Impact des mesures de similarité sémantique dans un algorithme de partitionnement : d'un cas biomédical à la détection de comportements de consommation

Jocelyn Poncelet, Pierre-Antoine Jean, Jacky Montmain, François Trousset, Sébastien Harispe, Nicolas Pecheur

▶ **To cite this version:**

Jocelyn Poncelet, Pierre-Antoine Jean, Jacky Montmain, François Trousset, Sébastien Harispe, et al.. Impact des mesures de similarité sémantique dans un algorithme de partitionnement : d'un cas biomédical à la détection de comportements de consommation. SFC 2019 - XXVIe Rencontres de la Société Francophone de Classification, Sep 2019, Nancy, France. hal-02292992

HAL Id: hal-02292992
https://hal.science/hal-02292992

Submitted on 23 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Impact des mesures de similarité sémantique dans un algorithme de partitionnement : d'un cas biomédical à la détection de comportements de consommation

Jocelyn Poncelet*, Pierre-Antoine Jean*, François Trousset*,
Sebastien Harispe*, Nicolas Pecheur**, Jacky Montmain*

*LGI2P - IMT Mines Alès - Université de Montpellier, Alès, France
{prénom}.{nom}@mines-ales.fr
**TRF retail - 116 Allée Norbert Wiener - Nîmes, France
{prénom}.{nom}@trfretail.com

Résumé. Les approches permettant de regrouper/segmenter des objets partageant ou non des caractéristiques similaires sont nombreuses. Des distances classiques sont utilisées selon l'hypothèse que ces caractéristiques sont indépendantes et définissent un espace métrique. Cependant, lorsque ces caractéristiques sont organisées dans une représentation des connaissances, ces métriques deviennent discutables. Cet article propose la comparaison de distances et de mesures de similarité au sein d'une approche de partitionnement ascendant hiérarchique. L'étude cherche à mettre en évidence l'intérêt d'approches sémantiques permettant de détecter des comportements de consommation. Un parallèle avec le domaine biomédical a été réalisé pour pallier le manque de données dans le domaine de la grande distribution et valider notre approche.

1 Introduction et état de l'art

Dans les secteurs industriels tels que la grande distribution, le partitionnement de clients est une problématique récurrente. Cela permet de recueillir des informations essentielles pour contrôler et définir des stratégies commerciales et marketing orientées autour d'indicateurs, *e.g.* RFM pour *Recency, Frequency, Monetary* ou CLV pour *Customer Lifetime Value* (Chen et al., 2009; Ching-Hsue et You-Shyang, 2009). Les approches de partitionnement traditionnelles considèrent une représentation vectorielle des clients où les dimensions sont essentiellement des mesures monétaires des comportements de consommation (fréquence, volume d'achats, etc.). Elles conduisent donc plus à une segmentation des clients orientée « valeur monétaire » qu'à une véritable caractérisation du client. Il existe pourtant un réel besoin d'analyses dirigées par la caractérisation des produits achetés pour identifier précisément les typologies de comportements de consommation afin de mieux comprendre et anticiper les changements dans les habitudes d'achats. Dans ce cadre, les dimensions de l'espace à partitionner pourraient correspondre aux produits en vente dans un supermarché avec pour chaque composante une

valeur booléenne (acheté ou pas) [1], une fréquence d'achats, etc. Ainsi, si trois clients achètent respectivement dans une boulangerie : des bonbons, des chewing-gums et une baguette ; ils seront considérés équidistants si l'on utilise, par exemple, une distance euclidienne dans cet espace de dimension 3. Pourtant, de façon intuitive, les personnes qui achètent des bonbons ou des chewing-gums ont des comportements proches et distincts du client qui achète du pain. Cette intuition repose sur l'idée qu'il existe un lien sémantique qui « rapproche » bonbons et chewing-gums : ce sont des friandises. Pour formaliser cette intuition, les mesures de similarité sémantique constituent une solution intéressante pour introduire la connaissance a priori associée à un domaine. Par ailleurs, la représentation vectorielle des clients à l'échelle d'une grande surface ($1,5 \times 10^6$ produits dans le calcul de voisinages) conduit à une projection très peu dense et difficilement interprétable en termes de segments. De nouveau, les mesures de similarité sémantique vont permettre d'introduire une notion d'abstraction qui résout ce problème.

Cette étude s'inscrit dans ce contexte et propose une comparaison des performances d'algorithmes de segmentation en fonction des distances et mesures de similarité utilisées. A des fins de validation, cette problématique est transposée dans le domaine du biomédical où jeux de données et structurations de la connaissance sont davantage disponibles et formalisés. Le protocole expérimental [2] décrivant le corpus, la métrique d'évaluation et la méthodologie élaborée sont exposés en section 2. Les résultats et la discussion sont détaillés dans la section 3.

2 Protocole expérimental

2.1 Description du corpus et métrique d'évaluation

À notre connaissance, aucun jeu de données public issu de la grande distribution n'est disponible pour réaliser des expérimentations. Toutefois, des jeux de données provenant du domaine biomédical proposent des similarités fortes avec les données de la grande distribution. Le jeu de données recherché doit représenter un ensemble d'objets (*i.e.* nos clients) au travers d'un ensemble de concepts hiérarchisés au sein d'un ordre partiel (*i.e.* nos produits) pour se conformer à la particularité de notre problématique (l'ordre partiel n'est autre qu'une hiérarchie d'abstraction dans laquelle les produits vendus seraient les concepts les plus spécifiques). En outre, la validation de nos approches dans un contexte multi-classes implique également qu'une étiquette unique soit associée à chacun de nos objets (*i.e.* l'étiquette de leur segment).

Les travaux de Zhou et al. (2014) proposent une liste de maladies/symptômes désambiguïsés dans la taxonomie MeSH [3]. L'analogie avec notre problématique du retail s'opère de la façon suivante : un symptôme est à la maladie ce qu'un produit est au consommateur. Ainsi, chaque maladie peut être vue comme un vecteur de symptômes, et un client peut être perçu comme un vecteur de produits achetés. La désambiguïsation des maladies et des symptômes dans la taxonomie du MeSH (c'est-à-dire la transposition des maladies et des symptômes dans

1. Ce type de représentation est, par exemple, utilisé dans le cadre de la recherche de règles d'association où les colonnes correspondent aux produits du magasin et les lignes aux différents clients.

2. Afin d'assurer la reproductibilité des évaluations, elles sont réalisées sur un jeu de données public et le code développé est mis à disposition à l'adresse suivante : https ://github.com/PAJEAN/diseases_segmentation.

3. Le *MeSH* pour *Medical Subject Headings*, est le thésaurus de référence dans le domaine biomédical. Il est notamment utilisé pour indexer les articles de PubMed.

la structure hiérarchique du MeSH) permet à la fois d'appliquer les mesures de similarité sémantique pour analyser la ressemblance/différence entre deux maladies et d'attribuer une étiquette unique aux maladies par le biais de leurs concepts plus abstraits. Ainsi, l'étiquetage des maladies sous un même concept abstrait permet d'évaluer la pertinence du partitionnement qui a été réalisé : deux maladies partageant plusieurs symptômes seront déclarées proches et devraient porter la même étiquette abstraite (comme deux consommateurs qui achètent des produits similaires devraient être classés dans un même segment de clientèle).

Le jeu de données finalement utilisé contient 1517 maladies et 223 symptômes. Les différentes méthodes de partitionnement sont évaluées et comparer à l'aide de la F_1-mesure. Cette mesure d'évaluation se base sur la moyenne harmonique entre la précision et le rappel calculés sur toutes les paires de maladies du jeu de données (Hatzivassiloglou et McKeown, 1993).

2.2 Description des expérimentations

Les expérimentations réalisées ont pour objectif d'étudier les performances de mesures sémantiques dans le cadre d'une problématique de segmentation. Le protocole expérimental mis en place s'appuie sur un partitionnement ascendant hiérarchique qui tient compte des similarités sémantiques. Leur regroupement est réalisé par la méthode de Ward qui minimise la distance à l'intérieur des groupes (distance *intra*-groupes) tout en maximisant la distance entre les groupes (distance *inter*-groupes) (Murtagh, 2014). Dans ce protocole, les performances obtenues avec ces partitionnements sont comparées à deux méthodes de référence : le *K-means* et l'utilisation de métriques sur des espaces vectoriels couplés au partitionnement ascendant hiérarchique.

***K-means* et mesures vectorielles** Dans les travaux de Zhou et al. (2014) chaque objet (maladie) est représenté par un vecteur de réels sur l'ensemble des concepts (symptômes). Chaque composante de ce vecteur mesure la force d'association du concept avec l'objet. Cette force d'association réelle est calculée sur la base d'un TF-IDF (Sparck Jones, 1972). À partir de ces vecteurs d'observation, les auteurs calculent la distance entre les objets avec une distance cosinus. Pour des résultats plus exhaustifs, la distance euclidienne est également expérimentée et une restriction exploitant uniquement des vecteurs binaires (la force d'association existe ou non) pour un objet donné est aussi proposée.

Mesures de similarité sémantique Les similarités sémantiques entre les objets sont calculées à partir d'une structuration taxonomique. Ces similarités comparent des groupes de concepts associés aux objets par le biais de mesures dites *groupwise* (Harispe et al., 2015). Ces mesures se basent elles-mêmes sur des mesures dites *pairwise* permettant de calculer la similarité entre deux concepts au sein de la taxonomie. Certaines d'entre elles exploitent le contenu informationnel (*Information Content*, IC) associé aux concepts, c'est-à-dire, la quantité d'information associée à un concept (plus un concept est spécifique, plus son contenu informationnel est grand). Il existe plusieurs mesures *groupwise*, comme il existe plusieurs mesures *pairwise* et plusieurs définitions pour l'IC. L'objectif de cette étude est de présenter les performances et l'intérêt d'approches sémantiques.

Les mesures *groupwise* permettent la comparaison des ensembles de concepts rattachés à des objets. Il en existe deux catégories principales : les mesures directes et indirectes. Les me-

sures *groupwise* directes comparent les ensembles de concepts sans tenir compte de leur position dans la taxonomie. Pour cette étude, la distance de Jaccard a été mise en place. Concernant les mesures *groupwise* indirectes, elles agrègent les similarités obtenues des mesures de similarités *pairwise*. Cette étude exploite la BMA pour *Best Match Average* (Pesquita et al., 2007). Pour les mesures de similarité *pairwise*, la littérature en relate deux principaux types : les mesures basées sur le contenu informationnel (IC) et, celles basées sur la notion de plus court chemin dans la taxonomie (Harispe et al., 2015). Dans cette étude, les mesures de similarité *pairwise* mises en place sont respectivement la mesure de Resnik (Resnik, 1995) et la mesure de Wu & Palmer (Wu et Palmer, 1994). Enfin, concernant le contenu informationnel (IC), nous distinguons les IC intrinsèques et extrinsèques. Les IC intrinsèques prennent en compte uniquement les propriétés topologiques de la structure taxonomique du graphe sémantique. Ce type d'IC est généralement lié à la position d'un concept dans la taxonomie. L'IC intrinsèque utilisé pour cette étude est celui de Seco (Seco et al., 2004). Pour les IC extrinsèques, introduits par Resnik (Resnik, 1995), ils étendent l'approche intrinsèque en prenant également en considération la fréquence d'un concept dans une base d'observation (*e.g.* corpus). L'IC extrinsèque de Resnik est utilisé pour cette étude. Dans le cadre des expérimentations, les objets sont les maladies et les symptômes, les concepts. Pour plus d'information concernant les mesures de similarité sémantique mises en œuvre grâce à la *Semantic Measures Library*, le lecteur est invité à se référer aux travaux de Harispe et al. (2014).

3 Résultats et discussion

Le tableau 1 présente les résultats obtenus sur le corpus établi pour cette étude (cf. sous-section 2.1). A noter que les vecteurs d'observation du *K-means* et les matrices de distance dans le cadre du partitionnement ascendant hiérarchique sont soit, exploités tels quels, soit, *normalisés* pour observer l'impact de la normalisation sur le processus de segmentation.

	F_1-mesure	F_1-mesure (*normalisées*)
K-Means		
Vecteurs binaires	0.114	0.156
Vecteurs TF-IDF	0.113	0.136
Mesures vectorielles		
Vecteurs binaires, distance euclidienne	0.078	0.074
Vecteurs TF-IDF, distance euclidienne	0.104	0.094
Vecteurs binaires, distance cosinus	0.104	0.109
Vecteurs TF-IDF, distance cosinus	0.123	0.140
Mesures de similarité sémantique		
Jaccard	0.086	0.083
Wu & Palmer, BMA	0.108	0.124
IC Seco, Resnik, BMA	0.104	0.127
IC Resnik, Resnik, BMA	0.127	**0.182**

TAB. 1 – *Résultats des partitionnements.*

La F_1-mesure permet de mesurer la performance d'une configuration donnée à regrouper des maladies à partir de leurs symptômes sous une même étiquette abstraite (classe de maladies). A partir des résultats obtenus, nous pouvons dresser deux constats. Le premier repose sur la normalisation des vecteurs d'observation et de la matrice des distances. Dans la majorité des cas, le processus de normalisation a un impact positif et significatif sur la F_1-mesure. Ces résultats démontrent l'importance d'un tel processus lors d'une phase de segmentation. Le second constat, quant à lui, porte sur la pertinence des mesures de similarité sémantique au sein d'un processus de partitionnement appliqué avec des objets caractérisés par un ensemble de concepts structurés au sein d'un ordre partiel. Les résultats montrent que les mesures de similarité sémantique sont plus performantes avec la BMA associée à la mesure *pairwise* de Resnik et l'IC de Resnik avec une amélioration de la F_1-mesure de 16% au regard de ceux obtenus avec le *K-means*. Nous sommes conscients que ces résultats sont spécifiques aux particularités du jeu de données (typologie du MeSH, nombre de symptômes associés aux maladies, nombre de symptômes différents dans le jeu de données). Toutefois, ils permettent d'apporter une réponse objective à l'intérêt des mesures de similarité sémantiques dans un processus de segmentation où les données sont structurées par un ordre partiel.

Pour reprendre l'analogie avec le domaine du retail, l'utilisation de cette méthodologie permettra de regrouper les clients en fonction de leurs habitudes d'achats. Les mesures de similarité sémantique dans le processus de partitionnement apporteront des informations plus intuitives permettant la proposition d'assortiments et de services adaptés à un segment de clientèle.Ce partitionnement dépendra alors des produits achetés et de la structuration de la taxonomie de produits (*e.g.* chewing-gums et bonbons sont des friandises). Ces informations permettront, pour un preneur de décision, de mieux connaître sa clientèle et de proposer des stratégies de fidélisation adéquates.

4 Conclusion et perspectives

Dans cet article, nous proposons une alternative potentielle aux méthodes de segmentation de clients, qui se basent sur des métriques appliquées à des espaces vectoriels, en considérant les relations de similarité pouvant exister entre les dimensions de ce même espace. Ainsi, nous mettons en avant l'intérêt d'employer une connaissance *a priori* au travers des mesures de similarité sémantique afin d'exploiter les liens existants entre les caractéristiques des objets comparés. Un parallèle avec le domaine biomédical qui offre des jeux de données structurées nous a permis d'amorcer la validation de la pertinence de notre partitionnement sémantique. Les expérimentations nous ont permis d'observer de meilleures performances associées aux mesures de similarité comparativement aux méthodologies de segmentation *classiques*. En terme de perspective, nous allons employer les mesures de similarité sémantique sur des données réelles issues de la grande distribution pour identifier les habitudes d'achats des consommateurs : par exemple, différencier les consommateurs qui viennent principalement pour des produits alimentaires de ceux qui achètent des produits ménagers. Pour aller plus loin, nous envisageons également de travailler sur l'évolution des segments pour être en mesure d'identifier les changements d'habitudes grâce à l'analyse des leurs trajectoires (Gaffney et Smyth, 1999) et prédire les nouvelles tendances (*e.g* vegan).

Références

Chen, Y. L., M. H. Kuo, S. Y. Wu, et K. Tang (2009). Discovering recency, frequency, and monetary (rfm) sequential patterns from customers' purchasing data. *Electronic Commerce Research and Applications 8*(5), 241–251.

Ching-Hsue, C. et C. You-Shyang (2009). Classifying the segmentation of customer value via rfm model and rs theory. *Expert Systems With Applications 36*(3), 4176–4184.

Gaffney, S. et P. Smyth (1999). Trajectory clustering with mixtures of regression models. *KDD 99*(2), 63–72.

Harispe, S., S. Ranwez, S. Janaqi, et J. Montmain (2015). Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*.

Harispe, S., D. Sánchez, S. Ranwez, S. Janaqi, et J. Montmain (2014). A framework for unifying ontology-based semantic similarity measures : a study in the biomedical domain. *Journal of Biomedical Informatics 48*, 38–53.

Hatzivassiloglou, V. et K. R. McKeown (1993). Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. *Association for Computational Linguistics*, 172–182.

Murtagh, F. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification 31*, 274–295.

Pesquita, C., D. Faria, H. Bastos, A. Falcao, et F. Couto (2007). Evaluating go-based semantic similarity measures. *Proc 10th Annual Bio-Ontologies Meeting 37*(40).

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of IJCAI-95*, 448—-453.

Seco, N., T. Veale, et J. Hayes (2004). An intrinsic information content metric for semantic similarity in wordnet. *16th European Conference on Artificial Intelligence*, 1–5.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation 28*, 11–21.

Wu, Z. et M. Palmer (1994). Verb semantics and lexical selection. *32nd. Annual Meeting of the Association for Computational Linguistics*, 133—-138.

Zhou, X., J. Menche, A. L. Barabási, et A. Sharma (2014). Human symptoms–disease network. *Nature communications 5*, 4212.

Summary

Segmenting approaches used to group objects that share similar features are numerous. On the hypothesis that characteristics are independent and defined in a metric space, conventional distances are used. When these characteristics are instead organized in a representation of knowledge, these metrics become questionable. This article proposes the comparison of distances within a hierarchical ascending partitioning approach. The study aims to highlight the interest of semantic approaches to detect customers behavior. A parallel with the biomedical field was made to overcome the lack of data related to the retail sector.