Approche semi-supervisée pour la désambiguïsation des affiliations dans les bases de données bibliographiques

Pascal Cuxac, Jean-Charles Lamirel, Valérie Bonvallot

Approche semi-supervisée pour la désambiguïsation des affiliations dans des bases de données bibliographiques

Cuxac Pascal[1], Lamirel Jean-Charles[2], Bonvallot Valérie[1]

[1] INIST-CNRS, Vandoeuvre lès Nancy, France

pascal.cuxac@inis.fr ; valerie.bonvallot@inist.fr

[2] LORIA-Synalp, Vandoeuvre lès Nancy, France

jean-charles.lamirel@loria.fr

Résumé : La désambiguïsation d'entités nommées est un défi dans de nombreux domaines tels que la scientométrie, les réseaux sociaux, l'analyse des citations, le Web sémantique etc ... Les ambiguïtés peuvent provenir de fautes d'orthographe, d'erreurs typographiques ou d'OCR, d'abréviations, d'omissions ... Ainsi, la recherche de noms de personnes ou d'organisations est rendue difficile par la multiplicité des formes utilisées.

Cet article propose deux approches pour lever l'ambiguïté sur les affiliations des auteurs d'articles scientifiques dans des bases de données bibliographiques : la première, considère un corpus d'apprentissage, et utilise un modèle bayésien naïf ; la deuxième sans ressource d'apprentissage, une approche semi-supervisée, associant une méthode de clustering recouvrant et un apprentissage bayésien.

Les résultats sont encourageants et les méthodes sont déjà partiellement appliquées dans un pôle de veille scientifique. Cependant, cette approche a des limites : par exemple, on ne peut pas traiter efficacement des données très déséquilibrées, mais des solutions sont envisageables pour de futurs développements.

Mots-clés : Veille scientifique, Bases de données, Affiliations, Désambiguïsation, Classification, Clustering, Semi-supervisé.

Abstract : The disambiguation of named entities is a challenge in many fields such as scientometrics, social networks, record linkage, citation analysis, semantic web...etc. The names ambiguities can arise from misspelling, typographical or OCR mistakes, abbreviations, omissions...So the search of names of persons or of organization is difficult, a single name can appear in different forms.

This paper proposes two approaches to disambiguate on the affiliations of authors of scientific papers in bibliographic databases: the first way, considers that we have a training corpus, and uses a Naive Bayesian model. The second way assumes that we have not resource learning, and uses a semi-supervised approach, mixing soft-clustering and Bayesian learning.

The results are encouraging and are already partially applied in a scientific survey department. However, we aware that our approach may have limitations: we can't process efficiently highly unbalanced data but solutions are possible for future developments.

Keywords : Scientific survey, Databases, Affiliations, Desambiguation, Classification, Clustering, Semi-supervised.

I. INTRODUCTION

Dans les bases de données bibliographiques, les affiliations des auteurs sont d'une importance primordiale. Ainsi, elles permettent aux laboratoires ou instituts d'obtenir une visibilité nationale voir internationale. Nous ne pouvons pas aborder la question des affiliations sans parler du «classement de Shanghai»[1], qui vise à évaluer les universités. Notre but ici n'est pas de nourrir la polémique [1][2], mais de souligner que le traitement des affiliations joue un rôle important dans le calcul de la «performance» des universités.

Moed [3] rapporte quelques problèmes liés aux noms d'auteurs et aux institutions : « *Les auteurs de la même institution, ou du même département, n'indiquent pas leur affiliation institutionnelle de la même manière* ». Selon le pays, il n'est pas toujours évident de savoir comment nommer un laboratoire à l'égard de ses autorités de tutelle. L'affiliation est également une information importante pour désambiguïser les noms des auteurs de bases de données bibliographiques. À cet égard, Wang note : «alors que le volume d'informations disponible augmente, les problèmes causés par des fautes d'orthographe, des orthographes différentes et des changements de nom ou d'affiliation s'aggravent» [4].

Une normalisation des données dans les bases de données bibliographiques est nécessaire pour mener des études infométriques, mais ce n'est pas une tâche triviale : la pratique,

[1] http://www.shanghairanking.co

intentionnelle ou non, d'omettre les affiliations institutionnelles, ou de donner des informations incomplètes ou erronées n'est pas rare [5].

Cet article propose une approche basée sur la méthode d'apprentissage bayésien naïf et sur un clustering recouvrant.

Ce document est structuré comme suit : la section 2 fait un rapide tour d'horizon des travaux. La section 3 décrit notre approche tout d'abord avec la méthode d'apprentissage supervisé puis avec la méthode semi-supervisé. Les résultats expérimentaux se trouvent à la section 4. La section 5 conclut et discute de travaux futurs.

II. ETAT DE L'ART ET DISCUSSION

Dans le cadre d'analyses bibliométriques, des statistiques sont produites à partir des affiliations d'auteurs, par laboratoires ainsi que par instituts ou universités. Cependant, ces analyses sont souvent confrontées à des problèmes de variabilité importante et d'hétérogénéité : un même laboratoire peut ainsi apparaître de plusieurs manières différentes si les auteurs utilisent différentes abréviations, des mots incomplets ou mal orthographiés (fautes de frappe, d'orthographe ...). En outre, certaines universités peuvent avoir plusieurs noms (par exemple Université Pierre et Marie Curie = Université Paris VI). Ce problème est connu depuis longtemps mais persiste encore de nos jours. Dans les années 1990, De Bruin et al. [6] pointent le problème de la variabilité des adresses d'auteurs dans les bases de données telles que SCI[2] (Science Citation Index). Ils mettent en évidence les cas de pays comme l'Allemagne ou la France, où l'hétérogénéité des données est particulièrement importante. Zitt [7] met l'accent sur l'importance de la normalisation des données (noms d'auteurs, affiliations) en portant une attention particulière aux pays comme la France où les affiliations multiples sont monnaie courante (par exemple, un laboratoire peut avoir une affiliation universitaire et une affiliation CNRS). Pour beaucoup d'analyses bibliométriques, l'unification des adresses institutionnelles est une tâche essentielle, souvent fastidieuse, à réaliser préalablement à toute étude [8][9].

Pour résoudre le problème, De Bruin et al. [6] proposent de traiter séparément tous les mots appartenant à des affiliations et d'utiliser, dans une deuxième étape, une stratégie de classification permettant d'unifier toutes les variations possibles des mots. Dans un ouvrage ultérieur [10], les mêmes auteurs utilisent un «regroupement par simple lien» pour délimiter les différents domaines de la science sur la base d'affiliations. French et al. [11] fournissent un fichier d'autorité après une étape de nettoyage (nom du pays, codes postaux, états, développement des abréviations, acronymes ...) et utilisent ensuite une classification fondée sur une « distance d'édition ». Des approches récentes abordent le problème par la seule utilisation de méthodes TALN, comme dans [12].

[2] http://www.webofknowledge.com/

Les termes « nettoyage », « normalisation », « désambiguïsation », « homogénéisation » et également « résolution d'entités » sont utilisés pour désigner les tâches de transformation de données de base en données propres ou normalisées pour l'alimentation de bases de données, la mise en relation avec d'autres ensembles de données ou encore le calcul d'indices statistiques (analyse bibliométrique par exemple). Si, comme nous l'avons vu, ces problèmes sont essentiels à la bibliométrie, ils sont aussi récurrents dans de nombreux autres domaines où l'hétérogénéité des données est un problème important. Cela peut être dans un fichier ou une base de données, mais aussi lorsque l'on combine des informations provenant de sources hétérogènes (par exemple couplage d'enregistrements ou « record linkage »). Erhard Rahm [13] classe les problèmes de qualité des données rencontrés dans les tâches de nettoyage (fig.1). Dans notre cas, nous pouvons assimiler le problème « multi-source » à une (ou plusieurs) base de données bibliographiques signalant des articles de journaux de différents éditeurs.

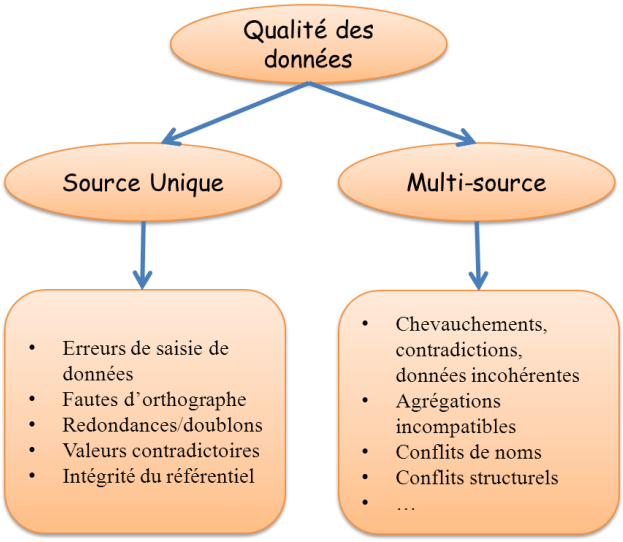

Fig. 1: Qualité des données lors d'une tâche de nettoyage des données (d'après [22])

L'approche présentée par Fellegi et Sunter [14] est une référence largement utilisée dans le couplage d'enregistrements pour identifier les mêmes entités dans des fichiers de données différents. Elle s'appuie sur le calcul de scores de similarité entre deux enregistrements. Des généralisations de cette méthode ont été récemment proposées [15][16]. Ventura [17] couple la désambiguïsation à des algorithmes de couplage d'enregistrements, en utilisant des forêts aléatoires, et applique cette méthodologie à une étude de cas sur des inventeurs de brevets dans le domaine de l'optoélectronique. Dans ce contexte de couplage d'enregistrements, Churches [18] montre que les modèles de Markov cachés probabilistes pour le prétraitement des données (noms et adresses) donnent des résultats précis avec des données complexes telles que les adresses résidentielles.

Lorsque des données d'apprentissage sont disponibles un grand nombre de ces études utilisent des métriques pour

mesurer les similarités entre ces données, comme par exemple Jaccard, soft-TF-IDF et surtout la distance d'édition [19]. Dans son document de synthèse, Bilenko compare la performance de différentes méthodes et conclut que la distance affine d'édition peut surpasser les autres avec des techniques EM [20].

Cependant, des approches probabilistes ont été proposées par plusieurs auteurs comme par exemple Carayol [21], qui propose une approche bayésienne pour traiter le problème de « qui est qui » dans les brevets européens. Il se pose la question de la transitivité que nous discuterons dans la conclusion. A l'inverse, certains auteurs proposent des approches basées uniquement sur des algorithmes non supervisés. C'est le cas de Niu [22] qui présente une nouvelle méthode pour la désambiguïsation d'entités à l'aide d'informations textuelles et des relations entre objets pour évaluer la similarité. Les entités sont des noms d'auteurs, et la relation entre objets est assimilée au réseau des co-auteurs.

La nouvelle méthodologie développée par Ashwani pour l'unification des noms d'auteurs utilise la fouille du Web pour obtenir les noms et trouver des pages de publications [23].

Comme on le voit, les applications sont nombreuses, que ce soit dans des bases de données bibliographiques, dans les entrepôts de données, dans les cas de couplage d'enregistrements multiples, dans le web sémantique, ou encore dans les bibliothèques numériques comme indiqué dans [24]. Pour conclure citons les actions de normalisation menées par l'intermédiaire de l' « International Standard Name Identifier » (ISNI) et le « Virtual International Authority File » (VIAF). Le but de ISNI[3] est d'identifier des identités publiques internationales d'individus ou de communautés et de fournir des outils pour la désambiguïsation. VIAF[4] est un projet de recherche de l'OCLC[5] (Online Computer Library Center), qui vise à harmoniser les listes d'autorité pour former une base de données de référence internationale.

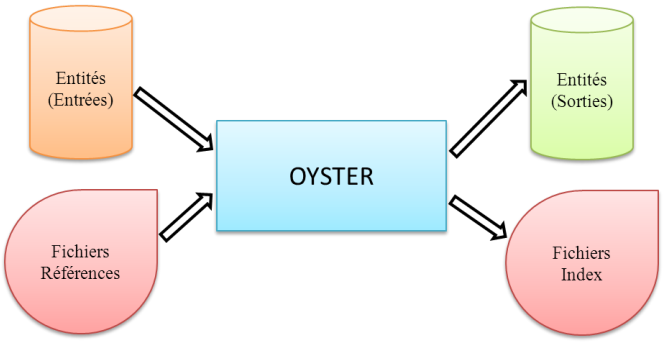

Fig. 2: Principe de fonctionnement d'Oyster (d'après [26])

Pour terminer n'oublions pas de mentionner le logiciel OYSTER[6] (Open sYSTem Entity Resolution). Ce logiciel open-source développé par John R. Talburt[7] au centre de recherche ERIQ (Entity Resolution and Information Quality, Université de l'Arkansas) est un système de résolution des entités utilisant des scripts XML [25] et fonctionnant sur le principe du schéma suivant (fig.2)

III. NOTRE APPROCHE

Nous présentons deux approches pour la désambiguïsation des affiliations : en premier lieu, une méthode d'apprentissage supervisé s'appuyant sur un corpus de références analysées manuellement, et dans une autre étape, une approche semi-supervisée dont le but est de s'affranchir du corpus d'apprentissage pour les cas où il n'y a aucune donnée de validation disponible.

A. Approche par apprentissage supervisé

Les méthodes d'apprentissage supervisé permettent de produire des règles à partir d'un corpus d'apprentissage, généralisant ce qui a pu être appris aux données inconnues. Dans la littérature, il existe de nombreuses méthodes telles que SVM, Rocchio, K-NN, Naive Bayes, HMM, arbres de décision... Notre approche supervisée est basée sur un algorithme Bayésien Naïf (NB).

Soit C un ensemble de classes, $C = \{c_1, c_2, \ldots, c_k\}$, le problème consiste à attribuer à une affiliation, une de ces catégories. En utilisant un ensemble N d'affiliations étiquetées $\{(a_i, c_i),\ 1 \leq i \leq N\}$, on construit une fonction de classification $\mathcal{F}$: A →C avec A = ensemble de toutes les affiliations.

La formule de Bayes pour une affiliation donnée permet de calculer sa probabilité d'appartenance à une classe particulière *c*:

$$P(c \mid a) = \frac{P(a|c) * P(c)}{P(a)}$$

avec :

$P(c \mid a)$ = probabilité de *c* étant donné *a*,
$P(a|c)$ = probabilité de *a* étant donné *c*,
$P(a)$ $P(c)$ = probabilité respectivement de *a* et de *c*.

Si nous simplifions en supposant que les étiquettes sont distribuées au hasard (ne dépend pas de la longueur de l'affiliation ou de la position au sein de l'affiliation), alors la probabilité d'appartenance d'une donnée à une classe *c*, est

$$P(a \mid c) = \prod P(w_i \mid c)$$

avec w_i = le *i*-ème mot de *a*.

[3] http://www.isni.org
[4] http://viaf.org/
[5] http://www.oclc.org/
[6] http://sourceforge.net/p/oysterer
[7] http://ualr.edu/eriq/people/john-talburt/

puis, en appliquant la règle de Bayes, nous pouvons classer une affiliation dans une classe *c* :

$$c = arg\ max\ P(a \mid c)\ P(c) = \arg\max P(c \mid a)$$

En dépit des deux principaux défauts connus de ces méthodes, qui sont, la non prise en compte de l'ordre des mots et l'hypothèse forte d'indépendance des mots conditionnellement à leur appartenance à une classe, la méthode NB représente une bonne alternative pour résoudre notre problème. Les résultats obtenus par cette mise en œuvre du théorème de Bayes sont valides et démontrés par Hand et Yu [27]. Domingos et al [28] ont montré que les erreurs de classification des méthodes NB sont minimisées par rapport à d'autres méthodes.

B. *Approche par classification semi-supervisée*

Lorsqu'aucune connaissance a priori n'est disponible, nous utilisons une méthode semi-supervisée. Dans ce cas, nous appliquons d'abord une méthode de clustering recouvrante. Nous utilisons ici la méthode des K-means axiales (une variante de la méthode des K-means proposée par Lelu, [29]), ce qui permet de produire des groupes présentant des caractéristiques particulières :

• ils peuvent être recouvrants : un objet ou une variable peut appartenir à plus d'une classe ;
• les éléments constituant une classe (objets et variables) sont classés par ordre décroissant de similarité avec le type de groupe idéal.

Dans la deuxième étape, nous ne retenons que les principaux représentants des classes, qui sont les documents aux valeurs les plus élevées de projection sur les axes représentant les classes et qui n'appartiennent qu'à une seule classe (noyaux de classes). Ces documents sont ensuite utilisés comme corpus d'apprentissage. Ensuite, nous calculons les mots les plus représentatifs de chaque classe et nous utilisons chacun de ces groupes de mots pour interroger le web (via Google). Le site qui représente la réponse la plus pertinente pour une classe donnée, est utilisé pour marquer la classe. Si nécessaire, nous procédons à une étape de fusion de classes.

Dans la dernière phase, nous entrainons la méthode NB avec le corpus défini dans la phase de regroupement et étiqueté par les noms de classes extraits du web. La phase de test est réalisée sur le corpus de documents éliminés après la phase de classification non supervisée.

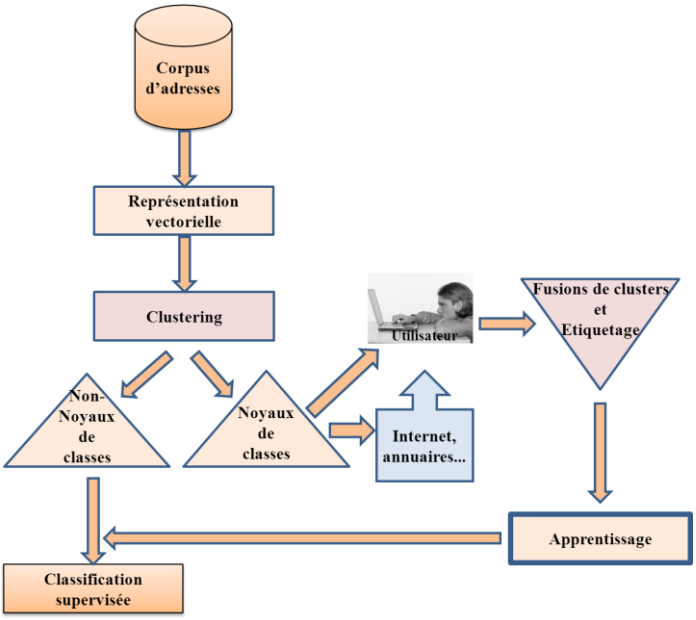

Fig. 3 : Schéma de la méthode de classification semi-supervisée

La figure 3 résume notre approche semi-supervisée avec les trois étapes : soft-clustering, étiquetage et fusion des clusters, et enfin la classification Naive Bayes.

IV. Expérimentations

Dans cette section, nous présentons les résultats obtenus avec trois corpus et les deux méthodes présentées précédemment.

A. *Les données*

Nous avons utilisé trois ensembles de données différents : un premier jeu de données de 10 057 affiliations français (ci-après noté A1), un second petit ensemble de données de 150 affiliations lorraines (région de France) extrait de la base PASCAL (noté A2), et un ensemble de données de 2266 affiliations françaises extraites du WOS[8] et de SCI (noté A3). Tous ces ensembles de données ont été prétraités en extrayant les mots des affiliations, utilisant l'espace comme séparateur (les signes de ponctuation, y compris les tirets sont retirés). Compte tenu de la difficulté que nous avons eu à couper les affiliations en mots dans les corpus A1 et A2, nous avons utilisé une technique basée sur les n-grammes après avoir converti les affiliations en une chaîne de caractères sans espace ni ponctuation. L'apprentissage supervisé est appliqué sur les données A1 et la méthode semi-supervisée est appliquée sur les deux autres ensembles de données.

[8] http://www.webofknowledge.com

LOR1A INRIA CNRS UMR 7503 BP 239	LORIA
LORIA INRIA Lorraine 615 rue du Jardin Botanique	LORIA
LORIA UHP Campus scientifique BP 239	LORIA
LAB-PLANETOL-GRENOBLE, GRENOBLE 9	IPAG
CNRS, UJF, OBSERV GRENOBLE, ASTROPHYS LAB, F-38400 ST-MARTIN-DHERES	IPAG
UNIV-GRENOBLE-1, CNRS, LAB ASTROPHYS GRENOBLE LAOG, UMR 5571, GRENOBLE	IPAG
OBSERV-GRENOBLE, F-38041 GRENOBLE	IPAG
UNIV-GRENOBLE-1, LAB ASTROPHYS GRENOBLE, INSU CNRS, GRENOBLE	IPAG
LAB-ASTROPHYS-GRENOBLE, GRENOBLE	IPAG
LAB-ASTROPHYS-OBSERV-GRENOBLE, GRENOBLE	IPAG
UNIV-STRASBOURG, INST PLURIDISCIPLINAIRE HUBERT CURIEN, CNRS, IN2P3, STRASBOURG	IPHC
UNIV-STRASBOURG, IPHC, CNRS, IN2P3, STRASBOURG	IPHC
ULP, IPHC, IN2P3, F-67037 STRASBOURG	IPHC

Fig. 4 : Echantillon de données avec différentes formes d'adresses

La figure 4 montre un exemple de données avec l'adresse dans la première colonne et le sigle du laboratoire dans la deuxième colonne. Nous pouvons voir trois laboratoires présentés de manière différente.

B. *Mesure de performance*

Les résultats sont évalués en termes de rappel, précision, F-mesure, puisque nous connaissons a priori les classes de toutes les affiliations :

Précision : $$P = \frac{VP}{(VP+FP)}$$

Rappel : $$R = \frac{VP}{(VP+FN)}$$

F-mesure : $$F = \frac{2*P*R}{(P+R)}$$

où VP, FP, FN, signifient respectivement « le nombre de vrais positifs », « le nombre de faux-positifs » et « le nombre de faux négatifs ».

C. *Apprentissage supervisé*

L'ensemble de données A1 a été divisé en un ensemble d'apprentissage (90% du corpus) et un ensemble de test (10% du corpus) successivement représenté par les mots d'affiliations et par les n-grammes. La figure 5 montre la répartition des affiliations dans les 53 classes qui en résultent (test et apprentissage) et met ainsi en évidence le fait que la classification qui en résulte est fortement déséquilibrée.

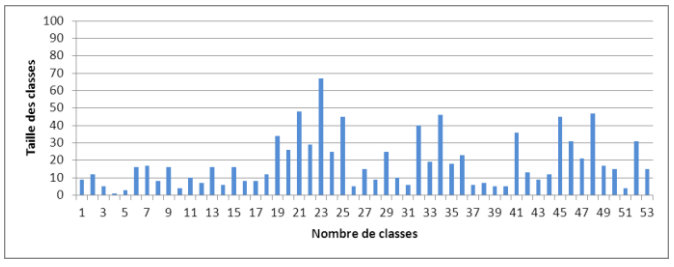

Fig. 5 : Distribution des affiliations dans les classes (corpus A1)

Le résultat de la classification supervisée sur l'ensemble de données A1 est donné tableau 1.

	Rappel	*Précision*	*F-mesure*
N-gram	0.92	0.94	0.93
Mots	0.81	0.88	0.85

Tab. 1 : Valeurs de R, P et F-mesure pour le corpus A1.

En raison du problème de la sélection des mots des affiliations dans le corpus A1, les résultats obtenus avec des n-grammes semblent meilleurs, avec un optimum de rappel de 0,92. Il convient également de noter qu'une recherche systématique des résultats des analyses NB avec une très forte probabilité, tout en étant en contradiction avec les résultats attendus (marquage manuel), nous permet de montrer que l'étiquetage manuel du corpus de test est parfois erroné (le modèle a donné la bonne réponse dans tous les cas !).

D. *Classification semi-supervisée*

Les ensembles de données A2 et A3 ont été utilisés pour cette expérience. Les figures 6a et 6b montrent la répartition des affiliations dans les classes résultantes (A2 : 19 classes ; A3: 10 classes). Elles mettent en évidence que le plus petit ensemble de données (A2) est fortement asymétrique, tandis que le plus gros (A3) est homogène.

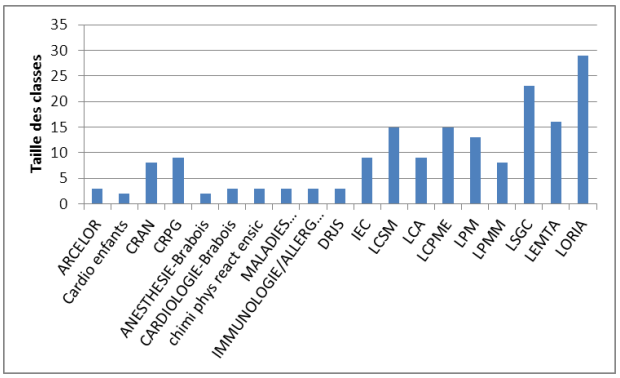

Fig. 6a : Distribution des affiliations dans les classes du corpus A2.

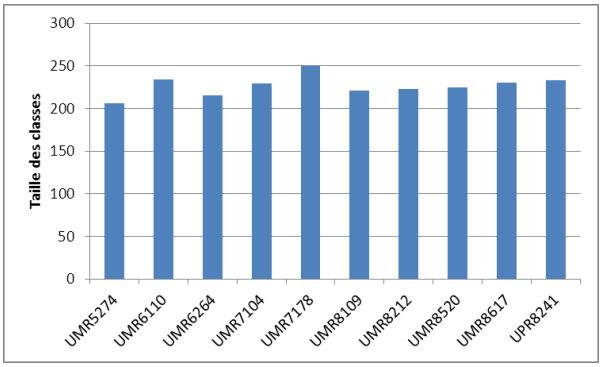

Fig. 6b : Distribution des affiliations dans les classes du corpus A3.

Cette répartition est évidemment intentionnelle, afin de se rendre compte de l'impact de la distribution des données sur les résultats

	Rappel	*Précision*	*F-mesure*
Corpus A2	0.44	0.81	0.55
Corpus A3	0.40	0.95	0.54

Tab. 2 : Resultat des k-means axiales pour les corpus A2 et A3

Avec ce type de données, les résultats d'un simple clustering comme les K-means ne sont pas performants (tableau 2). Cela est probablement dû à la mauvaise représentation des données. Comme nous le verrons dans la dernière section de ce document, une représentation vectorielle prenant en compte le contenu scientifique lié à chacune des adresses devrait améliorer le résultat de clustering.

	Rappel	*Précision*	*F-mesure*
Corpus A2	0.79	0.76	0.73
Corpus A3	0.98	0.97	0.97

Tab. 3 : R, P et F-mesure pour les corpus A2 et A3 dans le cas semi-supervisé

Comme le montre le tableau 3, si nous appliquons l'approche semi-supervisée les résultats sont très bons pour le jeu de données A3, mais significativement plus faible pour le corpus A2, où le rappel est moyen. Cela est dû au grand nombre de classes par rapport à la petite taille du corpus et à l'important déséquilibre de ces classes. En effet, dans ce cas, la méthode de clustering utilisée devient «aveugle» aux petites classes.

V. CONCLUSION ET DISCUSSION

Les résultats obtenus à l'aide de notre approche pour la désambiguïsation des affiliations sont très encourageants, que ce soit dans le contexte d'apprentissage supervisé ou en semi-supervisé. Nos expériences nous permettent également de montrer que notre méthode fournit une aide importante pour corriger les résultats d'étiquetage humain. Cependant, il est clair que nous devons pratiquer davantage d'expériences pour conclure sur la pertinence globale de la méthodologie. Il reste encore quelques faiblesses dans notre méthodologie, principalement liées à la méthode de classification utilisée dans le cas des classes très déséquilibrées. Nous avons donc l'intention de procéder à des essais avec d'autres méthodes de classification et de mettre en œuvre des techniques d'équilibrage de données. Un autre point important serait d'exploiter une méthode d'apprentissage qui devrait être en mesure d'apprendre itérativement avec des chaînes de caractères de longueur variable. Il serait nécessaire de développer une méthode automatique qui pourrait mettre en évidence les cas conflictuels, permettant ainsi à l'utilisateur de ne se focaliser que sur ce nombre réduit de cas. Les travaux futurs devraient également tenir compte de la structure XML (quand elle existe !) afin d'examiner séparément les villes, les noms de rue, les noms des laboratoires ...

Une autre possibilité est de considérer le contenu scientifique des documents, tels que les titres et les résumés des articles publiés. Une fois ces documents indexés, chaque adresse serait représentée par un vecteur de mots (décrivant les activités de recherche) permettant probablement une classification plus pertinente.

L'étude de la transitivité peut certainement permettre de détecter de faux positifs ou faux négatifs et isoler ainsi les résultats à vérifier.

Bien entendu tous les résultats obtenus (et validés) doivent être capitalisés dans une base de référence.

Nous proposons également de comparer nos résultats avec ceux obtenus en utilisant le logiciel OYSTER.

VI. BIBLIOGRAPHIE

[1] Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1), 133–143.

[2] Liu, N. C., Cheng, Y. & Liu, L. (2005). Academic ranking of world universities using scientometrics:A comment to the Fatal Attraction. Author's reply. *Scientometrics*, 64(1), 101–112.

[3] Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Springer.

[4] Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics*, 1-21.

[5] Hood, W., & Wilson, C. (2003). Informetric studies using databases: Opportunities and challenges. *Scientometrics*, 58(3), 587-608.

[6] De Bruin, R. E. & Moed, H. F. (1990). The unification of addresses in scientific publications. *Informetrics* 1989/90. Elsevier Science Publishers, Amsterdam, 65–78.

[7] Zitt, M., & Bassecoulard, E. (2008). Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues. *Ethics in Science and Environmental Politics*, 8, 49–60.

[8] Bourke, P., & Butler, L. (1996). Standards issues in a national bibliometric database: The Australian case. *Scientometrics*, 35(2), 199-207. doi:10.1007/BF02018478

[9] Osareh, F., & Wilson, C. S. (2000). A comparison of Iranian scientific publications in the Science Citation Index: 1985-1989 and 1990-1994. *Scientometrics*, 48(3), 427-442.

[10] De Bruin, R. E., & Moed, H. F. (1993). Delimitation of scientific subfields using cog nitive words from corporate addresses in scientific publications. *Scientometrics*, 26(1), 65–80.

[11] French, J.C., Powell, A.L. & Schulman, E. (2000). Using Clustering Strategies for Creating Authority Files. *Journal of the American Society for Information Science and Technology*, 51, 774–786.

[12] Galvez, C., & Moya-Anegón, F. (2006). The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). *Scientometrics*, 69(2), 323–345.

[13] Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, 23, 2000.

[14] Fellegi, I., & Sunter, A. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

[15] Sadinle, M., Hall, R., & Fienberg, S. (2010). Approaches to Multiple Record Linkage. cscmuedu.

[16] Sadinle, M., & Fienberg, S. E. (2012). A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record-Systems. *arXiv:1205.3217. http://arxiv.org/abs/1205.3217*

[17] Ventura, S. L., Nugent, R., & Fuchs, E. R. H. (2012). Methods Matter: Revamping Inventor Disambiguation Algorithms with Classification Models and Labeled Inventor Records. *SSRN eLibrary*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2079330

[18] Churches, T., Christen, P., Lim, K., & Zhu, J. X. (2002). Preparation of name and address data for record linkage using hidden Markov models. *BMC Medical Informatics and Decision Making*, 2, 9. doi:10.1186/1472-6947-2-9

[19] Huang, J., Ertekin, S., & Giles, C. L. (2006). Efficient name disambiguation for large-scale databases. *PKDD*'06 (p. 536–544). Springer-Verlag.

[20] Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5), 16-23. doi:10.1109/MIS.2003.1234765

[21] Carayol, N., & Cassi, L. (2009). *Who's Who in Patents. A Bayesian approach*. Consulté de http://hal-paris1.archives-ouvertes.fr/hal-00631750

[22] Niu, L., Wu, J., & Shi, Y. (2012). Entity Disambiguation with Textual and Connection Information. *Procedia Computer Science*, 9(0), 1249-1255.

[23] Aswani, N., Bontcheva, K., & Cunningham, H. (2006). Mining Information for Instance Unification. In *The Semantic Web - ISWC 2006, Lecture Notes in Computer Science* (Vol. 4273, p. 329-342). Springer Berlin / Heidelberg. http://www.springerlink.com/content/e01513178732676 2/abstract/

[24] Jiang, Y., Zheng, H.-T., Wang, X., Lu, B., & Wu, K. (2011). Affiliation disambiguation for constructing semantic digital libraries. *Journal of the American Society for Information Science and Technology*, 62(6), 1029-1041.

[25] Zhou, Y., Talburt, J. R., Su, Y., & Yin, L. (2010). OYSTER: A Tool for Entity Resolution in Health Information Exchange. Proceedings of the *5th International Conference on Cooperation and Promotion of Information Resources in Science and Technology* (COINFO 2010 E-BOOK), 358-364.

[26] Talburt J.R. (2011) : *Entity resolution and information quality*. 235 p., Elsevier.

[27] Hand, D. J., & Yu, K. (2001). Idiot's Bayes—Not So Stupid After All? *International Statistical Review*, 69(3), 385-398.

[28] Domingos, P. & Pazzani, M. (1996). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *Machine Learning* (p. 105–112).

[29] Lelu, A. (1993). *Modèles neuronaux pour l'analyse de données documentaires et textuelles*. Doctorat de l'université Paris 6.