Recherche de documents par apprentissage artificiel pour une méta-analyse sur les émissions d'un gaz à effet de serre

A. Cornuéjols[1] A. Duroy [1] Ch. Loyce[2] D. Makowski [3] Ch. Martin[1] A. Philibert[3]

[1] AgroParisTech, département MMIP et INRA UMR-518 16, rue Claude Bernard, F-75231 Paris Cedex 5 (France)
[2] AgroParisTech, Dept. SIAFEE, UMR Agronomie INRA / AgroParisTech 78850 Thiverval-Grignon, France
[3] UMR 211 INRA AgroParisTech 78850 Thiverval-Grignon, France {antoine.cornuejols,christine.martin}@agroparistech.fr

Résumé

La méta-analyse est un outil puissant pour réaliser des synthèses quantitatives d'articles scientifiques traitant d'un sujet commun, par exemple l'estimation des émissions de gaz à effet de serre. La qualité d'une méta-analyse dépend de plusieurs facteurs, et notamment de la qualité de la recherche bibliographique. Celle-ci génère fréquemment une liste de plusieurs centaines voire plusieurs milliers d'articles. En général, seule une petite partie de ces articles inclut des données exploitables. Dans ce cas, le tri manuel des articles et l'extraction des données constituent des étapes longues (souvent plusieurs mois) et fastidieuses. Ce papier décrit une technique permettant d'automatiser le tri des articles par apprentissage artificiel dans le but de réaliser des méta-analyses portant par exemple sur des problématiques environnementales.

Mots Clef

Apprentissage Artificiel, méta-analyse, réchauffement climatique

Abstract

Meta-analysis is a powerful tool to perform quantitative syntheses from scientific articles related to some given question such as what is the amount of nitrous oxide coming from fertilizers that ends up as a greenhouse effect agent ? The quality of a meta-analysis is greatly dependent upon the relevance and completeness of the prior bibliographical search. It is common that a first search yields hundreds or thousands of potentially relevant papers. In general, however, only a small number of these articles turn out to be of interest. Their filtering out mobilizes human experts for long periods of time and the process is subject to many and difficult to control defects. This paper describes an automatic process based on machine learning techniques that help to screen the vast amount of literature on a subject.

Keywords

Machine Learning, Meta-Analysis, Information Retrieval, Global Warming.

1 Introduction

Prédire l'évolution et les risques liés à des systèmes mettant en jeu des interactions complexes est un défi. Les enjeux sont particulièrement évidents dans l'estimation du changement climatique, des risques sanitaires, de l'utilisation des ressources naturelles ou des politiques agricoles.
Étant donnée l'importance croissante de ces questions pour nos sociétés, il est fréquent qu'existe déjà un (très) grand nombre d'études. Celles-ci sont cependant le plus souvent parcellaires et locales aussi bien d'un point de vue spatial que temporel. Prises séparément, il est difficile d'en extraire un scénario précis, ou même une estimation fiable de quelques paramètres clés, comme, par exemple, l'estimation des émissions de gaz à effet de serre (N_2O en particulier) par les sols agricoles. Nous avons besoin de solutions permettant de faciliter le tri des sources scientifiques, d'augmenter la fiabilité de ce tri et de rendre les procédures utilisées pour les exploiter plus transparentes.
La méta-analyse est une méthode permettant de réaliser des revues quantitatives de la littérature scientifique. Elle consiste à réaliser une analyse statistique d'un jeu de données constitué d'un nombre plus ou moins grand d'expérimentations traitant du même sujet. Cependant, la qualité d'une méta-analyse est en grande partie dépendante de la pertinence et de la complétude du corpus d'études rassemblé au préalable.
Ce papier présente une démarche générique pour opérer une sélection des articles pertinents. Elle est illustrée pour une méta-analyse cherchant à estimer la part des doses d'engrais azotés apportés sur les sols agricoles qui se retrouve convertie en N_2O, un puissant gaz à effet de serre [1].

Principe de l'approche. Étant donnée la masse des documents désormais disponibles dans l'univers numérique, ainsi que la complexité de chaque document par rapport à l'information recherchée se rapportant à un paramètre, une démarche raisonnable consiste à recourir à une séquence d'opérations destinées à raffiner la recherche.

1. **Sélection des articles par mots-clés**. Soit $\mathcal{S}_{mc}$, l'ensemble des articles retenus ainsi.

1. Son pouvoir de réchauffement global sur 100 ans est 310 fois plus élevé qu'une masse équivalente de CO_2 [4].

2. **Apprentissage portant sur les abstracts** des documents à trier (trois phases) :
   – *Apprentissage d'une redescription des abstracts* qui diminue le nombre de descripteurs afin de rendre la tâche d'apprentissage réalisable
   – *Étiquetage par l'expert* de N articles parmi ceux retenus à l'étape précédente ($\mathcal{S}_{mc}$), de telle manière qu'environ $N/2$ de ces articles soient étiquetés comme valides (contenant l'information recherchée pour la méta-analyse), et que les autres soient étiquetés négativement [2].
   – *Apprentissage supervisé d'une règle de décision* permettant, sur la base des abstracts redécrits, de distinguer les articles probablement pertinents de ceux qui ne le sont probablement pas.
3. **Apprentissage** pour chercher l'information **dans le contenu des articles** retenus après filtrage sur les abstracts.

La suite de ce papier est consacrée à la deuxième étape (section 2).
Le souci récurrent à chaque étape est d'éliminer au maximum les articles non pertinents (les *faux positifs*) tout en conservant cependant tous (ou le plus possible) les articles pertinents (*vrais positifs*). Or ces deux objectifs sont contradictoires.

2 Traitement des abstracts

2.1 Du texte à une représentation vectorielle

La source d'information la plus riche et la plus discriminante étant les abstracts, c'est sur eux que doit prioritairement porter la méthode de filtrage. Cependant, si les experts sont incapables de fournir une telle règle, ils sont néanmoins aptes à étiqueter les abstracts (et les documents associés) en deux classes : positif et négatif, ou « pertinent » et « non pertinent ». Il devient dès lors tentant de recourir à des techniques d'apprentissage artificiel supervisé pour essayer de découvrir de manière automatique une règle de décision dont la performance en reconnaissance se rapproche si possible de celle des experts.
Il n'existe pas encore de méthode d'apprentissage supervisé capable de s'appliquer à des textes quelconques. Ceux-ci sont en effet à la fois semi-structurés, avec une ligne argumentative non contrainte et une longueur variable, tout en impliquant potentiellement un vocabulaire considérable, de l'ordre de plusieurs dizaines de milliers de mots (en comptant les terminaisons et variations diverses). Il n'est donc pas possible de fournir en entrée à des systèmes d'apprentissage classiques des textes sous leur forme brute.
C'est pourquoi a été imaginée une approche radicale consistant à « projeter » un texte sur l'espace des mots du vocabulaire du domaine d'intérêt. Dans cette approche en « sac de mots » (*bag-of-words*), la structure du texte est ignorée et ne subsiste que l'information sur la présence ou non d'un mot dans le texte, parfois avec une information aussi sur la fréquence. Une difficulté est alors que la taille du vocabulaire implique des représentations dont la taille est aisément de l'ordre de plusieurs milliers, et donc souvent d'un ou deux ordres de grandeur supérieur au nombre d'exemples d'appprentissage disponible.
Il faut donc réduire très notoirement la dimension de l'espace de représentation, tout en conservant autant que possible l'information utile. Les grandes étapes de prétraitement de textes classique sont les suivantes.

1. *Étape de reconnaissance des mots*.
2. *Le retrait des mots standards (stopwords)*.
3. *La lemmatisation*.
4. *Le stemming*.

2.2 Sélection d'attributs

Si les traitements purement textuels permettent un gain appréciable en terme de réduction de taille de vocabulaire (souvent une division par 3 ou 4 du nombre de « mots »), cela n'est généralement pas suffisant pour permettre la mise en œuvre de techniques d'apprentissage. Il faut donc compléter ces traitements par des analyses fondées sur des mesures statistiques.
L'*analyse tf-idf* est l'une d'elle. Elle consiste à retirer les termes ne permettant pas, au vu d'une analyse statistique, de discriminer les textes au sein d'une collection de documents. Mais d'autres analyses sont spécifiquement dédiées à la découverte des attributs de description qui sont les plus corrélés à la distinction en classes selon certains critères.
Parmi les méthodes de sélection d'attributs (*feature selection*), on distingue généralement les méthodes *de filtre* (*filter methods*) qui évaluent la pertinence de chaque attribut ou descripteur indépendamment, et sans tenir compte de la méthode de classification supervisée utilisée en aval. Les *méthodes de filtre* calculent un score à chaque attribut de description indépendamment. Nous avons utilisée une méthode basée sur une mesure d'entropie : la *mdl-entropie* [3].

2.3 Recodage structuré et supervisé

Cependant, au lieu de redécrire les données grâce à des techniques de sélection décrites dans la section précédente, il peut être judicieux de procéder à un changement de représentation permettant si possible de rendre compte, au moins en partie, de la structure des textes étudiés.
Un moyen simple pour commencer à tenir compte de cette structure est de considérer non pas des mots isolés mais des ensembles de mots (*motifs*) particuliers pour décrire les documents. Une technique classique est de calculer les n-grams fréquents. On généralise les n-grams par les « motifs fréquents » qui ne sont pas tenus d'être composés d'éléments contigus.
Soit A l'ensemble des descripteurs a_i issus des étapes présentées en section 2.1. Un motif m est alors défini comme un sous-ensemble de A et un motif m est dit présent dans un document si tous les descripteurs qui le composent y figurent au moins une fois. Étant donné un motif m et une

2. Pour favoriser l'apprentissage, il est généralement préférable que les classes soient équilibrées dans l'échantillon d'apprentissage.

fréquence minimale f_{min} servant de seuil de décision, on dit que m est *fréquent* si et seulement si sa fréquence parmi les documents concernés est supérieure à f_{min}. L'algorithme Apriori [2], ou l'une de ses variantes, permet de calculer l'ensemble des motifs fréquents pour tout ensemble de documents fournis en entrée.

En sélectionnant les « motifs fermés fréquents » indépendamment sur chaque classe de documents que l'on souhaite pouvoir identifier (pertinent et non pertinent) on obtient une nouvelle « base » de descripteurs de dimension d'autant plus faible que la table initiale est creuse et que le seuil choisi est élevé. Il faut néanmoins veiller à ce que chaque document soit décrit par au moins un descripteur. Il y a donc un compromis à trouver au niveau de seuil de sélection pour, d'une part, réduire au maximum le nombre de descripteurs et, d'autre part, en conserver suffisamment pour que chaque exemple disponible initialement reste utilisable pour la suite des traitements.

2.4 Recherche d'une règle de discrimination par apprentissage supervisé

Après l'étape de sélection ou de recodage des attributs de description des documents, ne doit subsister qu'un petit vocabulaire de termes à l'aide duquel sont décrits les documents. Il est préférable que la taille du vocabulaire ainsi déterminée soit inférieure au nombre de documents de l'ensemble d'apprentissage. Tout document, d'apprentissage ou à classer, est alors codé soit en signalant pour chaque attribut si il est présent ou non dans le document : *codage booléen*, soit en reportant le nombre d'occurrences de chaque attribut dans le document : *codage par fréquence*.

Une fois le codage obtenu, supposé aussi concis et néanmoins informatif que souhaitable, il devient possible de recourir à des techniques d'apprentissage artificiel supervisé pour rechercher des règles de décision permettant de distinguer les documents pertinents de ceux qui ne le sont pas.

Ainsi, à partir d'un corpus d'entraînement $\mathcal{S} = \{(d_j, y_j)\}_{1 \leq j \leq m}$ de m documents d_j étiquetés dans une classe y_j par un expert, un algorithme d'apprentissage supervisé induit une règle de décision $h(\cdot)$ qui à partir d'un document d lui associe une étiquette $y = h(d)$.

De nombreuses *techniques d'apprentissage supervisé* existent (voir [2]). Les plus classiques incluent : l'apprentissage *par plus proches voisins*, l'apprentissage *bayésien naïf*, l'apprentissage par *réseaux de neurones artificiels*, l'apprentissage par *arbres de décision*, les méthodes à noyaux dont les SVM (*Séparateurs à Vastes Marges*), et des méthodes de méta-apprentissage comme le *boosting* ou le *bagging*. Ces algorithmes ont des propriétés propres et présentent des avantages et inconvénients les rendant plus ou moins adaptés à une utilisation donnée (voir le tableau 2.4). De plus, le choix parmi ces méthodes doit aussi prendre en compte les critères importants pour l'application visée, à savoir par exemple :

– La robustesse au bruit dans les données (données mal décrites, présence de données aberrantes, ...)

	Règle de décision (hypothèse) non linéaire	Transparence, intelligibilité	Espace de grande dimension	Besoin d'a priori fort	Facilité du paramétrage de la méthode	Généralisation / insensibilité au bruit	Multi-classe
k-ppv	√	-	-	-	√	-	√
Naïve Bayes	√	-	-	-	√	-	√
Arbre de décis.	√	√	√	√	√	-	√
Réseau de neur.	√	-	-	-	-	√	√
SVM	√	-	√	√	√	√	-

TABLE 1 – Comparaison de certaines propriétés de méthodes d'apprentissage supervisé classiques.

– La capacité à traiter des données numériques et/ou de nature symbolique
– La capacité à traiter des données décrites dans un espace de grande dimension
– L'intelligibilité des règles de décision produites
– La complexité calculatoire (temps calcul et espace mémoire nécessaires)

3 Application : Estimation des émissions de N_2O par les sols agricoles

Dans ses rapports successifs sur le réchauffement climatique, et concernant plus particulièrement les calculs d'émission de gaz à effet de serre dans l'atmosphère, le Groupe d'Experts Intergouvernemental sur l'Évolution du Climat (GIEC) a d'abord considéré que 1,25% des doses d'engrais azoté appliquées aux sols cultivés est converti en N_2O [1]. Plus récemment, le GIEC a revu ce taux d'émission à la baisse (1%) suite à une nouvelle méta-analyse [5]. Or l'impact de cet écart d'estimation est considérable sur les scénarios de réchauffement climatique. Il est donc crucial de parvenir à l'estimation la plus fiable et la plus précise possible, et cela dépend d'abord de la qualité du corpus de sources scientifiques rassemblé. Ce papier étudie l'estimation des émissions de N_2O liées aux quantités d'engrais azotés apportées aux cultures.

3.1 Corpus

La production mondiale d'articles scientifiques est en croissance vertigineuse depuis de nombreuses années (∼470 000 en 1988 contre ∼990 000 en 2008 selon le rapport de l'UNESCO sur la science 2010). La plus grande partie de cette production est désormais potentiellement accessible sur Internet et il existe différents sites Internet as-

surant le référencement de ces publications. Parmi ceux-ci, l'un des acteurs majeurs du marché est le *Web of Science* (Thomson Reuters) qui regroupe de multiples domaines scientifiques au sein de ses bases de données.
La recherche de documents sur le *Web of Science* s'effectue en précisant une thématique et des mots-clés. Par exemple, la requête avec les mots-clés « `nitrous oxide` » et « `crop` » retourne environ 1400 articles.
Partant des 1400 articles sélectionnés grâce à la première requête par mots-clés sur le *Web of Science*, les experts ont étudié les abstracts de 44 articles, étiquetant 20 d'entre eux comme « non-pertinent » et 24 comme « pertinent » formant ainsi le corpus utilisable pour l'entraînement et l'évaluation de notre méthode à chacune de ces étapes : d'abord filtrage sur la base des abstracts puis sur la base des contenus par l'analyse des légendes des figures et des tableaux.
Les fichiers PDF correspondant à ces documents représentent un spectre assez large de documents types liés à la thématique de recherche sur les émissions de gaz à effets de serre. Ils sont issus d'une dizaine de revues scientifiques différentes parmi les plus représentées dans le domaine de la biologie (Elsevier, etc.). Leur examen montre que l'agencement des parties de document ainsi que les formalismes utilisés sont très variés selon les publications, ce qui représente un défi supplémentaire pour la mise au point d'une méthode générique de récupération du contenu d'un document sous format PDF [3].

3.2 Mise en œuvre de la méthode

Le retrait de mots-standard (*stop-words*), la lemmatisation et le stemming ont été réalisés sur ces 44 abstracts, ramenant le nombre de termes de description à 1413.
Afin d'examiner comment varie la richesse du vocabulaire, ainsi que sa redondance, dans les abstracts, nous avons tracé la courbe du nombre t de termes différents en fonction du nombre n d'abstracts étudiés. Cette courbe a été obtenue par tirages aléatoires répétés de k abstracts parmi les 44 étiquetés avec calcul de la moyenne et de l'écart-type (voir Figure 1). La courbe, d'allure concave, montre que le nombre de termes croît sub-linéairement avec le nombre d'abstracts. Cependant, même s'il y a donc redondance, les mêmes termes étant utilisés dans plusieurs abstracts, il y a introduction significative de nouveaux termes (environ 30) encore au 42ème abstract.
Une fois cette première étape de réduction de la dimension des abstracts effectuée, les expériences réalisées avaient plusieurs objectifs :

1. *Déterminer le nombre optimal d'abstracts à étiqueter* pour obtenir une bonne performance de discrimination entre articles prometteurs et articles à rejeter. Étant donné le coût en temps expert de l'étiquetage des articles et la tendance des experts humains à augmenter leur taux d'erreur d'étiquetage avec le nombre à traiter, il est souhaitable de limiter au maximum la taille requise du corpus d'entraînement.
2. Mesurer *l'effet des changements de représentation* en comparant trois situations :
   – performance *sans changement de représentation* (pas de réduction au delà de la première étape)
   – performance en utilisant la méthode de *codage par mdl-entropie*
   – performance en utilisant le *recodage par motifs fréquents*
3. Comparer les performances sur ce problème de *diverses méthodes de classification supervisée*. Nous avons comparé trois méthodes ayant des fondements différents :
   – Les arbres de décision
   – Le classifieur naïf de Bayes
   – Les séparateurs à vastes marges (SVM), ici avec un noyau à base radiale

3. Faute de place, nous ne détaillons pas la méthode mise au point pour cette extraction, qui a cependant représenté une bonne moitié de notre travail et dont la qualité est cruciale pour le reste de l'analyse.

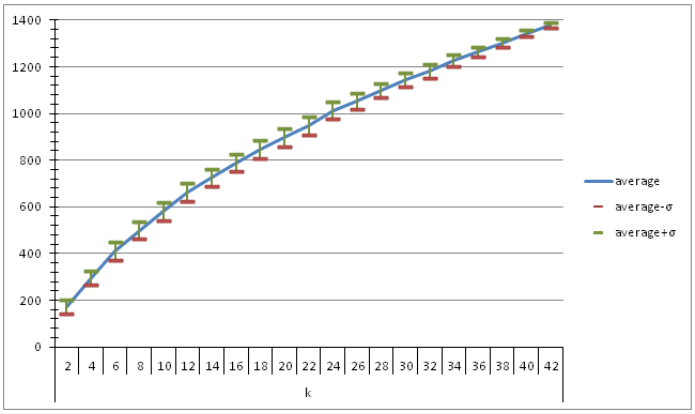

FIGURE 1 – Nombre moyen d'attributs n (et écart-type) en fonction du nombre k d'abstracts considéré pour 100 tirages aléatoires

3.3 Résultats

Sélection des attributs et recodage par motifs
Aussi bien les méthodes de recodage par sélection d'attributs (mdl-entropie) que les méthodes par construction de nouveaux descripteurs (motifs fréquents) dépendent du corpus d'entraînement. Il était donc important d'analyser l'impact de la taille du corpus d'entraînement (le nombre d'abstracts) sur le recodage obtenu.
La méthode de sélection de descripteurs par mdl-entropie apparaît satisfaisante à trois titres. Elle conduit à un codage de petite dimension, celui-ci est stable puisque les mêmes descripteurs sont sélectionnés lors des tirages, enfin les descripteurs sélectionnés semblent pertinents comme l'indique la présence de termes liés à des unités de mesure (« kg », « ha »), qui sont effectivement des cibles de la méta-analyse.
Le recodage par *motifs fréquents* présente grossièrement les mêmes caractéristiques : stabilité des motifs, pertinence

apparente, mais les nouveaux descripteurs ressortent en plus grand nombre, nombre qui varie fortement en fonction du taux de couverture utilisé.
Il est notable que 80 à 90 % des motifs retenus proviennent des documents de la classe « pertinent » tandis que seulement 10 à 20 % sont issus des documents non pertinents avec une intersection assez faible entre ces deux ensembles. Cela démontre que les documents « non pertinent » exploitent un vocabulaire plus diversifié et sont plus hétérogènes que les documents « pertinent ».

Comparaison des méthodes de classification supervisée et nombre optimal de documents à étiqueter
Une méthode de filtrage de documents pertinents sera performante et donc utilisée si elle demande aux experts un effort d'étiquetage faible, c'est-à-dire portant sur peu de documents, et si cela permet néanmoins d'obtenir une règle de classification efficace, ce qui veut dire en premier lieu maximisant le rappel (peu de faux négatifs) sans que cela soit trop au détriment de la précision (aussi peu que possible de faux positifs).
Nous avons comparé les performances des trois classifieurs : Séparateurs à Vastes Marges (SVM) (noyau gaussien avec $\sigma = 0.5$), arbres de décision (C4.5) et Naïve Bayes (NB) en termes de rappel et précision en faisant varier le codage utilisé.
Sans recodage (après lemmatisation et stemming), les performances ont été obtenues par répétition de tirages de n abstracts d'entraînement (avec équirépartition de « pertinent » et « non pertinent » dans les ensembles d'apprentissage et de validation). On a fait varier n de 10 à 42 (auquel cas il reste 2 abstracts de validation : un « pertinent » et un « non pertinent »).
Avec recodage, la procédure est la même. Pour chaque valeur de n considérée, on a répété 100 tirages aléatoires de n abstracts parmi les 44 étiquetés par les experts. Pour chaque tirage, le codage est calculé (parfois avec plusieurs valeurs de support pour le codage par motifs fréquents), puis utilisé pour coder les exemples d'entraînement et de test. La performance en rappel et précision est alors obtenue pour chaque valeur de n par calcul de la moyenne sur les 100 répétitions.
Les meilleurs résultats ont été obtenus avec les SVM.

Sans recodage		SVM
Rappel	(n = 20)	68%
	(n = 40)	80%
Précision	(n = 20)	80%
	(n = 40)	90%

Avec le recodage issue de la méthode mdl-entropie :

mdl-entropie		SVM	C4.5	NB
Rappel	(n = 20)	76%	**80**%	70%
	(n = 40)	**90**%	**90**%	86%
Précision	(n = 20)	**72**%	**72**%	68%
	(n = 40)	**75**%	74%	74%

Avec le recodage par motifs fréquents :

motifs fréquents		SVM	C4.5	NB
Rappel	(n = 20)	87%	**96**%	**96**%
	(n = 40)	96%	**100**%	98%
Précision	(n = 20)	**74**%	v62%	54%
	(n = 40)	v**83**%	65%	55%

De ces tableaux, il est possible de tirer les conclusions suivantes :

1. Le recodage améliore sensiblement les performances.
2. Le recodage par motifs fréquents est le plus performant, alors même qu'il implique un plus grand nombre de descripteurs que le codage par mdl-entropie. Cela peut être dû à ce que ce codage capture mieux les structures du texte. Plus probablement, cela est dû au fait que le codage lui-même distingue déjà beaucoup les deux classes de textes.
3. Il est possible d'obtenir un rappel de quasiment 100% par recodage par motifs fréquents (support 0.5) et utilisation des arbres de décision, avec une précision qui est alors de 65%.
4. Il est possible d'obtenir une bien meilleure précision en utilisant les SVM (83%), mais c'est au prix d'un rappel de 96% ce qui peut être rédhibitoire dans une tâche de méta-analyse qui ne veut pas laisser passer de documents pertinents.
5. On obtient déjà d'excellents résultats à partir de 20 abstracts d'entraînement. L'amélioration des performances étant faible ensuite.

La méthode qui maximise le rappel est donc celle qui repose sur un recodage par motifs fréquents et une classification supervisée par arbres de décision, avec apprentissage sur au moins une vingtaine d'articles.
Dans un contexte dans lequel on peut estimer à 10% environ la proportion des articles pertinents (soit environ 150 sur 1440 articles sélectionnés sur la base des mots clés), trois stratégies sont intéressantes à comparer pour évaluer l'effort demandé aux experts. Toutes les trois reposant sur un recodage par motifs fréquents, le plus performant.

1. Utilisation d'un *corpus d'entraînement de 40 abstracts* (20 « pertinent » et 20 « non pertinent ») et apprentissage d'*arbres de décision*. Le rappel est alors de 100% et la précision de 65%. La matrice de confusion sur les 1240 documents ne faisant pas partie du corpus d'entraînement est alors proche de :

Réelle / *Estimée*	+ (P)	− (N)
+	**VP** = 124	FP = $124\frac{0.35}{0.65} \approx 68$
−	FN = 0	**VN** = 1048

Pour obtenir 40 abstracts d'entraînement, dont 20 positifs (« pertinents »), les experts ont du préalablement examiné environ 200 documents. Après filtrage par recodage et apprentissage d'une règle de décision, il en ont encore 124 + 68 soit environ 192 documents à examiner pour finalement découvrir les 144 documents pertinents. Au total, les experts auront étiqueté environ *390 documents*, soit environ 27% des 1440 documents initiaux.

2. Utilisation d'un *corpus d'entraînement de 20 abstracts* (10 « pertinent » et 10 « non pertinent ») et apprentissage d'*arbres de décision*. Le rappel est alors de 96% et la précision de 62%. La matrice de confusion est alors proche de :

Réelle / *Estimée*	+ (P)	− (N)
+	**VP** = 128	FP = $134\frac{0.38}{0.62} \approx 82$
−	FN = 6	**VN** = 1124

Avec cette stratégie, les experts doivent préalablement examiner environ 100 documents pour obtenir le corpus d'entraînement, puis encore 128 + 82 $\approx$ 210 documents, soit *au total 310 documents*, ou 22% des 1440 documents initiaux. Le prix à payer pour cet effort moindre que dans la stratégie précédente est de courir le risque de manquer 6 documents pertinents.

3. Utilisation d'un *corpus d'entraînement de 40 abstracts* (20 « pertinent » et 20 « non pertinent ») et apprentissage par Séparateurs à Vastes Marges. Le rappel est alors de 96% et la précision de 83%. La matrice de confusion est alors proche de :

Réelle / *Estimée*	+ (P)	− (N)
+	**VP** = 119	FP = $124\frac{0.17}{0.83} \approx 25$
−	FN = 5	**VN** = 1091

Avec cette stratégie, les experts doivent préalablement examiner environ 200 documents pour obtenir le corpus d'entraînement, puis encore 119 + 25 $\approx$ 144 documents, soit *au total 344 documents*, ou 24% des 1440 documents initiaux. Cette stratégie est à michemin des deux précédentes en termes d'effort demandé aux experts.

On voit que les trois stratégies permettent une économie très sensible du travail demandé aux experts en terme d'examen des documents sans conduire à un taux de faux négatifs important (moins de 4%).

4 Conclusion et perspectives

Étant donnée la globalité, tant spatiale que temporelle, des processus environnementaux, c'est une multitude d'études fragmentaires et partielles qu'il faut savoir prendre en compte. Seul un processus automatisé défini avec soin peut à la fois permettre d'en faire l'analyse et de garantir l'objectivité et la reproductibilité de celle-ci. Ce papier a présenté une méthode visant à aider les experts à faire le tri des articles scientifiques portant sur une question donnée pour n'avoir à analyser que les articles pertinents : seulement ceux-ci mais tous ceux-ci.

Le souci est de maximiser le rappel tout en ayant une bonne précision. De notre étude, il ressort que le recodage par motifs fréquents est le plus performant, couplé à un apprentissage par arbre de décision ou par Séparateurs à Vastes Marges (SVM).

La méthodologie décrite a été testée sur un problème d'estimation d'émission de gaz à effet de serre, le N_2O, par kilo d'engrais épandu sur les terres agricoles. Elle a permis de voir qu'à partir d'environ 10 articles étiquetés « pertinent » et 10 « non pertinent », il était possible d'apprendre une règle de décision portant sur les abstracts permettant d'éliminer environ 80% des articles potentiels (ici 1440) sans risque (quasiment) d'éliminer des faux négatifs.

Remerciements. Nous remercions le Conseil Scientifique d'AgroParisTech pour le soutien au projet ExtraEx dans le cadre de l'appel d'offre annuel en 2012 sur les projets scientifiques, ainsi que le soutien par le Réseau National des Systèmes Complexes lors de l'appel à Réseaux pour l'année 2013.

Références

[1] Bouwman, A. (1996). Direct emission of nitrous oxide from agricultural soils. *Nutrient cycling in agroecosystems 46*(1), 53–70.

[2] Cornuéjols, A. and L. Miclet (2010). *Apprentissage Artificiel. Concepts et algorithmes (2nd Ed.)*. Eyrolles.

[3] Kohavi, R. and M. Sahami (1996). Error-based and entropy-based discretization of continuous features. In *Proceedings of the second international conference on knowledge discovery and data mining*, pp. 114–119.

[4] Mosier, A., J. Duxbury, J. Freney, O. Heinemeyer, and K. Minami (1998). Assessing and mitigating N2O emissions from agricultural soils. *Climatic Change 40*(1), 7–38.

[5] Stehfest, E. and L. Bouwman (2006). N2O and no emission from agricultural fields and soils under natural vegetation : summarizing available measurement data and modeling of global annual emissions. *Nutrient Cycling in Agroecosystems 74*, 207–228.