Classification Hiérarchique Multi-Etiquette de Larges Volumes de Données par Raisonnement Sémantique

Thomas Hassan[*,**], Rafael Peixoto[*,***]
Christophe Cruz[*,****] Aurelie Bertaux[*,♯]

[*]Le2i UMR6306, CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté
[**]thomas.hassan@u-bourgogne.fr, [***]rafpp@isep.ipp.pt
[****]christophe.cruz@u-bourgogne.fr, [♯]aurelie.bertaux@iut-dijon.u-bourgogne.fr

Résumé. Cet article présente un résumé de l'approche développée dans Peixoto et al. (2016b), et complémente l'évaluation qualitative de l'approche par de nouveaux résultats sur un jeu de données supervisé. Pour réaliser l'extraction d'informations de valeur à partir de larges volumes de données (big data), un processus de classification hiérarchique multi-étiquettes appelé "Semantic HMC" (Hierarchical Multi-label Classification) est présenté. Ce processus est composé de cinq étapes. Les trois premières étapes construisent automatiquement une hiérarchie d'étiquettes via une analyse statistique des données. Cet article s'intéresse aux deux dernières étapes, qui permettent de classer de nouvelles données selon la hiérarchie d'étiquettes. Une évaluation qualitative compare l'approche à l'état de l'art et montre que l'approche "Semantic HMC" est plus performante que les autres approches pour certains critères, tout en conservant les avantages de l'orientation big data en terme de performances.

1 Introduction

Werner et al. (2014) propose une méthode pour enrichir sémantiquement une ontologie utilisée pour classer des articles de presse, en utilisant un raisonneur sémantique basé sur la logique de description (DL). Bien que ce type de raisonneur soit valide pour effectuer des raisonnements même pour une un niveau de logique de description (expressivité) élevé, ils ne permettent pas de gérer de grandes quantités de données. Notre objectif est d'étendre le travail de Werner et al. (2014), et d'extraire des informations de valeur à partir de larges volumes de données textuelles (big data), en utilisant un processus de Classification Hiérarchique Multi-étiquette Sémantique ("Semantic HMC", Peixoto et al. (2016a)). Semantic HMC est un processus d'apprentissage d'ontologie non supervisé, basé sur les technologies du big data, le machine-learning, et le raisonnement sémantique basé sur des règles. Il est composé de 5 étapes :

— **Indexation** : extrait les termes des items (documents textes), et créé un index inversé des items.
— **Vectorisation** : calcule les vecteurs de fréquence des termes à partir de l'index inversé. L'ensemble de vecteurs de termes permet de générer une matrice de co-occurrence des termes.

— **Hiérarchisation** : détermine les termes les plus pertinents, i.e. les $Labels$ (étiquettes), et génère une hiérarchie de subsomption des labels à partir de la matrice de co-occurrence.
— **Résolution** : créé des règles de classification qui lient les nouveaux items aux labels à partir de la matrice de co-occurrence.
— **Réalisation** : remplit l'ontologie avec les nouveaux items et pour chaque item détermine les labels les plus spécifiques.

Peixoto et al. (2016a) se concentre sur les deux dernières étapes du processus, i.e. la classification des items à partir d'une ontologie en Logique Descriptive, en utilisant un raisonneur Sémantique basé sur les règles. Peixoto et al. (2016b) présente également une évaluation qualitative et une comparaison à d'autres algorithmes de classification multi-étiquette de l'état de l'art.

Cet article résume l'approche et les résultats décrits dans Peixoto et al. (2016b), non publiés en français. L'évaluation qualitative est étendue avec un autre jeu de données supervisé, qui permet de d'affiner les conclusions sur les capacités du processus dans la tâche de classification. Le reste de l'article est divisé en 5 sections : la section suivante présente le contexte de ces travaux. Les sections 3 et 4 décrivent les processus de création de règles et de classification ainsi que leur implémentation. La section 5 présente les résultats de l'évaluation qualitative. La dernière section conclue et décrit les axes de recherche futurs.

2 Contexte

2.1 Ontologies pour la tâche de classification

L'utilisation des ontologies pour la tâche de classification porte souvent sur la description du modèle de classification (étiquette, items, règles de classification). Galinina et Borisov (2013) utilise deux ontologies dans un système de classification : (1) une ontologie de domaine, indépendante de la méthode de classification, et (2) une ontologie dédiée à la méthode de classification basée sur un arbre de décision. Au delà de la description du domaine, les ontologies sont utilisées pour améliorer le processus de classification. Elberrichi et al. (2012) présente une méthode en deux étapes pour améliorer la classification de documents médicaux (MeSH - Medical Subject Headings). Leurs résultats montrent que l'utilisation d'une ontologie de domaine permet d'améliorer la performance de la méthode de classification de documents.

2.2 Raisonnement sémantique pour la tâche de classification

La plupart des travaux de la littérature se concentrent sur l'amélioration du processus de classification en utilisant les ontologies, ce qui permet d'améliorer la description des items. En revanche, ils ne tirent pas avantage des capacités des raisonneurs sémantiques pour classer automatiquement des items (Peixoto et al. (2016a)). L'utilisation de raisonneurs pour la tâche de classification peut améliorer le processus de classification (Moller et Haarslev (2009)). Fang et al. (2010) présente une méthode de classification de documents basée sur un raisonneur et des mesures de similarité. Ben-David et al. (2010) utilise les ontologies et un raisonneur pour définir le modèle de classification, effectuer la classification et représenter les résultats. Une ontologie de domaine décrit les entités sur lesquelles sont basées la classification.

Werner et al. (2014) utilise une ontologie de domaine dans un processus de classification de nouvelles économiques. L'ontologie est basée une hiérarchie de termes qui décrit le domaine et permet d'effectuer la classification. L'ontologie est enrichie avec les documents à classifier, et un raisonneur basé sur la logique de description (DL) tel que Pellet (Sirin et al. (2007)), FaCT++ (Tsarkov et Horrocks (2006)), ou Hermit (Shearer et Horrocks (2009)) est utilisé pour la tâche de classification.

L'approche de Werner et al. (2014) montre que les raisonneurs sémantiques basés sur la logique de description (DL) ne sont pas adaptés pour la montée en charge dans le cadre du traitement massif de données (big data). Des raisonneurs à échelle du web (Urbani (2013)) utilisent un raisonnement basé sur des règles, et permettent un passage à l'échelle par parallélisation et distribution de la charge de calcul sur des clusters de machines. La montée en charge peut ainsi s'effectuer via des techniques telles que Map-Reduce.

3 Classification multi-étiquette hiérarchique

Dans Peixoto et al. (2015), les 3 premières étapes du processus Semantic HMC sont décrites. Les sections suivantes décrivent les deux dernières étapes, i.e. la création des règles et la classification des items.

3.1 Résolution

La Résolution créé les règles de l'ontologie qui permettent de classer les items avec des labels, c'est à dire les conditions nécessaires pour qu'un $item_i$ soit classé avec un $label_j$, en fonction de l'ensemble de termes $\omega^{(term_j)}$ extrait de l'item $item_i$.

Lors de la Vectorization (Peixoto et al. (2015)), une matrice de co-occurrence des termes $cfm(term_i, term_j)$ est créée pour représenter chaque paire de termes dans le corpus d'items C. Soit $P(term_j|term_i)$ la proportion conditionnelle (nombre) d'items du corpus C communs aux termes $term_i$ et $term_j$, en fonction du nombre d'items qui comportent le terme $term_j$, tel que :

$$P_C(term_i|term_j) = \frac{cfm(term_i, term_j)}{cfm(term_j, term_j)} \tag{1}$$

Le processus de création de règles utilise des seuils pour déterminer la pertinence des termes et créer les règles (Werner et al. (2014)). Deux seuils sont définis :

— Seuil Alpha (α) tel que $\alpha < P_C(term_i|term_j)$, où $term_i \in Label$ et $term_j \in Term$.
— Seuil Bêta (β) tel que $\beta \leq P_C(term_i|term_j) \leq \alpha$, où $term_i \in Label$ et $term_j \in Term$.

Ces deux seuils sont définis par l'utilisateur dans l'intervalle $[0, 1]$. A partir de ces seuils, deux ensembles de termes sont déterminées (Figure 1) :

— L'ensemble Alpha ($\omega_\alpha^{(term_i)}$) est l'ensemble de termes $term_j$ qui co-occurrent avec $term_i \in Label$ tel que la proportion conditionnelle est supérieure au seuil α.
— L'ensemble Bêta ($\omega_\beta^{(term_i)}$) est l'ensemble de termes qui co-occurrent avec $term_i \in Label$ tel que proportion conditionnelle est supérieure au seuil β et inférieure au seuil α.

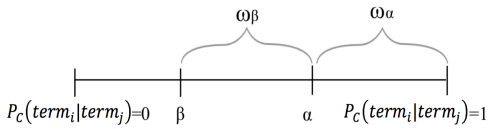

FIG. 1 – *Ensembles Alpha et Beta*

Si un item $item_i$ a au moins un terme dans l'ensemble $\omega_\alpha^{(term_i)}$ il est classé avec le label $term_i$, $term_i \in Label$. Pour chaque terme vérifiant cette règle, une règle de classification est générée au format SWRL (cf. section suivante). Pour $|\omega_\alpha^{(term_i)}| = |\{t_1, t_2\}|$, un exemple de règles SWRL générées est présenté dans le tableau 1.

Si un item $item_i$ a au moins δ termes dans l'ensemble $\omega_\beta^{(term_i)}$, il est classé avec le label $term_i$, $term_i \in Label$. Une règle est générée pour chaque combinaison $term_j \in \omega_\beta^{(term_i)}$ où le nombre de termes combinés est $\delta = \lceil |\omega_\beta^{(term_i)}| * p \rceil$, et $0 \leq p \leq 0.5$. Le tableau 2 présente un exemple de règles $beta$ générées pour $|\omega_\beta^{(term_i)}| = |\{t_1, t_2, t_3\}| = 3$ et $p = 0.5$, tel que $\delta = \lceil 3 * 0.5 \rceil = 2$.

TAB. 1 – *Exemples de règles Alpha générées*

Alpha rules
$Item(?it), Term(?t_1), Label(?t_1), hasTerm(?it, ?t_1) \rightarrow isClassified(?it, ?t_1)$
$Item(?it), Term(?t_2), Label(?t_2), hasTerm(?it, ?t_2) \rightarrow isClassified(?it, ?t_2)$

TAB. 2 – *Exemples de règles Beta générées*

Beta rules
$Item(?it), Term(?t_1), Term(?t_2), Label(?t_3),$ $hasTerm(?it, ?t_1), hasTerm(?it, ?t_2) \rightarrow isClassified(?it, ?t_3)$
$Item(?it), Term(?t_1), Term(?t_3), Label(?t_2),$ $hasTerm(?it, ?t_1), hasTerm(?it, ?t_3) \rightarrow isClassified(?it, ?t_2)$
$Item(?it), Term(?t_2), Term(?t_3), Label(?t_1),$ $hasTerm(?it, ?t_2), hasTerm(?it, ?t_3) \rightarrow isClassified(?it, ?t_1)$

L'ensemble des règles beta générées est la combinaison C_n^m de m termes parmi n éléments. Dans notre approche n est le nombre de termes possibles $|\omega_\beta^{(term_i)}|$, et m est le nombre minimum de termes δ dans chaque règle (exemple : $C_{20}^{10} = 184756$). Pour limiter le nombre de règles potentiellement générées pour chaque label, n est fixé tel que $n \leq 10$. Les termes sont sélectionnés en classant les termes de l'ensemble $\omega_\beta^{(term_i)}$ en fonction de leur proportion conditionnelle $P_C(term_i|term_j)$. L'ensemble de règles $beta$ est la combinaison C_n^m de m termes parmi n éléments. Dans l'approche, n est le nombre de termes possibles $|\omega_\beta^{(term_i)}|$, et m est le nombre minimal de termes δ présents dans chaque règle (exemple : $C_{20}^{10} = 184756$). Pour limiter le nombre total de règles pour chaque label, une valeur maximale de n est fixée,

i.e. $n \leq 5$. Les termes issus de $\omega_{\beta}^{(term_i)}$ sont alors sélectionnés par classement en fonction de la proportion conditionnelle $P_C(term_i|term_j)$.

Le résultat de la phase de résolution est l'ensemble de règles nécessaires à la classification d'un item pour chaque label.

3.2 Réalisation

La Réalisation est composée de 2 sous-étapes : le peuplement et la classification. Lors du peuplement, la base de connaissance est peuplée avec les nouveaux items et leurs termes pertinents (Abox). Chaque item est décrit par un ensemble de termes pertinents $\omega_{\gamma}^{(item_i)}$ tel que :

$$\omega_{\gamma}^{(item_i)} = \{term_j | \forall term_j \in Term \wedge \gamma < tfidf_{(item_i, term_j, C)}\} \tag{2}$$

où γ est le seuil de pertinence, $\gamma < tfidf_{(item_i, term_j, C)}$, $term_j \in Term$, $item_i \in Item$ et où $tfidf$ est calculé de la même façon que lors de la Vectorisation.

La classification effectue la classification hiérarchique multi-étiquette des items. Un raisonneur sémantique applique de façon exhaustive l'ensemble des règles à un ensemble de triplets (les items et leurs termes), et infère les labels correspondant à chaque item, i.e. leur classification. Le moteur d'inférence infère également les labels les plus spécifiques via une règle de subsomption à partir de la hiérarchie de labels. Le résultat est une classification hiérarchique multi-étiquette des items basée sur la structure hiérarchique des labels ("Hierarchical Multi-label Classification").

4 Implémentation

L'implémentation du processus consiste en une application Java distribuée et déployée sur un cluster Hadoop [1]. Un ensemble de librairies qui supportent nativement les différentes parties du processus sont utilisées. Dans les 3 premières étapes du processus, le modèle MapReduce (Dean et Ghemawat (2008)) est utilisé pour paralléliser le processus.

Lors de la Résolution, un job MapReduce créé les règles de classification à partir de la matrice de coocurrence (Peixoto et al. (2016a)). La principale différence avec Werner et al. (2014) est que plutôt que de traduire les règles en contraintes logiques de l'ontologie en logique de description, les règles sont transcrites au format SWRL (Semantic Web Rule Language) puis stockées dans l'ontologie (Rbox). Le principal intérêt des règles SWRL est de réduire la charge de calcul du raisonneur, donc d'améliorer la performance du système. L'ontologie composée de la hiérarchie de labels et des règles de classification est utilisée lors de la Réalisation pour classer les nouveaux items.

Lors de la Réalisation, l'ontologie est peuplée avec les nouveaux items et les terms extraits de ces items en tant qu'individus (Abox). Pour stocker et requêter la nouvelle base de connaissances (Tbox+Abox), le triple-store Stardog [2] est utilisé. La librairie OWL-API est utilisée pour générer les règles de classification et peupler l'ontologie.

1. `https://hadoop.apache.org/`
2. `http://docs.stardog.com`

5 Évaluation qualitative

Cette section présente une évaluation qualitative du processus Semantic HMC. Cette évaluation étend les tests de performance (passage à l'échelle) décrits dans Peixoto et al. (2016a), et se concentre sur la précision pour la tâche de classification.

Pour comparer l'approche à la littérature, deux jeux de données d'apprentissage supervisé issues du projet Mulan [3] sont utilisés dans cette évaluation. Le jeu de données Delicious est composé de données textuelles pré-étiquetées issues du site de bookmarking collaboratif du même nom (Tsoumakas et al. (2008)). Le choix de ce jeu de données se base sur le fait qu'il contient très peu de caractéristiques (500 mots) comparé au nombre d'étiquettes (983), ce qui rend la classification difficile (Papanikolaou et al. (2015)). Le jeu de données est divisé en un jeu d'entraînement (12910 items) utilisé pour apprendre l'ontologie, i.e. la hiérarchie d'étiquettes et les règles, et un jeu de test (3181 items).

Le jeu de données Bibtex est composé de données textuelles pré-étiquetées issues de métadonnées d'articles scientifiques au format bibtex(Tsoumakas et al. (2008)). Les résultats obtenus avec ce jeu de données sont inédits. Le motif pour l'utilisation de ce jeu de données est la différence importante avec le jeu de données Delicious par rapport au ratio de caractéristiques (features, i.e. termes) sur le nombre de classes (étiquettes). Comparativement au jeu de données Delicious, le nombre de caractéristiques est élevé comparé au nombre de classes (1836 caractéristiques pour 159 classes/étiquettes). Le jeu de données est divisé en un jeu d'entraînement (4880 items) et un jeu de test (2515 items).

5.1 Description

Des métriques standard basées-étiquette sont utilisées pour évaluer l'approche Semantic HMC sur les deux jeux de test : la micro précision moyenne, et le micro rappel (recall) moyen. Ces mesures sont calculées de la même façon que Madjarov et al. (2012), tel que :

$$Precision_{micro} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i} \tag{3}$$

$$Rappel_{micro} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i} \tag{4}$$

et la macro précision moyenne / le macro rappel (recall) moyen, tel que :

$$Precision_{macro} = \frac{1}{n}\sum_{i}^{n}(\frac{TP_i}{(TP_i + FP_i}) \tag{5}$$

$$Rappel_{macro} = \frac{1}{n}\sum_{i}^{n}(\frac{TP_i}{TP_i + FN_i}) \tag{6}$$

où TP_i est le nombre de vrais positifs, FP_i le nombre de faux positifs, et FN_i le nombre de faux négatifs pour le $Label_i$. Une métrique d'évaluation complémentaire est la F1-mesure, i.e.

3. http://mulan.sourceforge.net/datasets-mlc.html

la moyenne harmonique de la précision et du rappel où le poids de chaque composante est égal. Les micro et macro F1-mesure sont définies de façon similaires par la relation :

$$F1 = \frac{2 * (Precision * Rappel)}{(Precision + Rappel)} \tag{7}$$

Les paramètres (seuils) du processus Semantic HMC peuvent avoir un impact important sur les résultats. Les seuils Top et Bottom de la hiérarchisation influent sur la création de relations hiérarchiques entre paires de concepts, tandis que les seuils alpha et beta de la résolution influent sur la création des règles de classification. Des seuils bas engendrent un nombre de règles/relations hiérarchiques élevé au détriment de leur qualité, et inversement. Une étude exhaustive de l'impact des combinaisons de seuils possibles sur les métriques permet de déterminer les seuils les plus appropriés en fonction des résultats obtenus. Le tableau 3 décrit l'ensemble des valeurs utilisées pour l'évaluation ci-après. Pour comparer de façon objective les résultats des deux jeux de données, les paramètres sont identiques dans les deux tests.

TAB. 3 – *Valeurs des paramètres pour l'évaluation*

Paramètre	Etape	Valeur
Seuil Top	Hiérarchisation	50
Seuil Bottom	Hiérarchisation	40
Seuil Alpha	Résolution	20
Seuil Beta	Résolution	10
Classement des termes (n)	Résolution	5
p	Résolution	0.25
Seuil de Pertinence (γ)	Réalization	2

5.2 Résultats

La figure 2 est un exemple de relations hiérarchiques obtenues à partir du jeu de données Delicious.

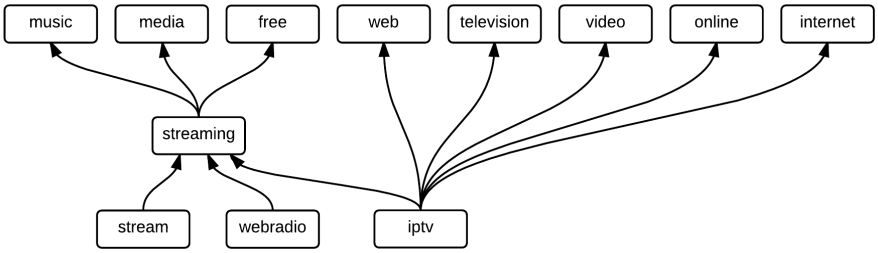

FIG. 2 – *Portion de la hiérarchie de labels automatiquement générée à partir du jeu de données Delicious*

Le tableau 4 montre les résultats de l'évaluation pour les métriques standard basées étiquettes sur les deux jeux de données. La moyenne harmonique (F1-mesure) est également reportée et sera ensuite utilisée pour comparer les résultats à d'autres approches.

TAB. 4 – *Résultats de l'évaluation sur le jeu de données Delicious*

		Précision	Rappel	F1-mesure
Delicious	Micro	0.284	0.74	0.410
	Macro	0.0676	0.178	0.0979
Bibtex	Micro	0,199	0,741	0,314
	Macro	0,188	0,444	0,264

TAB. 5 – *Comparaison aux approches de la littérature (Delicious/Bibtex)*

	Delicious		Bibtex	
Algorithme	Macro F1	Micro F1	Macro F1	Micro F1
SHMC	0.0979	**0.410**	0,264	0,314
CGS_p	0.104	0.297	0.258	0.363
TNBCC	0.088	N/A	0.189	N/A
Path-BCC	0.084	N/A	0.212	N/A
BR	0.096	0.234	0.307	0.457
CC	0.100	0.236	**0.316**	**0.462**
HOMER	0.103	0.339	0.266	0.429
ML-kNN	0.051	0.175	0.065	0.206
RFML-C4.5	**0.142**	0.269	0.016	0.123
RF-PCT	0.083	0.248	0.055	0.230

Le tableau 5 compare les valeurs de macro F1-mesure et micro F1-mesure obtenues avec différentes approches de l'état de l'art décrites dans Madjarov et al. (2012), Sucar et al. (2014),

Papanikolaou et al. (2015). Peixoto et al. (2016b) décrit chacune des approches de façon exhaustive.

Pour le jeu de données Delicious, le tableau 5 montre que l'approche Semantic HMC est légèrement plus performante que les autres approches pour la micro F1-mesure, et que la macro F1-mesure est comparable aux autres approches. Les résultats sur le jeu de données Bibtex montrent que l'approche est moins performante que les meilleures approches de l'état de l'art (Binary relevance, Classifier Chain et HOMER). Comparativement aux approches considérées, SHMC reste au niveau de la moyenne de macro F1-mesure (0.19) et de micro F1-mesure (0.32).

Suivant les critères sélectionnés, l'évaluation montre que la performance l'approche SHMC est comparable aux autres approches de l'état de l'art pour la tâche de classification multi-étiquette par apprentissage supervisé. L'orientation de traitement massif de données de l'approche SHMC conserve des performances proches de l'état de l'art pour la tâche de classification.

6 Conclusion

Cet article décrit un processus de classification hiérarchique multi-étiquette ("Semantic HMC") pour des données textuelles non structurées dans un contexte Big Data. Suivant l'évaluation de la performance du processus décrite dans Peixoto et al. (2016a), une évaluation qualitative du processus compare l'approche Semantic HMC à d'autres approches de l'état de l'art.

Les résultats montrent que l'approche basée sur l'apprentissage automatique d'une ontologie et le raisonnement sémantique est comparable aux autres approches de l'état de l'art. Ces résultats sont toutefois très dépendants du paramétrage du processus, et l'application à un jeu de données réel nécessiterait d'ajuster les paramètres et d'étudier de façon exhaustive leur impact. L'adaptation (maintenance) du modèle de classification en fonction des nouvelles données, et la sélection des paramètres optimaux basée sur ces indicateurs de performance font l'objet de travaux en cours. Contrairement aux autres approches du domaine de la recherche d'information, l'approche basée sur les ontologies permet par définition d'avoir une vue explicative des résultats (classifications), ce qui rend la révision des résultats par des experts possible.

Les travaux en cours et futurs se concentrent sur deux points : (1) l'application du processus à des données spécifiques au domaine de la veille économique, et (2) la maintenance de la base de connaissances à partir d'un flux de données dans un contexte Big Data.

7 Remerciements

Ce projet est financé par l'entreprise Actualis SARL, l'agence française ANRT, le conseil régional de Bourgogne Franche-comté et le Fonds Européen de Développement Economique et Régional (FEDER).

Références

Ben-David, D., T. Domany, et A. Tarem (2010). Enterprise data classification using semantic web technologies. In *The Semantic Web-ISWC 2010*, ISWC'10, Berlin, Heidelberg, pp. 66–81. Springer-Verlag.

Dean, J. et S. Ghemawat (2008). MapReduce : Simplified Data Processing on Large Clusters. *Communications of the ACM 51*(1), 1–13.

Elberrichi, Z., B. Amel, et T. Malika (2012). Medical Documents Classification Based on the Domain Ontology MeSH. *arXiv preprint arXiv :1207.0446*.

Fang, J., L. Guo, et Y. Niu (2010). Documents classification by using ontology reasoning and similarity measure. In *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, Volume 4, pp. 1535–1539.

Galinina, A. et A. Borisov (2013). Knowledge modelling for ontology-based multiattribute classification system. *Applied Information and Communication ...*, 103–109.

Madjarov, G., D. Kocev, D. Gjorgjevikj, et S. Džeroski (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition 45*(9), 3084–3104.

Moller, R. et V. Haarslev (2009). Tableau-Based Reasoning.

Papanikolaou, Y., T. N. Rubin, et G. Tsoumakas (2015). Improving gibbs sampling predictions on unseen data for latent dirichlet allocation. *arXiv preprint arXiv :1505.02065*.

Peixoto, R., T. Hassan, C. Cruz, A. Bertaux, et N. Silva (2015). Semantic HMC : A Predictive Model using Multi-Label Classification For Big Data. In *The 9th IEEE International Conference on Big Data Science and Engineering (IEEE BigDataSE-15)*.

Peixoto, R., T. Hassan, C. Cruz, A. Bertaux, et N. Silva (2016a). An unsupervised classification process for large datasets using web reasoning. In ACM (Ed.), *SBD'16 : Semantic Big Data Proceedings*, San Francisco (CA), USA.

Peixoto, R., T. Hassan, C. Cruz, A. Bertaux, et N. Silva (2016b). Hierarchical multi-label classification using web reasoning for large datasets. *Open Journal of Semantic Web (OJSW) 3*(1), 1–15.

Shearer, R. et I. Horrocks (2009). Hypertableau Reasoning for Description Logics. *Journal of Artificial Intelligence Research 36*(June 2008), 165–228.

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur, et Y. Katz (2007). Pellet : A practical OWL-DL reasoner.

Sucar, L. E., C. Bielza, E. F. Morales, P. Hernandez-Leal, J. H. Zaragoza, et P. Larranaga (2014). Multi-label classification with bayesian network-based chain classifiers. *Pattern Recognition Letters 41*, 14 – 22. Supervised and Unsupervised Classification Techniques and their Applications.

Tsarkov, D. et I. Horrocks (2006). FaCT++ Description Logic Reasoner : System Description. In U. Furbach et N. Shankar (Eds.), *Proceedings of the Third International Joint Conference (IJCAR)*, Volume 4130, pp. 292–297. Springer Berlin / Heidelberg.

Tsoumakas, G., I. Katakis, et I. Vlahavas (2008). Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, pp. 30–44.

Urbani, J. (2013). Three Laws Learned from Web-scale Reasoning. In *2013 AAAI Fall Symposium Series*, pp. 76–79.

Werner, D., N. Silva, C. Cruz, et A. Bertaux (2014). Using DL-reasoner for hierarchical multilabel classification applied to economical e-news. In *Proceedings of 2014 Science and Information Conference, SAI 2014*, pp. 313–320.

Summary

This article recaps a previously published approach in Peixoto et al. (2016b), and complements its quality evaluation with new results using a different dataset for supervised classification. To perform Value extraction from Big Data sources, a Hierarchical Multi-Label Classification process called Semantic HMC is presented. This process is composed of five scalable steps. The first three steps automatically constructs a label hierarchy from statistical analysis of Data. This paper focuses on the last two steps which perform item classification according to the label hierarchy. A quality evaluation compares the approach with the state of the art and shows that Semantic HMC outperforms other approaches in some areas while scaling in a Big Data context.