan et la donnée la plus proche. Formellement, si $dist(x, w, w_0)$ représente la distance euclidienne entre le point x et l'hyperplan $w.x + w_0 = 0$, alors la marge M est définie ainsi :

$$M = min\{dist(x_i, w, w_0) : i = 1, \dots, l\}$$

b- Données non linéairement séparables

Il existe bien des cas pour lesquels il est impossible de séparer entièrement les données avec un hyperplan. Afin de régler ce problème, il est possible d'appliquer une transformation aux données de sorte qu'une fois transformées, elles soient linéairement séparables. Les données se trouvent dans ce qu'on appelle l'espace de redescription (*feature space*). Le passage à un espace de redescription est réalisé via une *fonction noyau*. Francoeur (Francoeur D. 2010) a détaillé le principe des SVM dans le cas de données non linéairement séparables.

Dans ce qui suit, nous allons exposer les différents travaux de classification de textes basés sur cette méthode.

2.2.1.2. Catégorisation de textes par les SVM

Les SVM sont largement utilisées dans la catégorisation de documents.

Joachims (Joachims T. 1998) est le premier qui a introduit cette méthode pour la classification de textes. L'ensemble d'expérimentations exécutées par lui montre l'efficacité des SVM dans la CT par leur capacité de généralisation dans un espace de dimension impressionnant, autrement dit, les SVM n'ont pas besoin d'une sélection de descripteurs ce qui rend la tâche de la CT considérablement aisée.

Les SVM sont aussi utilisés par (Yetisgen-Yildiz M. Pratt W. 2005) comme la technique de classification principale pour évaluer les différentes représentations des titres issus du corpus appelé OHSUMED.

(Polavarapu N. et al., 2005) ont étudié les SVM sur un ensemble d'articles liées aux épidémies génomiques provenant de la base (Human Genome Epidemiology « HuGE »)[4]. Cette étude comporte la comparaison des SVM avec d'autres algorithmes d'apprentissage automatique, leurs résultats montrent que les SVM sont meilleurs par rapport aux autres techniques avec un taux de classification correcte élevé (96.8%).

Une série d'expérimentations lancées par (Jimeno-Yepes. A., *et al*., 2014) dont le but de comparer un algorithme d'apprentissage profond d'un réseau de neurones (en anglais: Deep Learning Networks « DBN »), une vue d'ensemble sur les DBN est donné par (Schmidhuber J. 2014), et l'algorithme des séparateurs à vaste marge (SVM) sur un corpus obtenu du site MTI ML[5] avec 24 727 articles pour l'apprentissage et 12 363 pour le test, en choisissant comme classes les 10 termes fréquents du MeSH. Contrairement aux autres travaux exposés précédemment, les résultats montrent que les DBN sont plus performants que les SVM, surtout, lorsqu'il s'agit d'un large corpus. Cependant, le temps de calcul que prend les SVM (il est de l'ordre de quelques minutes) est négligeable vis-à-vis le temps que les DBN ont pris (5 jours pour effectuer l'apprentissage).

[4] http://www.cdc.gov/genomics/hugenet
[5] http://ii.nlm.nih.gov/MTI ML

2.2.2. Classification naïve bayésienne

2.2.2.1. Principe

La classification naïve bayésienne repose sur la règle de Bayes. Cette règle dit que la probabilité *a posteriori* d'une hypothèse est égale à sa probabilité *a priori* multipliée par la vraisemblance des données étant donné l'hypothèse. Plus formellement, la célèbre règle de Bayes s'écrit comme suit :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

L'importance pratique de cette règle réside dans le fait qu'on peut ré-exprimer la probabilité *a posteriori* difficile à calculer, en termes de probabilités *a priori* et conditionnelles plus facile à obtenir. Le théorème de Bayes a beaucoup d'applications dans le domaine du traitement de l'information notamment en traitement de la parole, traitement des images et bien d'autres. Nous nous focalisons à son application dans la catégorisation de textes.

(Ngouana E. Mayaya S. 2005) décrivent un modèle bayésien pour la classification de textes :

La classification bayésienne naïve de textes est l'une des approches probabilistes de classification simple. Cette approche est basée sur un modèle probabiliste dérivant du théorème de Bayes qui fait l'hypothèse que les mots qui apparaissent dans un document sont indépendants les uns des autres. Ce qui n'est pas tout à fait le cas dans la pratique. Supposons que nous disposons de n catégories de documents, déterminer à quelle catégorie C_i sera associé un document D revient à calculer la probabilité d'appartenance du document D à la catégorie C_i. En se basant sur le théorème précédemment énoncé, on peut calculer cette probabilité de la façon suivante :

$$P(C_i|D) = \frac{P(D|C_i) * P(C_i)}{P(D)}$$

Pour calculer la probabilité qu'un document D soit associé à une classe C_i, on considère que le document D est composé d'un ensemble de mots que nous noterons $W_1, \ldots, W_m$, calculer $P(D|C_i)$ reviendrait à calculer le produit des probabilités d'apparition de chaque mot W_j dans la catégorie C_i. Ce calcul se justifie par l'hypothèse selon laquelle tous les mots apparaissent indépendamment les uns des autres dans un document. Ce qui permet finalement d'écrire :

$$P(D|C_i) = \prod_{j=1<m,\ i=1<n} P(W_j, C_i)$$

Après cette brève description du modèle de classification bayésienne naïve, nous exposons ci-après les travaux qui ont utilisé cette approche.

2.2.2.2. Catégorisation de textes par Naïve Bayes

Dans la littérature des travaux de catégorisation de textes, les classifieurs naïve bayes sont largement utilisés, par exemple (Joachims T. 1998).

Dans le domaine biomédical, (Larkey L S. Croft B W. 1996) ont appliqué une classification bayésienne naïve dans le contexte d'assigner les codes de classification internationales des maladies ICD9[6] aux comptes rendus de sortie des patients.

En bio-informatique, des expérimentations menées par (Matthews M., 2006) dans le but d'identifier les documents contenant les interactions entre deux protéines. Sa contribution rentre dans le cadre d'améliorer la catégorisation de texte au lieu d'utiliser les méthodes standards de représentation (représentation sac-de-mots), il a utilisé les techniques issus du Traitement Automatique du Langage Naturel (TALN) tels que : la reconnaissance des entités nommées, l'extraction de relation, etc … sur une collection de 2025 résumés PubMed classés en deux catégories (pertinent, non pertinent). La collection est subdivisée comme suit : 64% pour l'apprentissage, 20% pour le test et 16% pour la validation. Chaque document a subi un ensemble de traitements avec des techniques TALN. Les résultats sont présentés à un classifieur Naïve bayes (qui présente des meilleurs performances par rapport les SVM dans cette expérimentation) avec une sélection de 1500 descripteurs. Son système a obtenu des meilleures performances : une F-Mesure de 68.1 pour un système qui combine les attributs issus d'un système TALN avec les attributs standards de catégorisation de texte, comparé avec une F-Mesure de 62.0 pour un système qui utilise seulement des descripteurs standards et une F-Mesure de 61.9 qui utilisent l'extraction des relations.

2.2.3. Réseaux de neurones artificiels

Un réseau neuronal artificiel (RNA) est l'association, en un graphe plus ou moins complexe, d'objets élémentaires, les *neurones formels*. La notion du neurone formel est décrite par Mc Culloch et Pitts en 1943 (McCulloch W. Pitts W. 1943). (voir Figure 5).

[6] **ICD-9** est l'acronyme de : *International Classification Disease 9*.
`http://www.cdc.gov/nchs/icd/icd9.htm`.

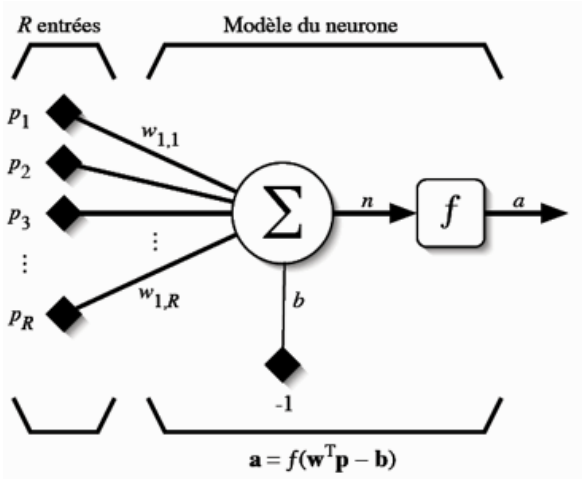


Figure 5. Représentation d'un neurone formel.

L'apprentissage d'un RNA se fait par plusieurs algorithmes d'apprentissage, (Svozil D. et al, 1997) ont expliqué en détail l'algorithme de la rétro-propagation (back propagation) pour l'apprentissage d'un perceptron multi-couches (PMC), (voir Figure 6).

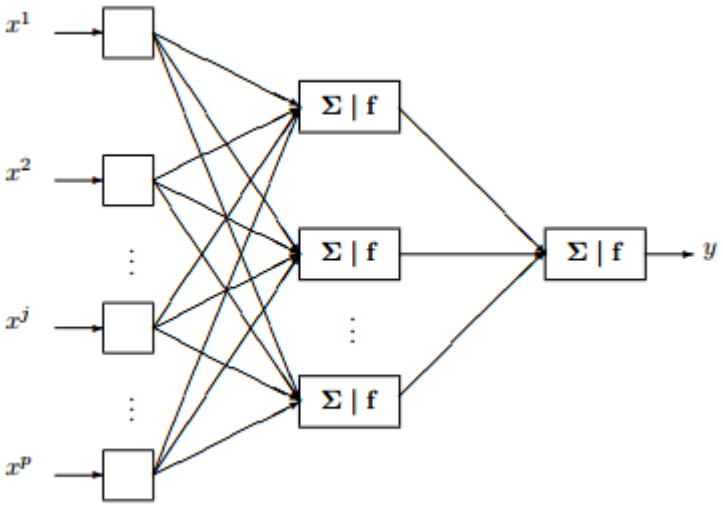


Figure 6. Exemple d'un perceptron multi-couches avec une couche cachée et une couche de sortie.

Dans le cadre de classification de documents biomédicaux, un travail effectué par (Ruiz M. E., Srinivasan P., 1998) en utilisant un réseau de neurones en comparant deux algorithmes d'apprentissage : i) algorithme de la rétro-propagation ; ii) algorithme de contre propagation (un apprentissage compétitif avec une architecture caractérisée par de deux couches : une couche de Kohonen et une couche de Grossberg) sur une collection de 2 344 articles de MEDLINE pour résoudre le problème de reconnaissance des termes MeSH d'un document particulier à partir des mots les plus importants dans le document. Le modèle développé reçoit en entrée les descripteurs de chaque document, la sortie c'est les étiquettes de MeSH, un

certain prétraitement de lemmatisation et élimination de mots vides est effectué avant la classification, les résultats obtenus dans cette expérimentation montre que l'algorithme d'apprentissage de la rétro-propagation donne des meilleures performances par rapport au contre propagation.

2.3. Métriques d'évaluation

Pour affirmer qu'une telle représentation ou tel algorithme d'apprentissage sont meilleurs par rapports aux autres, il fallait que la sortie d'un système de catégorisation de textes s'exécute d'une manière efficace par rapport à un autre. Nous allons discuter dans ce qui suit un modèle d'évaluation comportant des mesures simples et largement utilisées afin de mesurer l'efficacité d'un système CT.

2.3.1. Table de contingence

Lors de la catégorisation multi-classes de textes, c'est-à-dire lorsque $|C| > 2$, une approche commune consiste à « couper » le processus de catégorisation en sous problèmes. Chaque sous-problème concerne uniquement une classe et l'objectif est alors de juger si le nouveau texte « *appartient* » ou « *n'appartient* » pas à cette classe par opposition aux autres. Lors de l'évaluation de tels classifieurs à partir d'un ensemble de test, quatre nombres sont importants pour chaque classe (voir la table 2) :

1. le nombre de textes correctement classés comme appartenant à la classe i, noté VP_i (pour Vrai Positif) ;
2. le nombre de textes incorrectement classés comme appartenant à la classe, noté FP_i (pour Faux Positif) ;
3. le nombre de textes incorrectement rejetés, noté FN_i (pour Faux Négatif) ;
4. le nombre de textes correctement rejetés, noté VN_i (pour Vrai Négatif).

		Expert	
Classifieur		$Classe\ C_i$	$\neg\ Classe\ C_i$
	$Classe\ C_i$	VP_i	FP_i
	$\neg\ Classe\ C_i$	FN_i	VN_i

Tab 2. Table de contingence

2.3.2. Précision, Rappel, F-Mesure

- Rappel (R) :

Le Rappel présente la proportion de documents correctement classés par le système par rapport à tous les documents de la classe C_i.

$$Rappel(C_i) = \frac{Nombre\ de\ documents\ bien\ classés\ dans\ C_i}{Nombre\ de\ documents\ de\ la\ classe\ C_i}$$

$$R_i = \frac{VP_i}{VP_i + FN_i}$$

- **Précision (P) :**

La Précision c'est le rapport entre les documents classés parmi ceux classés par le système dans C_i. La précision mesure la capacité d'un système de classification à ne pas classer un document dans une classe, un document qui ne l'est pas.

$$Precision(C_i) = \frac{Nombre\ de\ documents\ bien\ classés\ dans\ C_i}{Nombre\ de\ documents\ classés\ dans\ C_i}$$

$$P_i = \frac{VP_i}{VP_i + FP_i}$$

- **F-Mesure :**

Elle permet de combiner les deux mesures précédentes (Rappel, Précision), défini comme suit :

$$F_1 = \frac{2 * P * R}{P + R}$$

3 Conclusion

Nous avons tenté le long de ce chapitre de présenter les algorithmes d'apprentissage les plus célèbres, ainsi que leurs applications dans le contexte de catégorisation de documents. Beaucoup d'auteurs affirment que les SVM et les classifieurs naïves bayésiens sont les classifieurs les plus appropriés à cette activité, en se basant sur des mesures d'évaluations et en comparant ces différentes techniques.

Chapitre IV : Evaluations Expérimentales

1 Introduction

Dans le chapitre II et III, nous avons présenté un état de l'art sur les méthodes de représentation du texte et les algorithmes d'apprentissage artificiel appliqués dans la catégorisation automatique de texte.

Dans cette partie nous nous intéressons à l'évaluation expérimentale. Cette évaluation consiste à comparer deux méthodes d'apprentissage, et de mesurer l'impact d'utiliser une méthode de sélection de descripteurs sur les résultats d'un classifieur.

D'abord, nous présentons la collection de textes biomédicaux utilisés pour la CT. Ensuite, nous allons décrire l'environnement matériel et logiciel sur lequel l'évaluation a été déroulée. Puis, nous exposons les méthodologies pour passer d'un document « semi-structuré » à une représentation vectorielle traitable par les méthodes d'apprentissage automatique. Aussi, une section est consacrée à la présentation et la discussion des résultats expérimentaux.

2 Collection MEDLINE d'évaluation

La collection des documents MEDLINE que nous utilisons pour l'évaluation est un sous-ensemble de la collection utilisé dans le challenge BioASQ[7].

BioASQ est une compagne d'évaluation financé par l'Union Européenne, cette compagne a pour objectif d'installer des compétitions dans l'indexation sémantique des références biomédicales dans une large échelle et les Systèmes Questions / Réponses (SQR), respectivement, dans deux tâches : (Tâche 1a) et (Tâche 1b). Nous nous intéressons à la *Tâche 1a* dans laquelle la classification de documents est abordée.

Cette compagne a pris une place entre Mars et Septembre 2013 et offre des collections larges d'apprentissage dès cette date jusqu'à présent. La seule différence entre ces différentes collections à travers ces années c'est les descripteurs MeSH qui sont utilisés pour l'annotation.

Les collections d'apprentissage sont proposées sous format JSON (Java Script Object Notation).

[7] The BioASQ Challenge. `http://www.bioasq.org`

- **Java Script Object Notation (JSON)**

Le JSON est un format léger d'échange de données aisément analysable ou générable par des machines. Il est complètement indépendant de tout langage, mais les conventions qu'il utilise seront familières à tout programmeur habitué aux langages descendant du C, comme par exemple : C lui-même, C++, C#, Java, JavaScript, Perl, Python et bien d'autres.

Chaque document JSON de la collection contient 6 champs : résumé (abstractText), les termes MeSH (meshMajor), l'identifiant de la référence dans PubMed (pmid), titre (title), l'année de la publication (year), comme c'est illustré ci-après :

```
{"articles": [
      {
            "abstractText": "text..",
            "journal": "journal..",
            "meshMajor": ["mesh1",...,"meshN"],
            "pmid": "PMID",
            "title": "title..",
            "year": "YYYY"
      },..., {..}
]
}
```

Tab3. Format du document JSON de la collection d'apprentissage.

La collection que nous avons téléchargée contient environ 6 millions de références (8 Go de données) annotée par 27 150 descripteurs MeSH. Le tableau (Tab 4) ci-après récapitule ces informations :

	Collection d'Apprentissage v. 2016b
Nombre d'articles	4 917 245
Moy. MeSH/articles	13.01
Nombres de descripteurs couverts	27 150
Taille (zip/unzip)	2.4Go/7.92Go

Tab 4. Informations sur la collection d'Apprentissage BioASQ.

3 Plan Expérimental

3.1. Configuration matérielle et logicielle

Nos expérimentations sont exécutées sur une machine (Packard bell, avec un processeur *Intel® Core™ i3*) avec 2 Go de RAM. Dans notre implémentation, la Tokenisation, l'élimination de mots vides, la racinisation (en anglais, Stemming avec l'algorithme de Porter) sont assurés par le software LingPipe[8]. La lemmatisation est testée par l'outil BioLemmatizer, qui est un software spécialisé pour effectuer une analyse morphologique de la littérature biomédicale développé par (Liu H. et al., 2012). Cependant, l'implémentation de cet outil

[8] http://alias-i.com/lingpipe/

nécessite l'utilisation d'une mémoire de taille 1Go au minimum et demande un temps d'exécution important.

Nous avons utilisé également l'environnement de fouille de données Weka-3.6.12 (Waikato Environment for Knowledge Analysis) pour développer et évaluer les performances des classifieurs. Les deux algorithmes d'apprentissage utilisés, avec les paramètres standards de Weka :

- Naïve Bayes : *weka.classifiers.bayes.NaiveBayes*
- Séparateurs à vastes marges: *weka.classifiers.functions.SMO*

La version JAVA[9] utilisée est : (Java Development Kit « JDK » : 1.8.0_74)

3.2. Segmentation du fichier JSON

La taille de la RAM (2 Go) est contraignante. Il est impossible de lire et charger un fichier de taille volumineuse avec une telle RAM. Pour pallier cette contrainte, nous avons adapté un algorithme pour découper cette collection à des fichiers réduits de taille 1Mo (environ 500 documents dans chaque fichier JSON).

3.3. Analyse du fichier JSON

Cette étape a pour objectif, l'extraction du résumé et du titre en décodant les éléments de chaque document JSON. Pour ce faire, nous avons utilisé l'API[10] « *java-json* ».

Les étapes de décodage d'un fichier JSON sont résumées dans l'algorithme suivant :

```
    Entrées : Le document JSON
    Sortie : Chaîne concaténant le titre et le résumé de la référence
DEBUT
    1 : {Lire et parcourir le document JSON ligne par ligne et le mettre dans
    une chaîne de caractère}
    2: Analyser le fichier JSON
    Pour chaque document JSON
        resume ← jsonElement(« abstractText ») ;
        titre ← jsonElement(« title ») ;
        identifiant ← jsonElement(« pmid ») ;
        texte ← resume+titre ;
        Pour chaque meshDescripteur dans le document
        |       fichier ← getName.jsonArray(meshDescripteur) ;
        Fin Pour
    Fin Pour
FIN
```

Algorithme 2 – Algorithme de décodage d'un fichier JSON.

[9] http://java.sun.com

[10] API est l'acronyme de Application Programming Interface.

3.4. Prétraitement

Cette étape sert à préparer les données afin de garder que l'information pertinente du texte. Nous avons comparé deux softwares pour préparer les données : LingePipe, BioLemmatize

Entrées : texte à prétraiter
Sortie : fichier texte prétraité
DEBUT

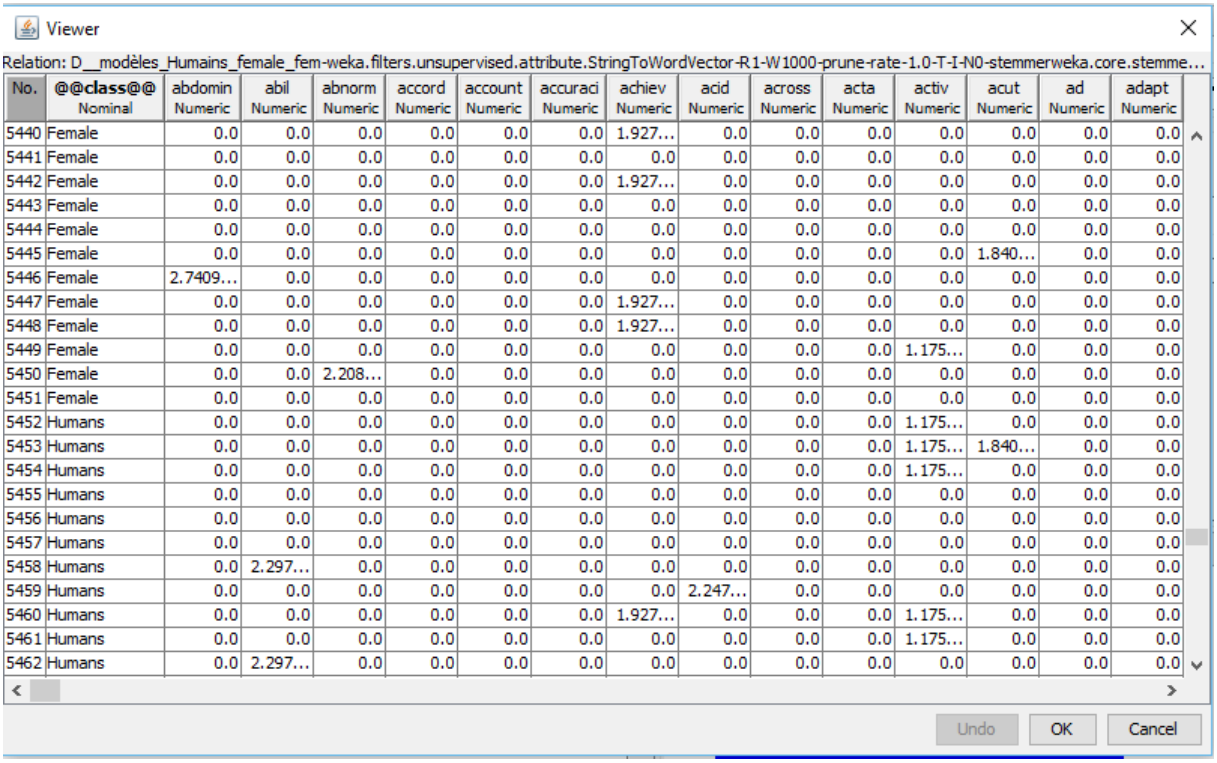
1 : Pour chaque $texte_i$ du corpus faire :
| texte Segmente [] ← tokenizer ($texte_i$) ;
Fin pour
2 : {pretraitement1 ← convertir les majuscules en minuscules et suppression des signes de ponctuations ;
pretraitement2 ← élimination de mots vides ;
pretraitement3 ← le Stemming avec l'algorithme de Porter ;}
3: Enregistrer les fichiers txt dans des répertoires, chacun d'entre eux a comme nom un descripteur MeSH ;

FIN

Algorithme 3 – prétraitement du texte avec le software LingPipe.

3.5. Vectorisation avec l'environnement Weka

La vectorisation d'un texte dans l'environnement Weka est assuré par le filtre : StringToWordVector (package *weka.filters.unsupervised.attribute*). Ce filtre applique une conversion du texte à un ensemble de descripteurs formés par des mots et leurs TF-IDF (voir Figure 7).



No.	@@class@@ Nominal	abdomin Numeric	abil Numeric	abnorm Numeric	accord Numeric	account Numeric	accuraci Numeric	achiev Numeric	acid Numeric	across Numeric	acta Numeric	activ Numeric	acut Numeric	ad Numeric	adapt Numeric
5440	Female	0.0	0.0	0.0	0.0	0.0	0.0	1.927...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5441	Female	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5442	Female	0.0	0.0	0.0	0.0	0.0	0.0	1.927...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5443	Female	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5444	Female	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5445	Female	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.840...	0.0	0.0
5446	Female	2.7409...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5447	Female	0.0	0.0	0.0	0.0	0.0	0.0	1.927...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5448	Female	0.0	0.0	0.0	0.0	0.0	0.0	1.927...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5449	Female	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.175...	0.0	0.0	0.0
5450	Female	0.0	0.0	2.208...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5451	Female	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5452	Humans	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.175...	0.0	0.0	0.0
5453	Humans	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.175...	1.840...	0.0	0.0
5454	Humans	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.175...	0.0	0.0	0.0
5455	Humans	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5456	Humans	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5457	Humans	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5458	Humans	0.0	2.297...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5459	Humans	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.247...	0.0	0.0	0.0	0.0	0.0	0.0
5460	Humans	0.0	0.0	0.0	0.0	0.0	0.0	1.927...	0.0	0.0	0.0	1.175...	0.0	0.0	0.0
5461	Humans	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.175...	0.0	0.0	0.0
5462	Humans	0.0	2.297...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 7. Une fenêtre de la plateforme Weka illustre les attributs et leurs valeurs dans chaque document (individu).

Pour pouvoir utiliser un classifieur sur cette matrice de données, Weka accepte des formats spécifiques (CSV, XRFF, et ARFF). Nous avons utilisé le format ARFF. Dans ce qui suit, nous donnons un aperçu sur ce format.

- ## Attribute-Relation File Format (ARFF)

C'est un fichier texte codé en ASCII décrivant une liste d'instances partageant un ensemble d'attributs. Le fichier ARFF est développé par l'université de Waikato pour son utilisation dans l'outil Weka pour l'apprentissage automatique.

- **Aperçu du format ARFF**

On distingue deux sections dans le fichier ARFF. La première section c'est l'entête d'informations (**Header**), suivie par les informations concernant les données (**Data**).

L'en-tête du fichier ARFF contient le nom de la relation, une liste d'attributs (les colonnes dans les données), et leurs types.

Voici un exemple d'un en-tête du corpus MEDLINE que nous avons utilisé :

```
@relation 'BioMed'

@attribute @@class@@ {Adolescent,Adult,Aged,Animals,Child,DNA,Humans,Male,Mice}
@attribute abnorm numeric
@attribute access numeric
@attribute accord numeric
```

```
@attribute accuraci numeric
@attribute achiev numeric
@attribute across numeric
@attribute acta numeric
@attribute action numeric
@attribute activ numeric
@attribute acut numeric
@attribute addit numeric
@attribute address numeric
@attribute adher numeric
@attribute adjust numeric
@attribute administ numeric
@attribute administr numeric
@attribute adolesc numeric
@attribute adult numeric
…
```

Les données (**Data**) du fichier ARFF ont cette forme suivante :

```
{0  Adult,65  0.693147,88  0.693147,107  0.693147,114  0.693147,128  0.693147,175
0.693147,220  0.693147,286  0.693147,343  0.693147,345  0.693147,347  0.693147,378
0.693147,407  0.693147,410  0.693147,412  0.693147,436  0.693147,443  0.693147,481
0.693147,487  0.693147,494  0.693147,508  0.693147,539  0.693147,568  0.693147,581
0.693147,607  0.693147,638  0.693147,675  0.693147,681  0.693147,696  0.693147,698
0.693147,731  0.693147,735  0.693147,830  0.693147,848  0.693147,852  0.693147,857
0.693147,919  0.693147,937  0.693147,956  0.693147,970  0.693147,1411  0.693147,1608
0.693147,1707 0.693147}

{0  Aged,107  0.693147,123  0.693147,148  0.693147,156  0.693147,333  0.693147,388
0.693147,391  0.693147,434  0.693147,452  0.693147,597  0.693147,632  0.693147,642
0.693147,801 0.693147,830 0.693147,831 0.693147,875 0.693147,882 0.693147}

…
```

Ensuite, ce fichier ARFF est passé à un algorithme de classification afin de développer le modèle de prédiction.

L'association des descripteurs MeSH aux références est un problème de classification multi-labels (voir multi-étiquettes), puisque plusieurs descripteurs peuvent être attribués à un article. Nous avons transformé ce problème à des sous problèmes de classification binaire. Autrement dit, pour chaque descripteur MeSH un classifieur est développé afin de décider si un document pourra être affecté à ce descripteur ou pas. Dans la prochaine section, nous présentons les résultats d'évaluation de 28 descripteurs MeSH de la collection.

4 Résultats et discussions

La collection d'évaluation utilisée est un sous ensemble de la collection BioASQ, nous avons choisi les descripteurs MeSH les plus fréquents dans l'annotation des citations. Comme résultats, on a obtenu une collection d'évaluation de 57 469 références (voir **Tab 5**).

descripteur	fréquence	Descripteur	fréquence
Humans	11283	**Rats**	664
Female	7195	**Cell Line, Tumor**	647
Male	7003	**Prognosis**	559
Animals	4582	**Infant**	528
Middle Aged	3991	**Cell Proliferation**	483
Adult	3921	**Pregnancy**	477
Aged	3003	**Disease Models, Animal**	469
Mice	1967	**Case-Control Studies**	435
Young Adult	1710	**Cell Line**	413
Adolescent	1606	**Magnetic Resonance Imaging**	408
Aged, 80 and over	1326	**Amino Acid Sequence**	377
Treatment Outcome	1284	**Apoptosis**	369
Child	1088	**Brain**	342
Risk Factors	1008	**RNA, Messenger**	331

Tab 5. Les descripteurs MeSH dans la collection d'évaluation avec les fréquences de leurs références dans la base d'apprentissage.

Nous avons conduit une série d'expérimentations, en comparant les deux méthodes Naïve Bayes (NB) et les Séparateurs à Vastes Marges (SVM) sans sélection de descripteurs et avec une sélection de descripteurs.

4.1. Résultats

- Expérimentation 1

La première expérimentation consiste à comparer les deux méthodes de classification sur l'ensemble de descripteurs sans faire appel à une méthode de sélection de descripteurs. Nous avons subdivisé la collection en deux bases : 70% pour l'apprentissage (40 229 références) et 30% pour le test (17 240 références).

Le tableau 6 (Tab 6) montre les performances de la classification des deux méthodes Naïve Bayes (NB) et les séparateurs à vaste marge (SVM) sans sélection d'attributs, en termes de Précision, Rappel et F-Mesure :

	Naïve Bayes			**SVM**		
	Précision	**Rappel**	**F-Mesure**	**Précision**	**Rappel**	**F-Mesure**
Humans	0.679	0.581	0.626	0.716	0.694	0.703
Female	0.475	0.58	0.522	0.472	0.485	0.475
Male	0.83	0.48	0.608	0.816	0.914	0.862
Aged	0.291	0.261	0.275	0.205	0.212	0.208
Adult	0.732	0.589	0.652	0.675	0.750	0.711
Middle Aged	0.502	0.466	0.483	0.42	0.482	0.449
Animals	0.953	0.757	0.844	0.905	0.899	0.902
Mice	0.912	0.816	0.861	0.908	0.902	0.905
Rats	0.601	0.778	0.678	0.73	0.744	0.737
Adolescent	0.291	0.554	0.381	0.256	0.126	0.169
Child	0.753	0.653	0.699	0.631	0.686	0.657

Risk Factors	0.666	0.663	0.665	0.56	0.605	0.582
Young Adult	0.326	0.479	0.388	0.175	0.127	0.147
Brain	0.674	0.632	0.652	0.538	0.516	0.527
Cell Line	0.5	0.744	0.598	0.543	0.678	0.603
Treatment Outcome	0.704	0.681	0.692	0.637	0.603	0.619
RNA, Messenger	0.886	0.796	0.858	0.89	0.827	0.857
Amino Acid Sequence	0.839	0.912	0.874	0.86	0.912	0.885
Pregnancy	0.831	0.868	0.849	0.911	0.824	0.865
Disease models, Animal	0.215	0.641	0.322	0.079	0.085	0.082
Case-Control studies	0.192	0.444	0.269	0.959	0.935	0.947
Infant	0.427	0.528	0.472	0.317	0.286	0.301
Cell line, Tumor	0.611	0.513	0.558	0.512	0.43	0.468
Cell Proliferation	0.504	0.409	0.452	0.427	0.43	0.428
Apoptosis	0.284	0.951	0.437	0.933	0.957	0.945
Magnetic Resonance Imaging	0.743	0.777	0.759	0.657	0.677	0.667
Aged, 80 and over	0.183	0.262	0.215	0.059	0.029	0.039
Prognosis	0.55	0.591	0.57	0.515	0.5	0.507
Moyenne (Avg)	**0.577**	**0.622**	**0.596**	**0.582**	**0.583**	**0.58**

Tab 6. Les performances de classification avec les SVM et NB.

- Expérimentation 2

Dans la deuxième expérimentation, nous appliquons la méthode de sélection de descripteurs chi-2 (en anglais, chi-squared). Cette méthode permet de tester l'indépendance entre deux variables. Dans la sélection de descripteurs, la méthode de test χ^2 mesure l'indépendance entre le descripteur et la catégorie. Plus la valeur de χ^2 est grande, meilleure sera la corrélation entre ces deux variables. D'une autre part, si cette valeur est nulle, cela implique que le descripteur et la catégorie sont totalement indépendants. Donc, on a appliqué ce test en éliminant tous les descripteurs ayant une valeur de test nulle. Le tableau ci-après illustre les performances des deux classifieurs en termes de précision, rappel et F-Mesure :

	Naïve Bayes			**SVM**		
	Précision	**Rappel**	**F-Mesure**	**Précision**	**Rappel**	**F-Mesure**
Humans	0.678	0.606	0.64	0.619	0.929	0.743
Female	0.482	0.559	0.518	0.536	0.125	0.203
Male	0.876	0.734	0.799	0.825	0.971	0.892

Aged	0.396	0.227	0.289	0.433	0.088	0.147
Adult	0.739	0.781	0.76	0.715	0.966	0.822
Middle Aged	0.518	0.785	0.624	0.5	0.963	0.658
Animals	0.96	0.786	0.864	0.906	0.991	0.946
Mice	0.912	0.84	0.875	0.905	0.912	0.908
Rats	0.632	0.773	0.696	0.748	0.729	0.738
Adolescent	0.331	0.559	0.416	0.508	0.126	0.202
Child	0.767	0.798	0.782	0.683	0.873	0.767
Risk Factors	0.666	0.663	0.665	0.596	0.563	0.579
Young Adult	0.389	0.335	0.36	0.467	0.071	0.123
Brain	0.693	0.737	0.714	0.602	0.621	0.611
Cell Line	0.565	0.793	0.66	0.574	0.736	0.645
Treatment Outcome	0.718	0.676	0.696	0.651	0.619	0.634
RNA, Messenger	0.883	0.847	0.865	0.873	0.908	0.89
Amino Acid Sequence	0.873	0.904	0.888	0.918	0.886	0.902
Pregnancy	0.876	0.934	0.904	0.912	0.838	0.767
Disease models, Animal	0.248	0.683	0.364	0.071	0.007	0.013
Case-Control studies	0.983	0.919	0.95	0.959	0.944	0.951
Infant	0.568	0.491	0.527	0.55	0.273	0.365
Cell line, Tumor	0.644	0.58	0.61	0.584	0.523	0.552
Cell Proliferation	0.571	0.456	0.507	0.548	0.497	0.521
Apoptosis	0.92	0.983	0.95	0.941	0.957	0.949
Magnetic Resonance Imaging	0.798	0.762	0.78	0.717	0.7	0.708
Aged, 80 and over	0.281	0.114	0.162	0	0	0
Prognosis	0.543	0.614	0.576	0.445	0.534	0.486
Moyenne (Avg)	**0.649**	**0.676**	**0.659**	**0.635**	**0.62**	**0.627**

Tab 7. Les performances de classification SVM et NB avec une sélection de descripteurs.

4.2. Discussion des résultats

Les expérimentations montrent que la représentation de textes par des racines lexicales et lemmes est une amélioration de la représentation sac de mots et montrent des performances compétitives par rapport à d'autres représentations utilisés dans la littérature (Jimeno-Yepes et al. 2015), (Yetisgen-Yildiz M., Pratt W. 2005).

La sélection de descripteurs a un impact sur les performances de la classification. Dans ces expérimentations, nous avons employé la méthode chi-2. Cette méthode permet d'améliorer les performances des classifieurs Naïve Bayes et SVM. Contrairement aux

résultats des travaux précédents (Jimeno-Yepes et al. 2015), (Spolaor N. Tsoumakas G. 2013). La figure 8 montre les performances des classifieurs, en termes de F-Mesure, sans ou avec une sélection de descripteurs, où il est clair que les performances des deux classifieurs sont améliorées après une sélection de descripteurs. De plus, le NB montre de bonnes performances par rapport au SVM, ce résultat est en accord avec le travail de (Matthews M. 2006). Aussi, les SVM prend plus de temps que les NB.

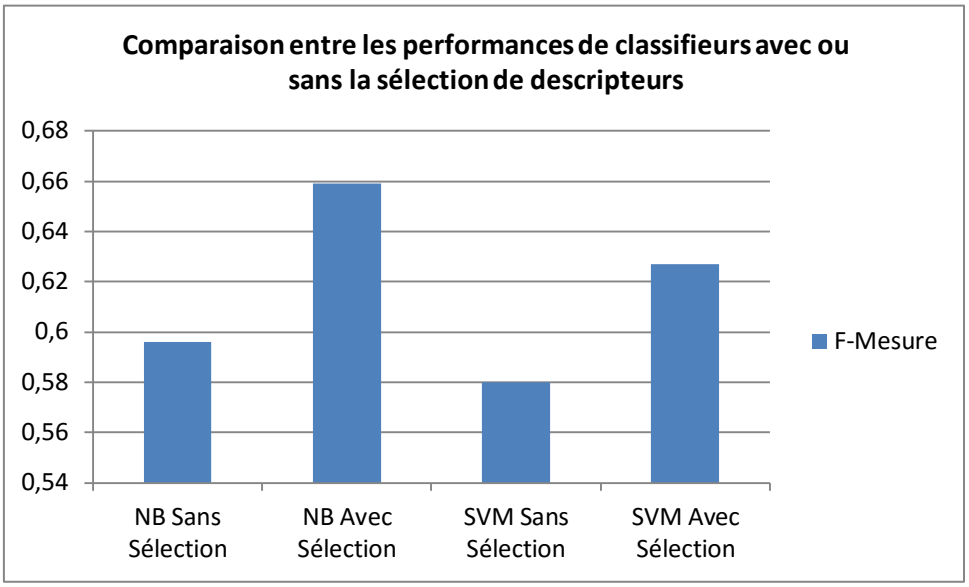


Figure 8. Les performances des classifieurs (NB et SVM), en termes de F-Mesure, sans ou avec une sélection de descripteurs.

La sélection de descripteurs utilisée avec un algorithme comme les SVM montre une diminution des performances de classification pour un ensemble de catégories par exemple : (Aged , 80 and over – Prognosis – Pregnancy – Young Adult) (voir Figure 11, 12), ce qui est en accord avec les travaux de (Jimeno-Yepes et al. 2015), (Spolaor N. Tsoumakas G. 2013).

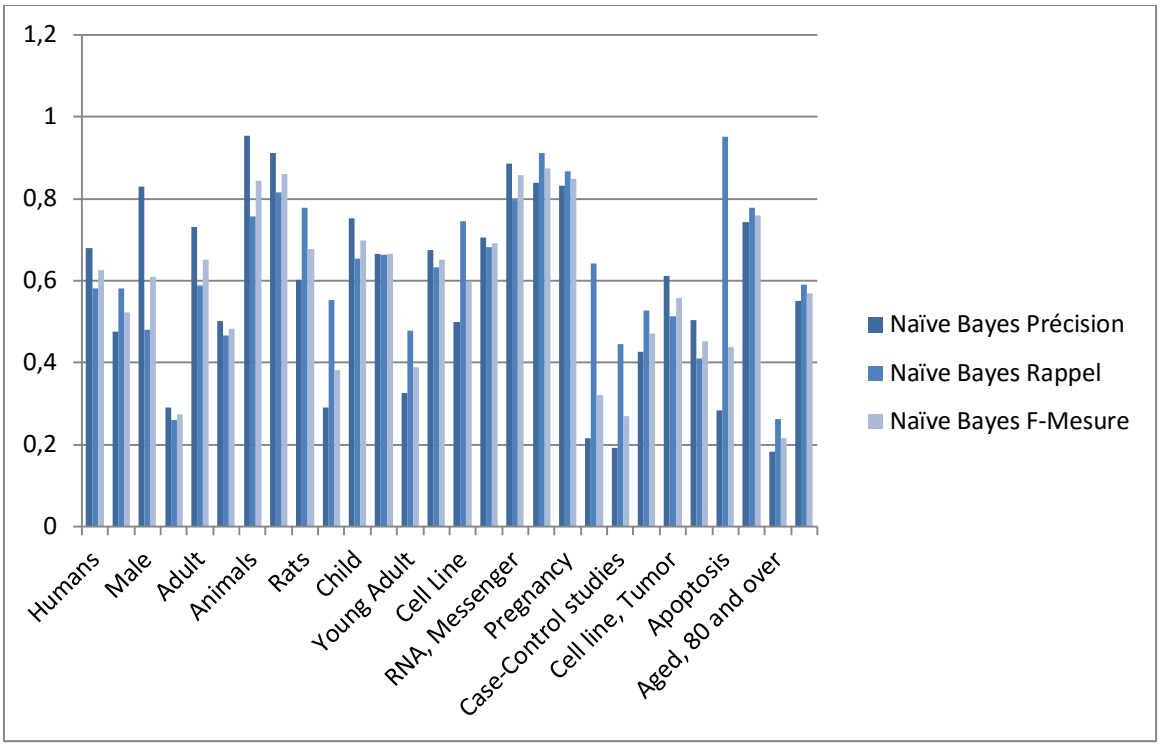


Figure 9. Performance du classifieur Naïve Bayes sans sélection de descripteurs.

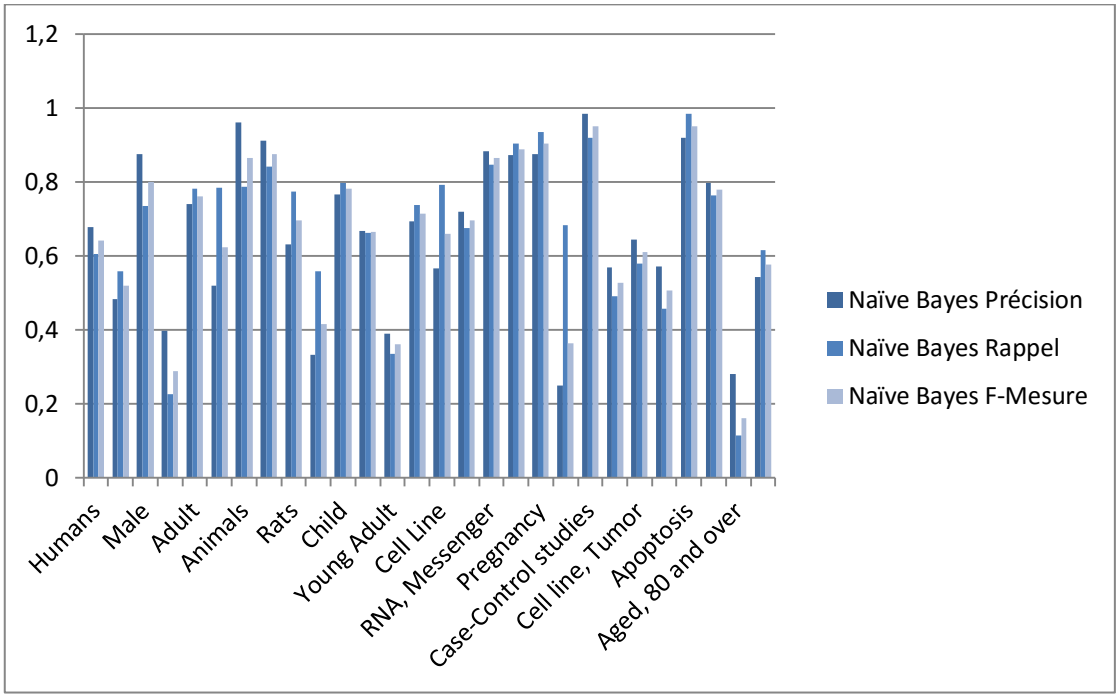


Figure 10. Performances du classifieur Naïve Bayes avec une sélection de descripteurs.

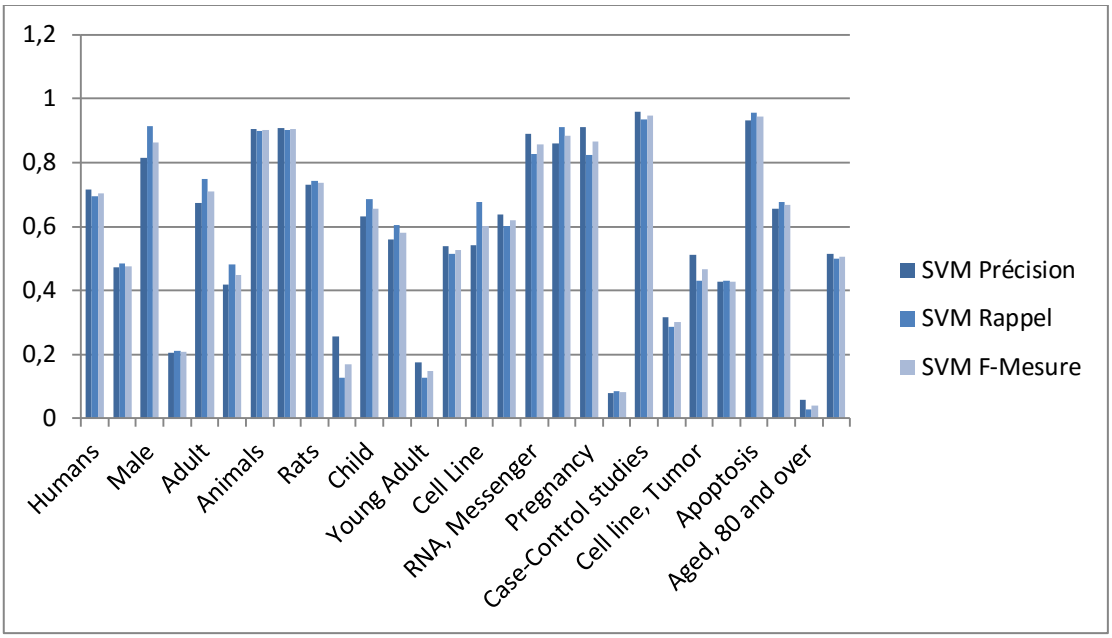


Figure 11. Performances du classifieur SVM, en termes de : précision, rappel, F-Mesure, sur chacune des catégories.

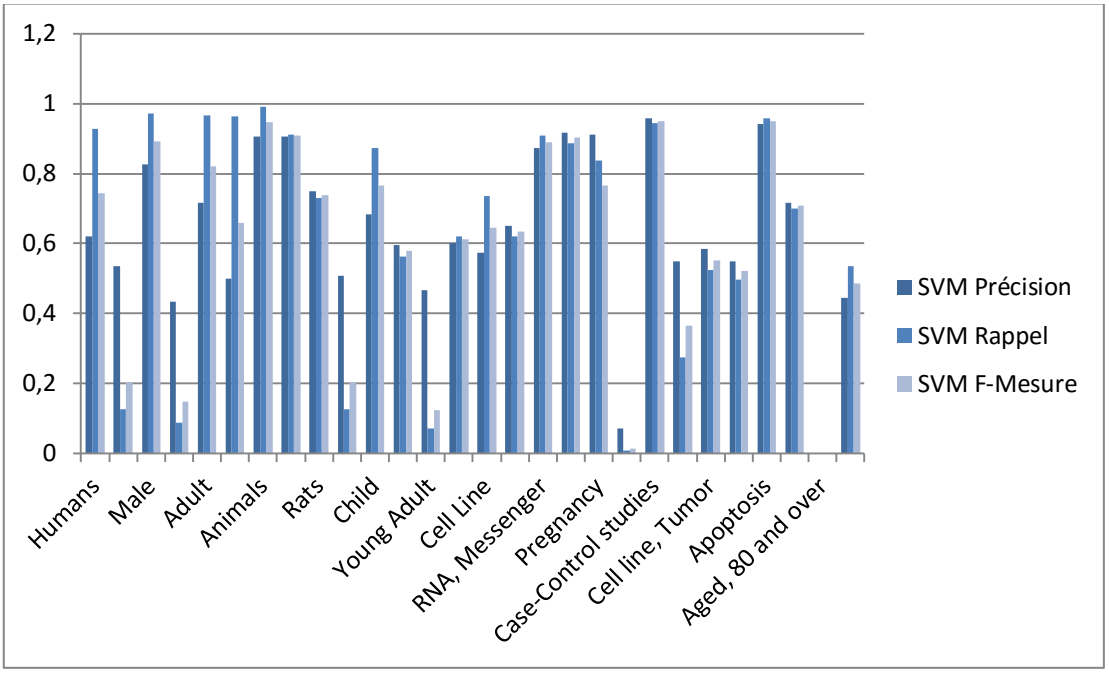


Figure 12. Performances du classifieur SVM avec une sélection de descripteurs sur chacune de catégories.

5 Conclusion

Les recherches dans la catégorisation des textes biomédicaux ont largement utilisé les SVM et les NB comme des algorithmes d'apprentissage. Dans cette partie d'évaluation, nous avons étudié l'application de ces deux algorithmes avec ou sans sélection de descripteurs pour la classification des références MEDLINE. Notre principale conclusion, c'est que : l'utilisation d'une méthode de sélection des descripteurs permet de réduire la dimensionnalité de l'espace de descripteurs, ce qui accélère le temps d'exécution de l'algorithme d'apprentissage. De plus, elle permet d'améliorer les performances des classifieurs. Cependant, cette sélection de descripteurs engendre parfois une diminution des performances du classifieur, ceci est interprété de la façon suivante : quand la méthode de sélection produit un nombre réduit de descripteurs, il est probable que la méthode de sélection ne prend pas en considération assez de descripteurs pour cette étiquette, une solution possible pour remédier ce problème consiste à augmenter la taille de la base d'apprentissage. Aussi, Nous avons trouvé que l'utilisation d'une représentation sac de mots améliorée montre de bonnes performances. Ce résultat corrobore les travaux précédents qui montrent l'efficacité d'une telle représentation.

Conclusion générale et perspectives

La catégorisation de texte est une activité primordiale dans la chaîne du Text Mining en général, et dans la Recherche d'information en particulier. Cette activité permet d'extraire des connaissances dissimulées dans une large collection de textes et les utilisent dans divers processus telle que l'indexation.

Dans notre mémoire nous avons étudié la tâche de la CT pour indexer les références biomédicales de la MEDLINE®.

L'utilisation d'une représentation standard (sac de mots) affinée avec les techniques du prétraitement donne des performances compétitives et encourageantes car elle caractérise parfaitement les propriétés statistiques des mots d'un texte vis-à-vis d'autres représentations utilisés dans d'autres travaux. Cependant, la représentation d'un texte produit un nombre impressionnant de descripteurs, qui peuvent entraver les performances d'algorithme d'apprentissage.

Les méthodes de sélection de descripteurs permettent de retrouver les mots représentatifs et éliminent ceux qui ne le sont pas. Nous avons utilisé une méthode de sélection connue sous l'appellation : *chi-2*, qui est basée sur le calcul d'un score entre deux variables. Cette méthode a une influence significative et a permis d'améliorer les performances des algorithmes d'apprentissage employées.

Les SVM et NB sont parmi les algorithmes d'apprentissage largement étudiés dans le contexte de la CT. Ces deux méthodes ont été utilisées avec les paramètres standards de l'environnement Weka. Les résultats obtenus avec les méthodes d'apprentissage automatique déjà présentées sont compétitives et encourageantes. Le NB a montré de bonnes performances par rapport aux SVM. Néanmoins, la comparaison de nos résultats avec d'autres travaux nous paraît non crédible, parce que les jeux de données utilisés dans l'évaluation ne sont pas les mêmes.

Les perspectives de ce travail sont nombreuses. Les perspectives à court terme, consistent à appliquer des méthodes de classification Multi-labels sur le jeu de données. à long terme, Nous comptons appliquer d'autres algorithmes d'apprentissage sur le corpus entier de la MEDLINE sachant que le traitement d'un tel espace vectoriel demandera beaucoup des ressources matérielles plus robustes.

Enfin, ce projet était l'occasion de mettre en application l'ensemble des acquis de ma formation Master « Informatique Biomédicale », et m'a permis d'acquérir de nouvelles compétences.

Références bibliographiques

(Apte C., et al. 1994) Apte C. Damerau F. Weiss S. 1994. Towards Language Independant automated learning of text categorization models. Proceedins of the Seventeeth Annual International ACM-SIGIR Conference on research and Development in Information Retrieval, Dublin, Ireland, pp. 21-30.

(Arfken G. 1985) Arfken G. 1985. "Gram-Schmidt Orthogonalization." §9.3 in *Mathematical Methods for Physicists, 3rd ed.* Orlando, FL: Academic Press, pp. 516-520.

(Aronson A.R., *et al.*, 2004) A.R. Aronson, J.G. Mork, C.W. Gay, S.M. Humphrey, and W.J. Rogers. 2004. The NLM Indexing Initiative's Medical Text Indexer. In Medinfo 2004: proceedings of the 11th World Conference on Medical Informatics,[San Francisco, september 7-11, 2004], page 268. OCSL Press.

(AZZOUG W. 2013) Azzoug Wassila. 2013. Contribution à la definition d'une approche d'indexation sémantique de documents textuels. Université M'HMED Bougara – Boumerdas, Algérie.

(Ba-Duy D. 2012) Dinh Ba-Duy. 2012. Accès à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques, Thèse doctorale, Université Toulouse 3 Paul Sabatier.

(Basili R. Moschitti A. 2005) Basili Roberto, Alessandro Moschitti. 2005. Automatic Text Categorization: from Information Retrieval to Support Vector Learning. Université de Rome, pp. 4-7.

(Bassil Y. 2012) Bassil Youssef. 2012. A Survey on Information Retrieval, Text Categorization, and Web Crawling. Journal of Computer Science & Research (JCSCR), Vol. 1, No. 6, pp. 1-11, December 2012.

(Borko H., et al., 1963) Borko H. Bernick M. 1963. Automatic document classification. Journal of the association for Computing Machinery, pp. 151-161.

(Borko H., et al., 1964) Borko H. Bernick M. 1964. Automatic document classification. Part II. Additional experiments. Journal of the association for Computing Machinery, 11(2):138-151. April, 1964.

(Cornuéjols A. *et al.* 2003) Cornuéjols A. Miclet L. Kodratoff K. 2003. Apprentissage Artificiel: Concepts et Algorithmes. Editions Eyrolles.

(Dahmani H. 2012) Dahmani Houria. 2012. Thèse: Classification des documents médicaux basée sur le Text Mining. Université Sâad Dahlab – Blida.

(Francoeur D. 2010) Francoeur D. 2010. Machines à vecteurs de support - Une introduction. Université de Sherbrooke. pp. 7-25.

(Hamill K. Zamora A. 1980) Hamill K. Zamora A. 1980. The use of titles for automated document classification. Journal of the American Society for Information Science, 33, pp. 396-402.

(Hirotoshi T. 2002) Taira Hirotoshi. 2002. Text Categorization using Machine Learning (Thèse de doctorat). Nara Institute of Science and Technology, pp. 10-11.

(Humphrey S. Miller N. 1987) Humphrey S. Miller N. 1987. Knowledge-based Indexing of the medical literature: The indexing Aid Project. Journal of the American Society for Information Science, 38(3), pp. 184-196.

(Ingersoll G. S. et al. 2013) Ingersoll Grant S. 2013. Morton Thomas S. Farris Andrew L. "Taming text: How to find, organize and manipulate it". By Manning Publications Co. pp. 175-177.

(JALAM R. 2003) Jalam Redwan. 2003. Apprentissage automatique et catégorisation de texte multilingues. (Thèse de doctorat). ERIC, Université Lumière Lyon2.

(Jimeno-Yepes A., et al., 2011) Jimeno-Yepes A, Wilkowski B, Mork JG, Van Lenten E, Demner Fushman D, Aronson AR. 2011. A bottom-up approach to MEDLINE indexing recommendations.*AMIA Annual Symposium Proceedings*. 2011:1583-1592.

(Jimeno-Yepes A., et al., 2014) Antonio Jimeno Yepes, Andrew MacKinlay, Justin Bedo, Rahil Garvani and Qiang Chen. 2014. Deep Belief Networks and Biomedical Text Categorisation. In Proceedings of Australasian Language Technology Association Workshop. pages 123−127.

(Joachims T. 1998) Joachims T. 1998. Text categorization with support vector machines: learning with many relevant features. In Proceedings of ECML-98, 10th European Conference on machine learning (Chemnitz, DE, 1998), pp. 137-142.

(Khan A. et al. 2010) Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan. 2010. A Review of Machine Learning Algorithms for Text-Documents Classification. JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY, VOL. 1, NO. 1, pp. 5-20. February 2010.

(Lewis D D. 1992) Lewis David D. 1992. Representation and Learning in Information Retrieval. Ph.D. thesis, University of Massachusetts at Amherst, pp. 46-50. February, 1992.

(Lipka N. 2013) Lipka N. 2013. Modeling Non-standard text classification tasks. (Thèse de doctorat). Faculty of Media, Bauhaus-Universität Weimar, Germany.

(Liu H. et al., 2012) Liu H. Christianien T. Baumgartner W. A. Jr. Verspoor K. 2012. BioLemmatizer : a Lemmatization tool for morphological processing of biomedical text. Journal of biomedical Semantics 3. pp. 3.

(Lovins J. 1968) Lovins J. 1968. Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11, 22-31.

(Matthews M. 2006) Matthews M. 2006. Improving biomedical text categorization withNLP. Proceedings of the SIGs, the Joint BioLINK-Bio-Ontologies Meeting, pp. 93-96.

(McCulloch W. Pitts W. 1943) McCulloch W. Pitts W. 1943. A Logical Calcul of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, Volume 5. pp. 116-133.

(Mitchell T.1997) Mitchell T.1997. Machine Learning. McGraw Hill, New York, USA.

(NAKACHE D. 2007) Nakache Didier. 2007. Thèse : Extracion automatique des diagnostiques à partir des comptes rendus médicaux textuels. Conservatoire National des Arts et des Métiers. Septembre 2007.

(Névéol A. *et al.*, 2009) Névéol Aurélie, Sonya E. Shooshan, Susanne M. Humphrey, James G. Mork, Alan R. Aronson. 2009. A Recent Advance in the Automatic Indexing of the Biomedical Literature. J Biomed Inform. Page: 2:28.

(Ngouana E., Mayaya S. 2005) Ngouana E., Mayaya S. 2005. Classification Bayesienne naïve de textes. Faculté Polytechnique de Mons.

(Paice C. 1996) Paice C. 1996."Method for evaluation of stemming algorithms based on error counting", Journal of the American Society for Information Science, 47 (8), pp. 632-349.

(Polavarapu N. et al., 2005) Polavarapu N. Navathe S. B. Ramnarayanan R. ul Haque A. Sahay S. Liu Y. 2005. investigation into biomedical literature classification using support vector machines. In proceedings IEEE Computational Systems Bioinformatics Conference (CSB'05) ,pp. 366 – 374.

(Porter M. F. 1980) Porter M. F. 1980. An algorithm for suffix stripping. Program14,3,130-1367.

(Reese M R. 2015) Reese M Richard. 2015. Naturel Langage Processing with JAVA. Published By Packt Publishing Ltd, pp. 54.

(Roussey C. 2001) Roussey C. 2001. Une méthode d'indexation sémantique adaptée aux corpus multilingues. Thèse de doctorat, INSA de Lyon.

(Ruiz E. M., Srinivasan P., 1998) Ruiz M. E., Srinivasan P. 1998. "Automatic Text Categorization Using Neural Network",In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, pp. 59-72.

(Salton G. and Buckley C. 1988) Salton G. and Buckley C. 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 5, 513–523., pp. 323–328.

(Schmidhuber J. 2014) Schmidhuber J. 2014.Deep Learning in Neural Networks: An Overview. The Swiss AI Lab IDSIA. University of Lugano & SUPSI – Switzerland.

(Scott S. Matwin., 1999) Scott S, Matwin S. 1999. Feature engineering for text classification. In: ICML, Volume99. pp. 79-83.

(Sebastiani F. 2002) Sebastiani Fabrizio. 2002. Machine Learning in automated text categorization. ACM Computing Surveys, vol. 34, pp. 1-47.

(Spolaor N. Tsoumakas G. 2013) Spolaor N, Tsoumakas G. 2013. Evaluating feature selection methods for multi-label text classication. In: BioASQworkhsop.

(Svozil D. et al, 1997) Svozil Daniel, Vladimir Kvasnicka, Jiri Pospichal. 1997. Introduction to Multi-layer feed-forward neural networks. Chemometrics and Intelligent Laboratory Systems 39. pp. 43-62.

(Yetisgen-Yildiz M., Pratt W. 2005) Yetisgen-Yildiz M., Pratt W. 2005. The Effect of Feature Representation on MEDLINE Document Classification. In Proceedings: AMIA Symposium. pp. 850-853.

Résumé

MEDLINE est la base bibliographique de référence dans le domaine biomédical. Cette dernière connait une croissance explosive dans les dernières années. L'indexation de cette ample base avec le thésaurus MeSH devient de plus en plus une tâche difficile pour un staff qualifié de la US-NLM. La catégorisation de texte (CT) à base des algorithmes d'apprentissage, étudiée dans le contexte d'indexation des références de MEDLINE, est une façon efficace afin d'aider ce groupe d'expert.

Les algorithmes d'apprentissage supervisé, comme les SVM et la classification Naïve Bayes appliqués sur la représentation standard (sac de mots, ou en anglais : bag-of-words) affinée par des techniques de prétraitement montre des performances compétitives : une F-Mesure de 59.6% pour le classifieur NB, et F-Mesure de 58% pour les SVM avec des paramètres standards. Cependant, la représentation d'un texte peut engendrer un espace de dimension impressionnant entravant les classifieurs.

La sélection de descripteurs est une tâche populaire qui consiste à retrouver les mots représentatifs et éliminent ceux qui ne le sont pas. Nous avons appliqué la méthode de sélection chi-2 (chi-squared) avec les techniques SVM et NB. Cette méthode accomplis des bonnes performances, une F-Mesure de : 62.7% pour les SVM et 65.9% pour le classifieur NB.

Mots-clés: MEDLINE, chi-squared, SVM, Naïve Bayes, MeSH thesaurus, catégorisation de texte, Algorithmes d'apprentissage supervisé, prétraitement, sélection de descripteurs, sac de mots, Indexation.

Abstract

MEDLINE is the well-known database of bibliographic citations in biomedicine. It knows an explosive growth in recent years. The indexing of this large database using the MeSH thesaurus is performed by a relatively small group of highly qualified indexing staff at the US-NLM. However, their task is becoming difficult due to the ever increasing size of MEDLINE. Text Categorization based Machine learning – which is studied in the context of indexing MEDLINE citations – is an efficient way that could help this group of experts.

Supervised Machine Learning Algorithms, such as: SVM and the Naïve Bayes classifier applied on a refined bagof-words representation with preprocessing methods shows competitive performances: an F-Measure of 59.6% for a NB classifier and an F-Measure of 58% for SVM with standards parameters. However, text representation may produce a large number of features, which can hinder the performance of learning algorithms.

Features Selection is a popular task to find representative words and remove unimportant ones. We applied one of those approaches called chi-squared with the SVM and the NB. We achieved better performance: an F-Measure of 62.7% for SVM and 65.9% for NB.

Keywords: MEDLINE, MeSH thesaurus, chi-squared, SVM, Naïve Bayes, Text Categorization, Machine Learning, Features Selection, preprocessing, bag-of-words, Indexing.

ملخص

تعتبر MEDLINE قاعدة البيانات المرجعية للنصوص المتعلقة بالمجال البيوطبي. هاته القاعدة – التي هي بمثابة مكتبة إلكترونية يتم الوصول إليها عن طريق شبكة الأنترنت – تقوم بتخزين النصوص على وسائط الإعلام الإلكترونية وتعرف مؤخرا إرتفاعا هائلا في عدد النصوص المضافة لها. يتكفل مجموعة من الخبراء بتصنيف النصوص فيها عن طريق استخدام تجميعية تحتوي عددا من التعبيرات القياسية المتفق عليها تعرف بالـ MeSH. أصبحت عملية التصنيف للخبراء صعبة نظرا لعدد النصوص المرتفعة، فكان اللجوء إلى تكنولجيا الذكاء الإصطناعي أمر ضروري لتسهيل التصنيف و تقديم الدعم لهؤلاء الخبراء.

خوارزميات الذكاء الإصطناعي مثل: SVM و Naïve Bayes يمكن استخدامهما لتصنيف النصوص بعد تغيير هيئة النص إلى مجموعة من الكلمات الأكثر تمثيلا له و هذا بعد تصفيته. قدمت هاته الطريقة نتائج لا بأس بها. إنّ من المشاكل التي تعترض تمثيل النص، هو العدد المرتفع للخاصيات الممثلة للنص. لذا أصبح من الضروري القيام بعملية اختيار للخاصيات التمثيلية.

اختيار الخاصيات التمثيلية هي مهمة ايجاد الخاصيات الأكثر تمثيلا للنص و الاحتفاظ بها. استخدمنا طريقة من الطرق في اختيار الخاصيات تعرف بـ Chi-2 و قد ساعدت في تحسين النتائج المحصل عليها بعد التقييم.

الكلمات المفتاحية: التصنيف، MEDLINE، Naïve Bayes، اختيار الخاصيات التمثيلية، الذكاء الإصطناعي، SVM، MeSH.