Actes de l'atelier

Exploration des traces dans un monde du tout numérique : enjeux et perspectives

INFORSID 2023

La Rochelle, 30 mai 2023

Comité de programme :

Damien Mondou	L3i - La Rochelle Université	damien.mondou@univ-lr.fr
Ronan Champagnat	L3i - La Rochelle Université	ronan.champagnat@univ-lr.fr
Didier Vye	UMR LIENSs - La Rochelle Université	didier.vye@univ-lr.fr
Cyril Faucher	L3i - La Rochelle Université	cyril.faucher@univ-lr.fr

https://inforsid2023.sciencesconf.org/resource/page/id/18

Présentation de l'atelier

L'atelier Exploration des traces dans un monde du tout numérique : enjeux et perspectives a pour objectif principal de présenter des travaux innovants sur l'analyse et l'extraction d'informations issues des traces numériques que nous laissons tous derrière nous lors de nos interactions avec les technologies numériques. L'atelier explore les différentes formes de traces numériques, que ce soit lié à la robotique, l'e-éducation, la mobilité, les réseaux sociaux et bien d'autres encore.

Les participants ont l'opportunité d'explorer les enjeux et les défis liés à la collecte, au traitement et à l'analyse de ces données. Ils sont également amenés à discuter des perspectives futures de la collecte et de l'utilisation des traces numériques, ainsi que de l'impact potentiel sur les individus et les organisations. L'atelier s'appuie sur des méthodes de recherche innovantes, telles que l'analyse de données massives, l'apprentissage automatique, l'IA explicable et la visualisation de données. Cet atelier s'intéresse également à une ouverture aux enjeux des traces numériques dans le cadre de travaux de recherche en géographie.

Comité de programme

Damien Mondou	L3i - La Rochelle Université	damien.mondou@univ-lr.fr
Ronan Champagnat	L3i - La Rochelle Université	ronan.champagnat@univ-lr.fr
Didier Vye	UMR LIENSs - La Rochelle Université	didier.vye@univ-lr.fr
Cyril Faucher	L3i - La Rochelle Université	cyril.faucher@univ-lr.fr

Table des matières

Identifier les traces pertinentes dans la documentation scientifique d'une entreprise à des fins de communication institutionnelle

Le cas de la Compagnie TotalEnergies

Charlotte Darricades[1,3], Christian Sallaberry[2], Sébastien Laborie[2], Eric Kergosien[1], et Patrice De La Broise[1]

[1] Universite de Lille, GERiiCO
charlotte.darricades@univ-lille.fr, eric.kergosien@univ-lille.fr, patrice.de-la-broise@univ-lille.fr
[2] Universite de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA christian.sallaberry@univ-pau.fr, sebastien.laborie@univ-pau.fr
[3] Pôle d'Études et de Recherche de Lacq, TotalEnergies SE, BP 47, 64170 Lacq
charlotte.darricades@totalenergies.com

Résumé : Nous faisons l'hypothèse que des traces d'activités thématiques présentes dans les productions scientifiques d'une entreprise (le cas de TotalEnergies) sont utiles à des fins de développement social d'une entreprise en pleine mutation. Nous présentons une nouvelle approche permettant d'identifier ces traces par l'exploitation d'une cartographie qui fait état de l'organisation de l'entreprise, la mise en oeuvre d'entretiens avec les chercheurs et les communicants. Cela nous permettra ensuite de construire un premier squelette d'ontologie comme le préconise la méthodologie SAMOD. Enfin, nous détecterons et analyserons des traces dans les corpus documentaires scientifiques.

Mots clés : Corpus textuels, traces d'activité scientifique, communication institutionnelle

1 Introduction

Afin de gagner l'adhésion de ses collaborateurs et du grand public face à ses mutations, chaque entreprise doit bâtir une stratégie de communication. Mettre la R&D au cœur de cette stratégie pourrait permettre d'atteindre efficacement cet objectif et permettrait d'ancrer les projets de R&D dans de nouveaux processus de communication. Les grandes entreprises qui disposent de services R&D engendrent une grande variété de documents scientifiques. C'est le cas par exemple pour la Compagnie TotalEnergies où différents départements de la R&D produisent des publications scientifiques, des rapports d'études, des brevets, etc. En parallèle à cette activité, les communicants doivent mettre en valeur l'activité de Recherche dans la stratégie de communication institutionnelle. En effet, dans une démarche de valorisation de la R&D d'une entreprise, le service Communication du Pôle d'Études et de Recherche de Lacq (PERL) de la Compagnie doit pouvoir faire émerger des données pertinentes de façon automatique, comme par exemple des thématiques en lien avec la transition énergétique telles que la Capture et le Stockage du Carbone (CCS), Biogaz, Agrivoltaisme, etc.

D'un côté, l'information scientifique et technique (IST) est indispensable au travail des chercheurs, plus particulièrement à la construction de leur communication scientifique [4]. D'un autre côté, elle est également indispensable pour la stratégie de communication générale d'une entreprise. Les traces dans les productions scientifiques quelles qu'elles soient peuvent par conséquent être utiles à la rédaction d'articles de stratégie de communication. Se pose alors les problématiques suivantes : Quelles sont les différentes traces pertinentes liées aux productions scientifiques ? Comment les communicants peuvent-ils exploiter ces traces d'informations de R&D qui leurs seront utiles à des fins de stratégie de communication institutionnelle ?

Dans cet article, nous proposons d'expérimenter une nouvelle approche globale permettant d'identifier des traces : de la prise en compte de l'organisation de l'entreprise jusqu'à la production d'indicateurs pour les communicants. Cette approche débute par l'exploitation d'une cartographie

qui fait état de l'organisation de l'entreprise, des flux et des documents scientifiques existants. Une méthode de travail est ensuite déclinée pour permettre l'analyse du corpus documentaire.

Dans la section 2, nous allons tout d'abord aborder les méthodologies de construction d'une base de connaissance. Puis, dans la section 3, nous proposons une approche qui se décline en 2 phases distinctes, un état des lieux sous forme de cartographie suivi d'entretiens avec les chercheurs et les communicants. Enfin, nous conclurons en présentant quelques perspectives dans la section 4.

2 Travaux connexes : méthodologies pour la construction d'une base de connaissance

Nous cherchons à créer une ontologie métier pour représenter les connaissances contenues dans les différents documents scientifiques produits par les équipes du PERL et que différents types d'utilisateurs seraient susceptibles de rechercher (communicants et chercheurs du PERL principalement). Nous avons ainsi besoin d'une méthodologie de création d'une ontologie métier extensible, qui prenne en compte les besoins des utilisateurs experts. Un nombre important de travaux propose une méthodologie pour construire une ontologie. Parmi ceux-ci, nous pouvons notamment citer les méthodes Tove [5], Methontology [3], Sensus [9], Otk [7], Terminae [1], NeOn [8] ou encore Samod [6]. Toutes ces méthodes commencent par une phase d'acquisition de connaissances du domaine ou du métier, de rédaction de spécifications fonctionnelles ou de questions de compétence. Methontology et Otk sont très similaires. Les deux commencent par l'acquisition de connaissances et la rédaction de spécifications. Elles se poursuivent en modélisant le domaine d'abord d'une manière informelle puis dans un langage formel. Enfin, les deux proposent une évaluation de l'ontologie produite. Methontology recommande d'ailleurs un guide d'évaluation publié dans un document annexe. NeOn et Samod proposent des consignes pour créer des ontologies modulaires. Certaines méthodologies, notamment Otk, NeOn et Samod proposent un développement modulaire de l'ontologie qui est construite petit à petit, soit en ajoutant, à chaque itération, la modélisation d'une partie supplémentaire du métier/domaine, soit en modélisant toutes les parties du métier/domaine d'abord, puis en les fusionnant. NeOn se distingue des autres méthodologies présentées car elle fournit de nombreuses approches pour élaborer une ontologie ou un réseau ontologique. Elle demande aux ontologues de réaliser préalablement une analyse approfondie du projet afin de pouvoir choisir la bonne combinaison des processus et activités proposés. Les trois méthodes Otk, NeOn et Samod intègrent également une phase d'évaluation de l'ontologie produite, et cela lors de l'étape finale. Samod semble ressortir du lot selon nos critères car elle propose une première étape permettant de créer un premier squelette d'ontologie à partir des besoins exprimés par les utilisateurs cibles. La méthode préconise d'impliquer fortement les experts concernés afin de préciser et d'étendre les besoins, et le modèle ontologie produit, de façon itérative. L'aspect itératif impliquant les experts est primordial dans un secteur spécifique tel que le nôtre. Enfin, Samod intègre des phases de tests à différentes étapes du processus. À termes, nous prévoyons ainsi de formaliser des requêtes informelles exprimées par les experts en requêtes SPARQL afin de tester à la fois le modèle ontologique, et celui-ci une fois peuplé par les données collectées dans les différentes sources de données. Nous sommes encore dans une phase d'analyses et de tests des différentes méthodes existantes, et nous confirmerons notre choix une fois ce travail d'analyse terminé.

3 Contribution : Construction d'une méthodologie test issue de différentes méthodes existantes

Dans un premier temps, nous allons produire un état des lieux à travers un sociogramme (cartographie) qui fait état de l'organisation actuelle du PERL et de sa structure documentaire. Il permet d'en faire ressortir ses productions scientifiques avec un corpus de documents hétérogènes. Puis, nous allons identifier les besoins des utilisateurs, soit des chercheurs et des communicants des différents départements de la R&D en organisant des entretiens avec eux. Cette méthode nous permettra depuis une cartographie de produire un plan et une base de données afin d'offrir un outil utile aux communicants. Ces premiers travaux nous permettront ensuite de construire un premier squelette d'ontologie comme le préconise la méthodologie SAMOD. La particularité ici est que nous travaillons à l'enrichissement d'une cartographie au fur et à mesure des entretiens menés, afin

d'identifier et formaliser précisément les concepts à modéliser, les données du corpus à mobiliser pour instancier l'ontologie, et les cas d'usages pour tester la robustesse de la base de connaissance produite. La première étape de la méthode SAMOD consiste à collecter et formaliser les besoins des experts. Ils sont présentés sections 3.1 et 3.2. L'étape 2 de la méthode est détaillée section 3.3 et la dernière étape de tests est présentée section 3.4.

3.1 Etat des lieux

La figure 1, en annexes, présente donc un extrait de cette cartographie. Ainsi, le PERL dépend de plusieurs directions dans l'organigramme de TotalEnergies. De nombreuses informations au sujet de la R&D circulent entre les directions. Chaque direction (Direction OT, R&D Lines, UP Line, PERL) possède un service de communication représenté ici par 2 emojis bleu et rouge.

Le PERL a une communauté de chercheurs considérable (environ 80 chercheurs) dans plusieurs domaines. Beaucoup de documents scientifiques circulent tels que le rapport annuel, les rapports de projets de R&D, les brevets et les demandes d'inventions, les notes de synthèse R&D et les publications scientifiques.

Prenons l'exemple des publications scientifiques. Nous allons donc nous intéresser à cette partie de la cartographie afin d'observer les types de publications scientifiques existants rédigés par les chercheurs PERL. Premièrement, il y a les articles dits « normaux », on peut dire que ces articles sont les plus communs dans le domaine de la communication scientifique. Ils exposent les résultats obtenus à l'issue d'une proposition de méthodologie et de son expérimentation. Deuxièmement, il y a les "articles de revue" où le chercheur établit un état de l'art relatif à sa problématique afin de donner une cohérence à son sujet. Troisièmement, il y les "articles de perspectives" où le chercheur-auteur va établir une revue en y ajoutant des pistes de solutions. Pour finir, les articles de "correction ou ajout" qui viennent compléter un article dit "normal" précédemment publié. À savoir que ces publications sont généralement rédigées en anglais ou en français, d'où les drapeaux bleu et rouge sur cette cartographie. Ainsi, le corpus de publications scientifiques est varié, il présente également une forte hétérogénéité de structures. De plus, il est chargé en traces thématiques de R&D qui pourraient potentiellement intéresser les communicants. Il est important de récolter toutes les traces afin de ne pas faire d'impasse et d'obtenir le plus de données sur la R&D.

3.2 Construction d'un premier modèle ontologique à partir d'entretiens

Afin de construire un premier modèle ontologique, nous organisons des entretiens semi-directifs en face à face avec des chercheurs de chaque service du PERL ainsi que des communicants de chaque direction précédemment évoquée. Nous allons donc préalablement préparer une grille d'entretiens et mettre en forme la première cartographie pour être le plus précis possible durant l'échange avec chaque expert. Ces entretiens auront tout d'abord pour but d'aligner le vocabulaire scientifique au vocabulaire des communicants. Ensuite, ils nous permettront de relever l'ensemble des traces qui sont utiles selon le chercheur et selon le communicant pour la stratégie de communication institutionnelle, et enfin de formaliser de façon claire les attentes des communicants en termes de vulgarisation de la communication scientifique. Ces entretiens semi-directifs nous offrent ainsi l'opportunité de connaître le rôle des chercheurs et l'influence de leurs publications dans la stratégie de communication institutionnelle. À partir de l'ensemble de ces éléments, nous allons pouvoir construire un premier squelette de l'ontologie, que nous pourrons ensuite enrichir et instancier.

3.3 Extraction d'informations pour l'instanciation de l'ontologie

Comme nous l'avons décrit précédemment, nous disposons d'un corpus documentaire hétérogène. Également, nous disposons d'une cartographie qui permet de connaître les producteurs de ces documents ainsi que leurs relations entre eux (ex. relations hiérarchiques, relations de collaborations...). Enfin, le communiquant désireux d'analyser le corpus dispose d'indicateurs qui vont lui permettre d'ajuster sa politique de communication. Une première stratégie pour produire ces indicateurs consisterait à indexer en amont tout le corpus documentaire sans tenir compte d'un contexte d'analyse donné. Il va de soi que cette méthode n'est pas efficace car d'une part elle pourrait produire des indicateurs non-utiles au communiquant à un instant t, et d'autre part certains

indicateurs pourraient ne pas être pertinents lorsqu'ils sont appliqués de façon globale sur certains types de documents. Par conséquent, tenant compte des différents types de documents ainsi que de notre cartographie, nous proposons plutôt une seconde stratégie qui consiste à cibler l'indexation en fonction d'un contexte d'analyse du communiquant. Par exemple, si le communiquant désire avoir une vue globale de son corpus au sujet des tendances de thématiques abordées au sein de son organisation, un indexeur spécifique pourra être sélectionné plutôt que tous les indexeurs possibles (ex., un extracteur de concepts) et certaines parties de documents pourront être analysées plus particulièrement par cet indexeur (ex., les titres et les résumés des articles scientifiques, les références bibliographiques pour les rapports d'activités...). En effet, il existe actuellement de nombreux outils d'indexation de documents, notamment textuels, allant d'outils standards "clé en main" jusqu'à des outils configurables par des experts informaticiens. En voici quelques exemples :

- Les outils standards "clé en main" : Voyant[1] , VOSViewer[2] , Bibliometrics[3] , Sketchengine [4], Cortext[5], Gargantext[6], etc. Il s'agit globalement d'outils dédiés aux spécialistes de la langue qui désirent, via une interface graphique adaptée, pratiquer l'extraction terminologique, l'alignement multilingue, la visualisation des occurrences de termes en contexte ou encore l'édition de ressources linguistiques.
- Les outils avancés "paramétrables" : TextRazor[7], Lexalytics[8], etc. Ils combinent des techniques de traitement du langage naturel avec des bases de connaissances pour extraire des entités informationnelles dans des documents, tweets ou pages web. Ils peuvent être utilisés en version standard ou avec un minimum de paramétrage, tout comme certains proposent des librairies de fonctions intégrables dans des programmes ad-hoc. En version standard, ils reconnaissent, par exemple, des organisations, des dates, des lieux, des prix, des adresses, des personnages célèbres, des œuvres d'art, etc.,
- Les outils experts "avec programmation" : Gate[9], Spacy[10], Google NLP[11], etc. Il s'agit de boîtes à outils logicielles utilisées pour le traitement du langage naturel dans différentes langues. Différents services sont mis à disposition des programmeurs d'applications dans des langages tels que Java et Python.

Voici un exemple d'article de recherche (https://bit.ly/3LDLDhx) qui, après traitement par des outils existants, présente déjà des résultats d'extraction intéressants en ciblant leurs traitements uniquement sur le résumé de l'article (voir les figures 2 et 3 en annexes).

En fait, nous avons pour objectif d'associer des outils d'annotation à notre base de connaissance métier relative à l'énergie et à l'environnement. Nous devrons nous appuyer sur des outils suffisamment ouverts qui intègrent de tels types de ressources externes. Dans le cadre d'un premier travail de veille, nous avons identifié TextRazor et Gate qui proposent ce type de service. Le travail d'annotation devra nous aider à peupler semi-automatiquement l'ontologie produite en amont.

3.4 Outils d'analyses et de visualisation

Dans le cadre de notre projet, chaque communicant devrait pouvoir construire son propre tableau de bord, afin d'avoir une vision globale du corpus documentaire lui facilitant, par la suite, sa prise de décision. Pour ce faire, nous pourrions :

- exploiter les fonctionnalités complémentaires des outils d'extraction identifiés plus haut comme, par exemple, Voyant qui permet l'affichage de nuage de mots et/ou VOSViewer qui présente des réseaux bibliométriques,

[1] https://voyant-tools.org/
[2] https://www.vosviewer.com/
[3] https://www.bibliometrix.org
[4] https://www.sketchengine.eu/
[5] https://www.cortext.net
[6] https://www.gargantext.org
[7] https://www.textrazor.com/
[8] https://www.lexalytics.com/
[9] https://gate.ac.uk/
[10] https://spacy.io/
[11] https://cloud.google.com/natural-language

– ou bien, sur la base d'extractions réalisées en amont, utiliser des outils d'analyse et de visualisation externes comme, par exemple, les outils de la Business Intelligence tels que Tableau, Qlik ou Power BI. Notons l'émergence d'outils No-Code spécialisés dans le NLP : SimpleX (https://sx.simpledecisions.io/landing/about-us), par exemple, est une console de text mining dédiée au traitement et à la visualisation de données textuelles.

4 Conclusion / perspectives

Cette première méthode permettant d'identifier ces traces par l'exploitation d'une cartographie, avec la mise en œuvre d'entretiens pour construire un modèle ontologique, suivis d'une étape d'analyse des corpus documentaires scientifiques pour instancier ce modèle et obtenir une base de connaissance métier. Nous sommes actuellement dans la phase d'exploration à travers le processus de construction des connaissances avec en aval une réflexion sur la définition de l'objet de recherche (via la problématique des communicants et l'identification des sources pertinentes pour produire des supports de communication), et en amont les données (recueil des traces sur la R&D et traitement) ainsi que sur les choix finaux concernant le dispositif méthodologique [2]. Selon la méthodologie finale choisie, celle-ci pourrait permettre de saisir une nouvelle approche de la communication scientifique inter-métiers.

References

1. Nathalie Aussenac-Gilles, Sylvie Despres, and Sylvie Szulman. The TERMINAE Method and Platform for Ontology Engineering from Texts. In Paul Buitelaar and Philipp Cimiano, editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, volume 167 of *Frontiers in Artificial Intelligence and Applications*, pages 199–223. IOS Press, February 2008.
2. Sandra Charreire-Petit and Florence Durieux. Explorer et tester : les deux voies de la recherche. In Raymond-Alain Thiétart éd, editor, *Méthodes de recherche en management*, chapter 3, pages 76–104. Dunod, Paris, 2014.
3. Mariano Fernández-López, Asuncion Gomez-Perez, and Natalia Juristo. Methontology: from ontological art towards ontological engineering. *Engineering Workshop on Ontological Engineering (AAAI97)*, 03 1997.
4. Cécile Gardiès and Isabelle Fabre. Communication scientifique et traitement documentaire de l'ist. quelles méthodes du travail intellectuel ? *Les Cahiers du numérique*, 5:85–104, 2009.
5. Michael Grüninger and Mark Fox. Methodology for the design and evaluation of ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing*, 07 1995.
6. Silvio Peroni. Samod: an agile methodology for the development of ontologies. 01 2016.
7. S. Staab, R. Studer, H.-P. Schnurr, and Y. Sure. Knowledge processes and ontologies. *IEEE Intelligent Systems*, 16(1):26–34, 2001.
8. Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. *The NeOn Methodology for Ontology Engineering*, pages 9–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
9. William Swartout, Ramesh Patil, Kevin Knight, and Tom Russ. Toward distributed use of large-scale ontologies. *Association for the Advancement of Artificial Intelligence*, 01 1997.

Annexes

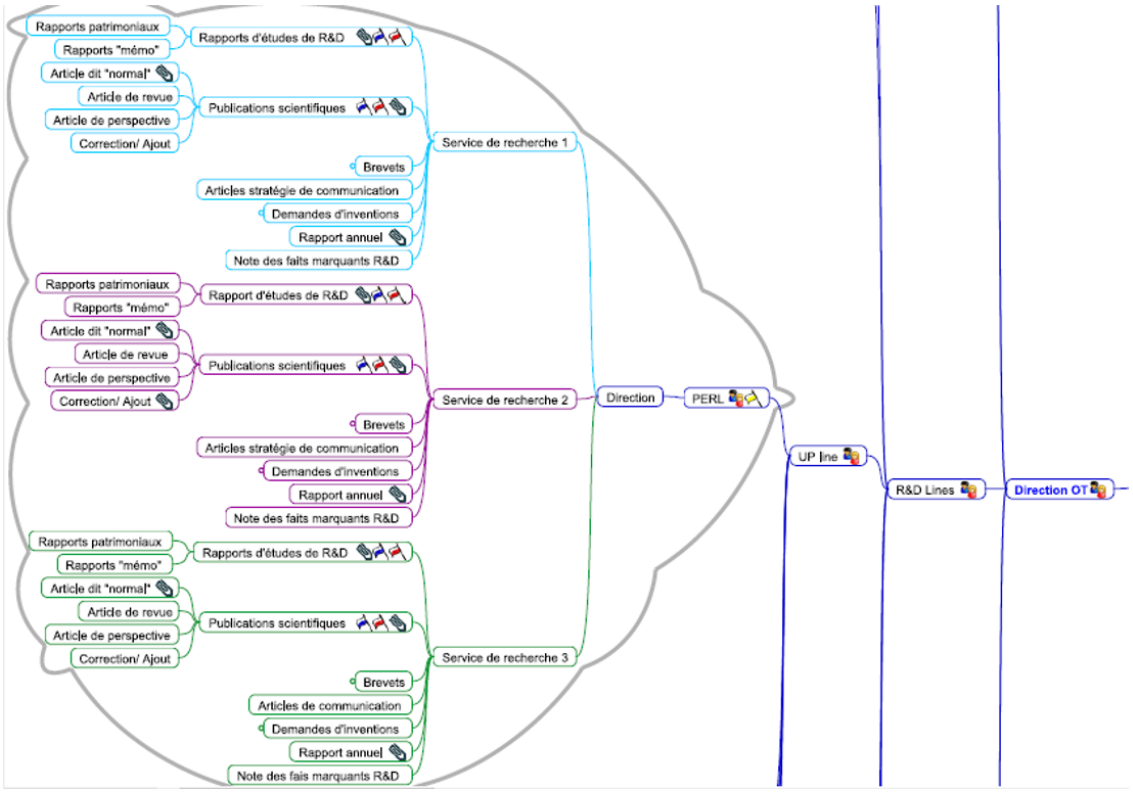

Fig. 1. Extrait de la cartographie du PERL

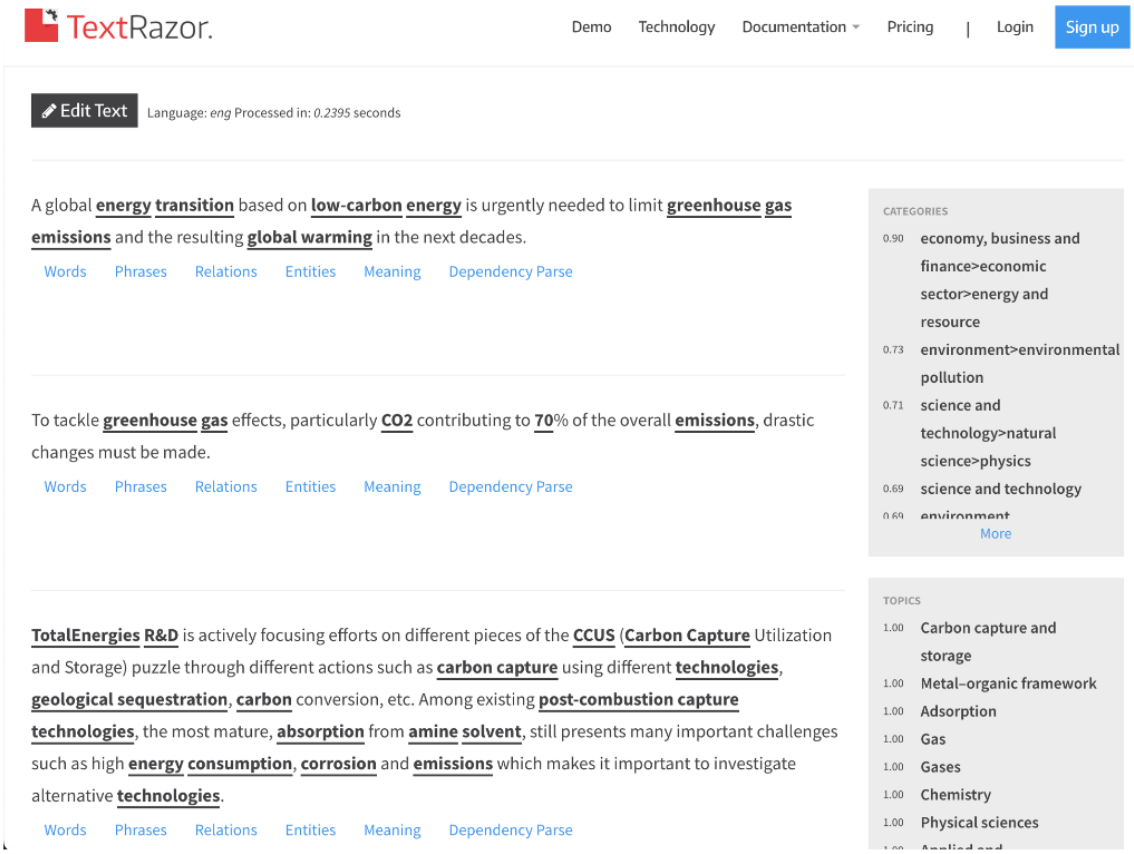

Fig. 2. Test de l'outil TextRazor

Fig. 3. Test de l'outil Voyant

Détecter des communautés sur Twitter et les analyser

Philippe Suignard, Mathieu Brugidou

EDF R&D, 7 boulevard Gaspard Monge 91120 Palaiseau
prenom.nom@edf.fr

Résumé : L'importance grandissante des sujets liés à l'énergie et sa place dans les réseaux sociaux ont motivé la R&D d'EDF à mieux comprendre qu'elle était l'influence de ces réseaux sociaux dans la formation des opinions publiques. Pour cela, nous nous sommes intéressés à ce qui s'était dit sur Twitter à propos de la thématique des "énergies renouvelables" pendant l'année 2021, année marquée par les élections régionales au mois de juin. Après avoir présenté la collecte des données, l'article présente une méthode pour détecter/visualiser des communautés, puis une méthode pour décrire ou analyser ces communautés. Enfin, l'article présente une série de perspectives et de travaux futurs.

Mots clés : réseau social, twitter, détection de communautés, énergie éolienne, opinion

1 Introduction

L'année 2021 a été marquée par la tenue des élections régionales les 20 et 27 juin et par la campagne électorale qui s'est focalisée, pour partie, sur l'écologie et plus particulièrement sur l'hostilité aux éoliennes, notamment dans le nord de la France [7]. Cette campagne s'est également déroulée sur les réseaux sociaux, avec de nombreux échanges entres personnes, associations, partis politiques, médias, etc. Toutes ces discussions peuvent être vues comme des "traces numériques", thème de cet atelier, qu'il convient d'analyser avec les méthodes et outils à notre disposition [5]. Dans cet article, nous nous intéressons plus particulièrement aux discussions sur Twitter sous l'angle de la détection de communautés, "groupes humains dont les membres sont unis par un lien social"[1]. Nous cherchons à détecter ces communautés pour ensuite les analyser.

2 La collecte de données

La collecte des tweets a été réalisée à l'aide de l'API fournie par Twitter[2]. Twitter propose différents types d'habilitations pour récupérer les tweets, l'idée générale étant que plus on paie et plus on peut récupérer de tweets. Dans notre cas, nous utilisons la version gratuite qui permet seulement de revenir jusqu'à 7 jours en arrière. Une collecte a donc été lancée automatiquement tous les jours pour récupérer les tweets émis la veille. Les tweets sont récupérés à l'aide du moteur de recherche de Twitter : il s'agit des tweets écrits en français contenant les mots liés au vocabulaire des énergies renouvelables[3].

Environ 900k tweets ont ainsi été émis par 150k comptes ou twittos différents. Les 100/1000/10000 twittos les plus prolifiques représentent respectivement 10% / 29% / 60% du volume total.

3 La détection des communautés

3.1 Notion de communautés sur les réseaux sociaux

Dans tout type de réseau social, les comptes ou individus sont connectés entre eux via différents types de connexions : connexions de type suiveur/suivi, connexions de type action (comme retweeter, liker ou répondre). A partir de ces informations, il est possible de constituer un graphe,

[1] Définition Wikipédia : https://fr.wikipedia.org/wiki/Communaut%C3%A9
[2] https://developer.twitter.com
[3] éolien(s), éolienne(s), hydraulique, transition énergétique, enr, photovoltaïque, panneau(x) solaire (s) ou biomasse

les sommets étant les comptes, les liens entre sommets étant les connexions entre les comptes (retweet, réponse, ...). A partir de ce graphe, il devient possible de calculer des communautés.

Une des approches les plus simples à mettre en oeuvre et les plus efficaces s'appuie sur le principe d'homophilie, c'est-à-dire la tendance qu'ont les individus à s'associer à des personnes similaires à elle-même [1]. En s'appuyant sur le graphe des retweets, il devient possible de détecter les personnes qui se "ressemblent" et ainsi détecter des communautés [6].

3.2 Détection des communautés en fonction de la morphologie du graphe des retweets

Une question importante consiste à savoir quels comptes considérer. En effet, il existe des comptes (ou des bots) qui ne font que retweeter des comptes spécifiques (comme les comptes politiques) pour faire monter leur importance. Pour limiter ce phénomène, on garde uniquement les comptes qui ont retweeté au moins 2 comptes différents et qui ont émis un tweet (qui ne soit pas un retweet). Ainsi le graphe obtenu est composé de **13 082** noeuds ou comptes liés par **86 249** liens. Les liens ne sont pas dirigés : si A retweete B "x" fois et B retweete A "y" fois, on considère qu'il y a un lien non dirigé entre A et B avec un poids de "x+y".

L'outil utilisé pour calculer les communautés et les visualiser est Gephi [2]. L'algorithme Louvain [3] permet de détecter ces communautés (les couleurs sur la Figure 1). Un algorithme de positionnement automatique de type force-ressort appelé "Force Atlas" intégré à Gephi permet de placer ensemble les comptes qui sont fortement liés entre eux. La taille des comptes est proportionnelle au calcul du PageRank [10], c'est-à-dire que plus les comptes sont importants (au sens du graphe) et plus leur taille sera grande (ce qui ne veut pas forcément dire qu'un compte qui émet beaucoup de tweets aura une grande taille).

Fig. 1. Carte des communautés de la thématique des "énergies renouvelables"

Ainsi, sept communautés principales sont identifiées sur la Figure 1, de la gauche vers la droite :

- en bleu : une communauté des pro éolien composée d'associations militantes pour le déploiement de l'éolien en France ;
- en orange : la communauté de la "France Insoumise", pro éolien également ;
- en marron foncé : la communauté liée au gouvernement et aux différents ministres concernés par l'énergie ;

- en vert clair : une communauté pro nucléaire et assez neutre sur les éoliennes ;
- en rouge : une communauté anti éolien et défenseur des paysages ;
- en violet : une communauté anti éolien, fortement à droite portée par des hommes politiques comme E. Zémmour, F. Philippot
- en vert foncé : une communauté également anti éolien, fortement à droite portée par M. Le Pen.

4 Qualification des communautés

La description des communautés faite au paragraphe précédent est réalisée de manière experte en allant regarder quelles personnes ou associations composent chaque communauté. On se propose ici d'effectuer une analyse plus fine et de manière automatique pour aider un analyste, par exemple. La démarche est la suivante :

- Pour une communauté donnée, il faut partir des comptes de cette communauté ;
- Récupérer les tweets qui sont émis par un membre de cette communauté et retweetés par un autre membre de cette communauté ;
- "Nettoyer" les tweets (éliminer les mentions et les URL), puis les découper en "token" ou mots;
- Extraire les associations saillantes de mots, grâce à la méthode statistique nPMI [4], pour "Normalized Pointwise Mutual Information". En résumé, la méthode cherche des mots A et B qui ont tendance à être d'avantage utilisés ensemble qu'avec d'autres mots, comme "transition énergétique". Un seuil permet de décider si deux mots doivent être agglutinés ou non. La méthode peut être réitérer pour trouver des chaines de mots plus longues ;
- Calculer la fréquence de ces chaînes de mots, pondérée par le nombre de retweets. Ainsi on obtient les chaînes de mots les plus caractéristiques pour une communauté donnée.

Une fois calculés, ces mots ou expressions, peuvent être visualisés à l'aide d'un logiciel de visualisation de nuages de mots[4].

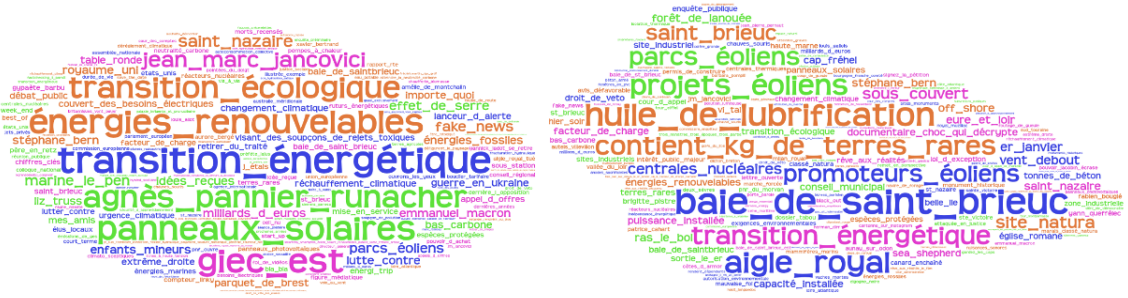

Fig. 2. Nuages de mots et d'expressions de la communauté pro "énergie renouvelable", à gauche et de la communauté "anti-éolien et défenseur des paysages", à droite.

La Figure 2 présente deux nuages de mots pour deux communautés très différentes : - les "pro énergie renouvelable" à gauche qui mettent en avant la "transition écologique/énergétique" nécessaire et l'installation de "panneaux solaires" et de "parc éoliens", pour lutter contre "le réchauffement/l'urgence climatique" ; - les "anti éoliens" à droite qui mettent en avant des "huile de lubrification" qui seraient contenues dans les moteurs des éoliennes, les "kg de terre rares" qu'il faut utiliser pour construire ces éoliennes, les dégâts sur la faune (aigle, cigogne) et qui dénoncent les nombreux "projets et/ou parcs éoliens" en construction sur l'ensemble du territoire comme en "baie de St-Brieuc" ou à "St Nazaire".

[4] https://wordart.com/create

5 Perspectives

Plusieurs perspectives sont envisagées pour améliorer et compléter toutes ces analyses comme :

- la détection des "narratifs" : les narratifs ou récits sont utilisés pour mieux diffuser une idée (vraie ou fausse d'ailleurs). Ces narratifs peuvent être des phrases courtes, des idées, des "punch-lines" et peuvent prendre des formes plus diverses que les simples arguments ;
- l'analyse des "Thread" : les threads ou fils de discussion sont des moyens pour un utilisateur de diffuser une information plus longue qu'un simple tweet [9]. D'un point de vue technique, il s'agit d'un tweet initial auquel son auteur va répondre (parfois jusqu'à une centaine de fois). Ces longs fils sont des moyens pour leur auteur d'exposer une théorie, pour "debunker" les tweets d'une communauté adverse, etc.
- l'aspect temporel : les tweets ont une connotation temporelle très forte qu'il conviendrait de mieux prendre en compte : détection des périodes de forts échanges, détection des périodes de renouvellement important du vocabulaire [8].

References

1. Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.
2. Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362, 2009.
3. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
4. Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.
5. Jean-Philippe Cointet. *La cartographie des traces textuelles comme méthodologie d'enquête en sciences sociales*. Habilitation à diriger des recherches, École normale supérieure, November 2017.
6. Michael D. Conover, Bruno Goncalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 192–199, 2011.
7. Tristan Haute and Marie Neihouser. Élections régionales en hauts-de-france, 20-27 juin 2021. *BLUE*, 2(1):42–47, 2022. Publisher: Groupe d'études géopolitiques.
8. Cyril Labbé, Dominique Labbé, and Pierre Hubert. Automatic segmentation of texts and corpora. *Journal of Quantitative Linguistics*, 11(3):193–213, 2004.
9. Julien Longhi. Le thread, un texte cousu de fil numérique ? *Le Français Moderne - Revue de linguistique Française*, March 2022.
10. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

Visualisation interactive de trajectoires d'activités touristiques

Application à des données extraites de Twitter

Maxime Masson, Siwar Abdelhedi, Christian Sallaberry, Rodrigo Agerri, Marie-Noëlle Bessagnet, Annig Lacayrelle, et Philippe Roose

Laboratoire LIUPPA, Collège STEE, Université de Pau et des Pays de l'Adour Avenue de l'Université, 64000 Pau, France
maxime.masson@univ-pau.fr, siwar.abdelhedi@univ-pau.fr, christian.sallaberry@univ-pau.fr

Résumé : Dans cet article, nous proposons une approche innovante pour l'analyse et la visualisation de données issues des réseaux sociaux. Nous avons conçu un tableau de bord interactif multidimensionnel que nous avons testé sur le domaine spécifique du tourisme.

Mots clés : Réseaux Sociaux, Visualisation de données, Traitement Automatique du Langage (TAL), Tourisme

1 Introduction

Ce travail est réalisé dans le cadre d'un projet régional transfrontalier : le projet APs (APs signifiant "Augmented Proxemics services"). Ce projet vise à collecter, traiter, analyser puis valoriser des données issues des réseaux sociaux, relatives à la pratique du tourisme, aux flux de visiteurs et à l'utilisation du patrimoine culturel dans la région du Pays Basque, un territoire hautement touristique situé entre la France et l'Espagne. Dans le cadre de ce projet, nous avons mis en place un cadre de travail générique pour le traitement et l'analyse de données issues des réseaux sociaux : **le framework APs**. Le cycle de vie de ce dernier est exposé dans la Figure 1. La phase de **collecte** (1) couvre l'ensemble du processus de recherche et d'extraction des données, c'est-à-dire qu'elle produit un corpus de posts de réseaux sociaux à partir d'une définition spécifique du domaine ciblé. Pour y parvenir, nous avons conçu une méthodologie de collecte générique et itérative. Elle a fait l'objet d'un article séparé, voir [2] pour plus de détails. La **transformation** (2) fait référence aux diverses modifications et enrichissements appliqués aux données collectées précédemment afin d'accroître leur valeur ajoutée et de les préparer pour les étapes suivantes. Les informations extraites au cours de cette étape (par exemple: le sentiment ou les entités nommées spatiales présentes dans les posts) instancient notre modèle de trajectoire, qui est l'élément central de notre cadre de travail. **L'analyse proxémique** (3) utilise le modèle précédemment instancié pour calculer les métriques proxémiques (appelées distances), qui sont des indicateurs bruts calculés en fonction des besoins. Enfin, la **valorisation** (4) permet de visualiser les résultats de l'analyse précédente (indicateurs bruts exprimés en métriques proxémiques) pour les utilisateurs finaux (tels que les acteurs du tourisme). Pour les professionnels du tourisme, des cartes multidimensionnelles permettent de visualiser les tendances et les associations de thèmes et de lieux dans les réseaux sociaux. Cet article sera principalement consacré à cette étape 4, qui constitue l'une des originalités du framework APs.

2 La visualisation interactive de trajectoires issues de Twitter

2.1 La visualisation en Business Intelligence (BI)

Les outils de visualisation de données en Business Intelligence sont des logiciels ou des plateformes qui permettent de créer des graphiques, des tableaux de bord, des rapports et des visualisations interactives à partir de données provenant de différentes sources, pour une compréhension rapide et facile. En BI, ces outils sont utilisés pour aider les entreprises à comprendre leurs données, à détecter des tendances, à prendre des décisions éclairées et à communiquer les résultats aux

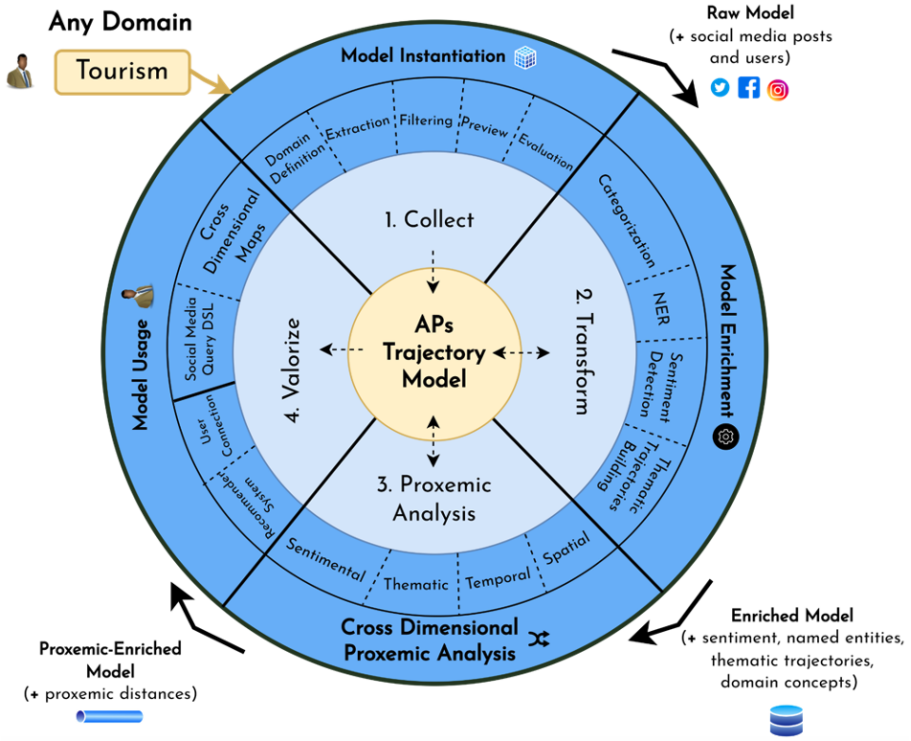

Fig. 1. Cycle de vie du Framework APs [1]

décideurs. Cet objectif correspond également aux attentes des professionnels du tourisme et de l'aménagement du territoire. De nombreux outils de visualisation de données sont disponibles en BI, parmi lesquels on peut citer le Tableau, Microsoft Power BI, SAP Lumira, Domo, Google Data Studio et QlikView. Notons également l'émergence d'outils No-Code spécialisés comme SimpleX qui, est une console de text mining dédiée au traitement et à la visualisation de corpus textuels.

2.2 Les systèmes d'information géographique

Un système d'information géographique (SIG), est une solution logicielle qui permet de visualiser et d'analyser des entités géographiques ainsi que les phénomènes qui y sont liés. Les attentes des professionnels du tourisme et de l'aménagement du territoire concernent très précisément ce besoin de représentation du territoire et des activités touristiques. Il existe plusieurs systèmes d'information géographique disponibles comme ArcGIS, QGIS, GeoServer, MapInfo, GRASS GIS et SAGA GIS. Notons également les travaux d'Aline Menin [3] et de Cécile Saint-Marc [4].

2.3 La proposition APs dédiée aux trajectoires

2.3.1 Visuel 4D : espace/temps/thème/individus

Nous avons proposé un tableau de bord multidimensionnel qui intègre les dimensions spatiales, temporelles, thématiques et individuelles. Cette interface permet d'afficher plusieurs visualisations telles qu'**une carte thématique** qui présente la fréquence des concepts dans les tweets, **une carte spatiale** qui montre la fréquence des villes dans les tweets, **une frise chronologique** (timeline) qui affiche le nombre de tweets par jour, subdivisé en matin (bleu), après-midi (orange) et soir (violet), et **une carte des utilisateurs** qui présente l'emprise temporelle des visites, le nombre de followers et le nombre de posts pour chaque utilisateur.

L'interface de notre tableau de bord multidimensionnel offre une interactivité avancée pour faciliter l'exploration des données. Au niveau de la localisation thématique, l'utilisateur peut cliquer sur un thème spécifique et mettre à jour les fréquences en fonction du thème sélectionné dans la carte spatiale et la frise chronologique. De même, au niveau de la localisation spatiale, la sélection

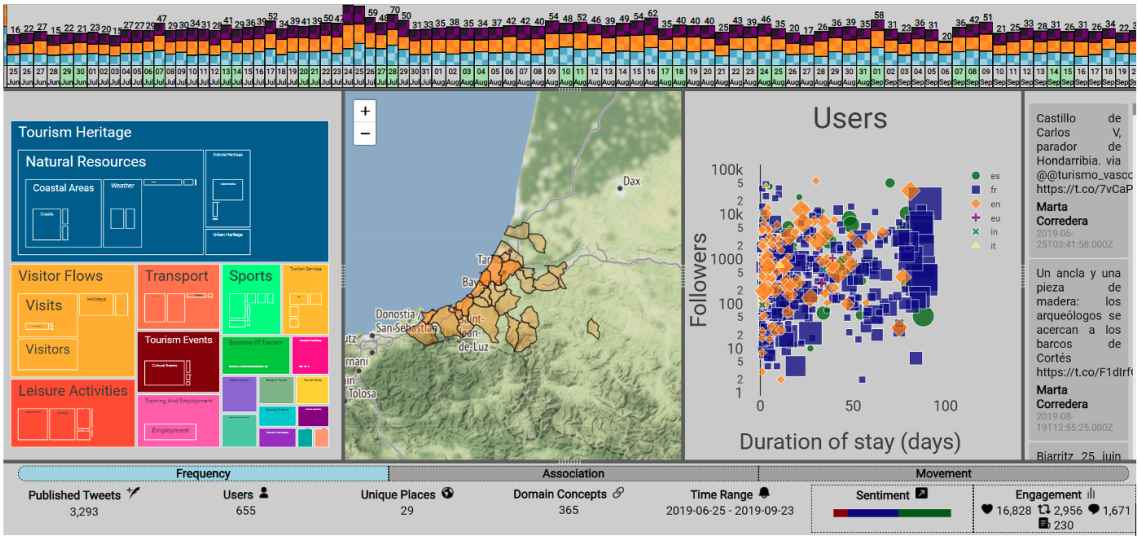

Fig. 2. Tableau de bord APs avec mode d'affichage "fréquence"

d'une ville spécifique met à jour les fréquences dans la carte thématique et la frise chronologique. Enfin, au niveau de la localisation temporelle, l'utilisateur peut sélectionner une plage temporelle spécifique avec une granularité fine, que ce soit par jour ou par section de jour (après-midi, matin, soir). Cette sélection permet de mettre à jour les fréquences correspondantes dans la carte thématique et la carte spatiale, et peut être combinée avec un nouveau filtrage spatial ou thématique pour affiner encore plus l'exploration des données.

2.3.2 Visuel à base de graphes d'associations

En mode association, la carte thématique est transformée en un graphe, dans lequel les nœuds représentent les concepts et leur taille est proportionnelle au nombre d'occurrences dans les tweets. Les arêtes représentent la fréquence d'association à laquelle les concepts sont trouvés dans le même post. De même, un graphe est superposé à la carte spatiale pour représenter les relations entre les lieux. Les nœuds représentent les lieux et leur taille est proportionnelle au nombre d'occurrences, tandis que les arêtes représentent la fréquence d'association à laquelle les lieux sont trouvés dans le même post.

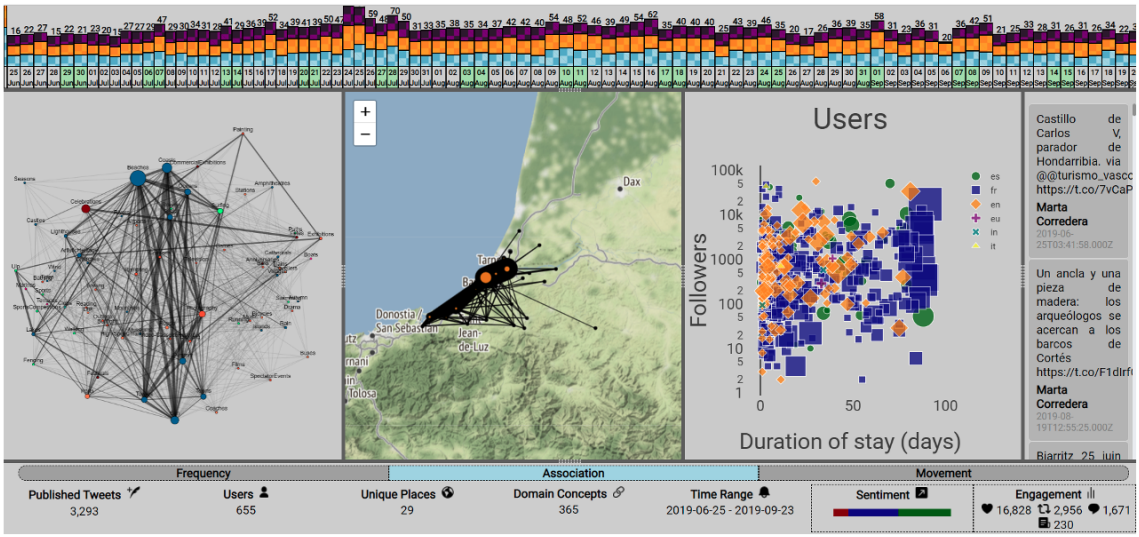

Fig. 3. Tableau de bord APs avec mode d'affichage "association"

3 Conclusion

Le projet APs nous a permis de développer une première série de tableaux de bords interactifs dédiés aux trajectoires d'activités touristiques. Nous travaillons à la conception d'une nouvelle plateforme sur laquelle l'utilisateur final pourra choisir, paramétrer et agréger différents visuels proposés sur une palette de services.

References

1. Maxime Masson, Philippe Roose, Christian Sallaberry, Rodrigo Agerri, Marie-Noelle Bessagnet, and Annig Le Parc Lacayrelle. Aps: A proxemic framework for social media interactions modeling and analysis. In Bruno Crémilleux, Sibylle Hess, and Siegfried Nijssen, editors, *Advances in Intelligent Data Analysis XXI*, pages 287–299, Cham, 2023. Springer Nature Switzerland.
2. Maxime Masson, Christian Sallaberry, Rodrigo Agerri, Marie-Noelle Bessagnet, Philippe Roose, and Annig Le Parc Lacayrelle. A domain-independent method for thematic dataset building from social media: The case of tourism on twitter. In Richard Chbeir, Helen Huang, Fabrizio Silvestri, Yannis Manolopoulos, and Yanchun Zhang, editors, *Web Information Systems Engineering – WISE 2022*, pages 11–20, Cham, 2022. Springer International Publishing.
3. Aline Menin. *eSTIMe : un environnement de visualisation pour l'analyse multi-points de vue des mobilités quotidiennes.* PhD thesis, Université Grenoble Alpes, 2020.
4. Cecile Saint-Marc. *Formalisation et géovisualisation d'événements historiques issus de risques naturels pour la compréhension des dynamiques spatiales : application aux inondations ayant touché le système ferroviaire français.* Theses, Université Grenoble Alpes, June 2017.

Fouille de processus pour l'amélioration d'un jeu sérieux

Sébastien Amoury et Karell Bertet

Laboratoire Informatique, Image et Interaction, La Rochelle Université, La Rochelle, France
sebastien.amoury@univ-lr.fr, karell.bertet@univ-lr.fr

Résumé : Nous évaluons, dans cet article, l'efficacité de différents algorithmes de fouille de processus pour les traces numériques recueillies lors d'interactions Homme-Robot dans le contexte d'un jeu sérieux, ainsi que l'apport des modèles générés par ces algorithmes sur l'amélioration des activités. L'extraction de modèles à partir des logs de chaque joueur permet d'entrapercevoir des comportements communs entre différents joueurs, mais aussi des comportements moins communs, voir uniques, des "pépites" propres à chaque joueur. Cela permet aussi de mettre en lumière les différents problèmes que les joueurs ont pu rencontrer (problèmes de reconnaissance vocale avec un robot, mauvaise réponse ou réponse non-prévu) et des pistes pour améliorer l'interaction. Néanmoins l'extraction de modèles via la fouille de processus sur un jeu de données aussi large sont difficile à exploiter. Nous proposons ici quelques perspectives pour les rendre plus exploitables.

Mots clés : Fouille de processus, Fouille de séquences, Interactions Humain-Robot

1 Introduction

A l'heure du "tout numérique", la génération de données et l'exploitation de ces dernières est un enjeu de taille pour la recherche ou l'industrie. L'interaction Humain-Robot est un domaine de recherche où l'interactivité y est présente sous nombre de formes, ce qui suscite fortement l'intérêt chez le public. Chez les personnes jeunes, particulièrement friandes de ce genre de choses qui y voient une manière d'accéder à du contenu numérisé et souvent de manière ludique, de l'engouement est généré. Chez les personnes plus âgées, cela représente une curiosité à découvrir. Cet intérêt, de plus en plus croissant pour des expériences reposant sur ce type d'interaction amène des secteurs à s'adapter et proposer bien plus ce genre d'expérience afin d'attirer le public au sein de ces secteurs, comme par exemple les lieux de culture tels que les musées, peu fréquentés par une population jeune. Ce type d'expérience interactive mêlant ludique et culturel s'inscrit dans le domaine des jeux sérieux (serious game), qui consiste à mêler le côté culturel (d'un endroit, d'une exposition etc.) au côté ludique d'une interaction avec différents terminaux numériques tels que des tablettes, des smartphones ou bien des robots, et ce afin de proposer des expériences plus enrichissantes et appréciables qu'une visite simple.

Ces différentes interactions vont générer des traces numériques, aussi appelés logs, principalement sous un format textuel, comme présenté sous la forme du tableau 1. Après leur collecte sous forme de journal d'activités, les logs peuvent être exploités grâce à diverses techniques issues de la fouille de processus.

- **CaseID :** il correspond à l'identifiant de l'utilisateur qui effectue l'action ;
- **Activité :** l'événement que l'utilisateur a déclenché ;
- **Timestamp :** la date et l'heure où l'utilisateur a déclenché son événement ;

caseID	Activité	Timestamp
joueur 1	repondsQuestion1	13/10/2017 14:25:01
joueur 2	demandeIndice3	13/10/2017 14:26:54
joueur 1	repeteQuestion2	13/10/2017 14:28:30

Tableau 1. Exemple de données exploitables par un algorithme de fouille de processus

La fouille de processus permet d'extraire des graphes ou des réseaux de Petri afin de visualiser l'enchaînement des différents processus exécutés lors d'une interaction. Un analyste peut ainsi tirer des conclusions de ce qu'il perçoit, afin de faire évoluer cette interaction.

L'idée derrière la fouille de processus est de découvrir les processus (Process Discovery), vérifier la conformité ou la qualité des modèles découverts (Conformance Checking) et l'amélioration de processus réels (Enhancement) par l'extraction de connaissances à partir des données de parcours.

Malgré sa jeunesse dans le paysage informatique (le début remonte aux années 1990), la fouille de processus a obtenu rapidement un intérêt croissant dans divers domaines d'application du fait de ses finalités. On peut observer son utilisation dans le cadre d'analyse de processus dans un hôpital [5], dans l'industrie [10] ou bien encore dans le jeu-vidéo [2]. Des travaux ont également été effectués dans le cadre des jeux sérieux comme on peut le voir dans les travaux de Hernández et al. [3].

Le fonctionnement général d'un algorithme de fouille de processus est divisé en trois étapes : (1) l'analyse des données de navigation dans les fichiers de logs ; (2) l'identification des processus et de leurs relations; (3) construction d'un modèle. Les modèles obtenus permettent de décrire le processus et les intéractions sous forme d'un diagramme.

Les fichiers de logs constituent une excellente source de données et peuvent être considérés comme des traces brutes qu'il faudra par la suite transformer en traces modélisées (sous forme d'un journal d'événements) afin de pouvoir les utiliser. Ces traces modélisées doivent porter les informations nécessaires représentant les activités afin de pouvoir en effectuer leur analyse.

Pour résumer, il faut sessioniser les actions utilisateurs afin de pouvoir en faire une analyse. Des champs additionnels, porteurs d'informations, peuvent être ajoutés aux trois précédemment cités (paramètres optionnels, courte description etc.). Mais que se passe t-il lorsque nous souhaitons générer un modèle à partir d'une masse de données volumineuse?

Nous allons, dans un premier temps, effectuer un état de l'art de la fouille de processus qui permettra de regrouper les différents modèles de processus existants, ainsi que les algorithmes présents dans la littérature. Nous aborderons ensuite notre cas d'étude et les perspectives qu'elle amène, puis nous concluerons.

2 État de l'art

2.1 Les différents modèles de processus

Les algorithmes de fouille de processus permettent d'obtenir différents types de modèles que nous allons présenter, chaque algorithme permettant d'obtenir en général un ou plusieurs des modèles présentés ici (il existe des équivalence entre certains modèles).

Tout d'abord nous avons les ***réseaux de Petri***, présenté pour la première fois par Murata [6], qui sont des modèles mathématiques avec représentation graphique, qui permettent d'exprimer des séquences d'évènements, avec des synchronisations et des partages de ressources. Ils sont composés de deux types de noeuds : des places et des transitions, reliés par des arcs. Si un arc relie une place à une transition il est dit "entrant", s'il relie une transition à une place, il est dit "sortant". Pour décrire l'état courant de ce type de système, on utilise un système de jeton qui sont associés aux places. L'évolution dans ce système est régie par un ensemble de règles qui décrivent les transitions (récupérer un objet dans un jeu par exemple).

Nous avons ensuite les ***Workflow Nets***, introduits par Van der Aalst [8] en 1997 et repris ensuite en 2011 [9], qui sont une sous-classe des réseaux de Petri. Ses caractéristiques les plus importantes sont d'avoir une place dédiée pour un "début" du processus et une "fin" du processus. Un tel modèle implique que pour chaque succession d'événements depuis la place initiale, il existe une séquence de transitions qui permet d'atteindre la place finale.

Il existe également les ***Directly-Follows Graphs*** (DFG) qui sont des graphes où chaque noeud représente une activité existante dans le journal d'événements et les arêtes dirigées sont présentes entre les noeuds s'il existe au moins une trace dans le journal d'événement où l'activité "source" est suivie par l'activité "cible". Il est également facile de représenter des mesures comme la fréquence

(nombre de fois où l'activité source est suivie par l'activité cible) et des indicateurs de performance (moyenne de temps écoulé par exemple entre les deux activités).

2.2 Les différents algorithmes de fouille de processus

Au travers de la littérature, nous pouvons observer l'existence de plusieurs algorithmes de fouille de processus, chacun présentant ses forces et ses faiblesses. En effet, l'évolution croissante du domaine a permis d'affiner les différentes techniques et permettent de travailler sur des traces de qualités différentes (traces bruitées, incomplètes, non structurées) afin d'en obtenir tout de même une représentation du modèle de processus. Généralement, les différentes techniques de découverte de processus se basent sur l'identification des liens entre les activités dans un ensemble de traces. On peut identifier quatre relations qui permettent de définir deux activités quelconques a_1 et a_2 :

- la succession directe, indique que a_1 est immédiatement suivi de a_2 dans certaines séquences d'événements, elle est notée $a_1 \geq a_2$;
- la causalité indique que pour toutes les traces, a_1 est situé toujours avant a_2, il n'y a jamais a_2 avant a_1. Il faut la différencier de la succession directe dans le sens où a_1 et a_2 peuvent être espacés d'une où plusieurs activités. Elle est notée $a_1 \rightarrow a_2$;
- parallèle si dans une trace on obtient $a_1 \geq a_2$ et dans une autre $a_2 \geq a_1$. Elle est notée $a_1 \| a_2$;
- le choix lorsqu'il n'y a jamais $a_1 \geq a_2$ ni $a_2 \geq a_1$. Il est noté $a_1 \neq a_2$.

L'un des premiers algorithmes à avoir vu le jour pour la découverte de processus est l'algorithme α par Van der Aalst et al. [7]. Il consiste à identifier les différentes relations observées entre les activités dans les logs et se fonde sur les 4 relations énoncées précédemment. Il est en revanche incapable de découvrir des boucles ou des activités dupliquées. Certaines améliorations à cet algorithme ont été apportées avec l'apparition de α^+, puis α^{++}. Cette famille d'algorithmes souffre toujours cependant du bruitage dans le journal d'événements.

Nous avons ensuite l'Heuristic Miner, développé par Weijters et al. en 2006 [12] et amélioré en 2011 [11]. Comme son nom l'indique, il se base sur une approche heuristique afin de résoudre les différents problèmes rencontrés par l'algorithme α. La différence fondamentale entre α et Heuristic Miner vient du fait qu'Heuristic Miner se base sur des mesures statistiques afin de déterminer les relations entre activités. Les inconvénients de cet algorithme sont que les modèles produits contiennent des anomalies et ne sont pas capables de rejouer la majorité des logs.

Un autre algorithme est l'Inductive Miner, découvert en 2014 par Leemans et al. [4]. Il a été développé afin de produire des modèles de processus qui permettent de rejouer la majorité des logs et qui ne contiennent pas d'anomalies. Il permet en outre de traiter efficacement un potentiel bruit dans les logs en supprimant les traces non fréquentes et permet de traiter le problème des logs non complets et est a l'heure actuelle l'algorithme le plus utilisé en fouille de processus. En revanche il n'est pas en mesure d'identifier des motifs complexes et non locaux de contrôle de processus.
Un comparatif des algorithmes est disponible dans le tableau 2.

	Bruité	**Incomplet**	**Log réel**
α^+	non supporté	non supporté	non supporté
Heuristic Miner	supporté	non supporté	non supporté
Inductive Miner	supporté	supporté	supporté

Tableau 2. Tableau d'analyse des supports de logs pour les algorithmes cités

3 Cas d'étude

Notre cas d'étude est un jeu sérieux (serious game) ayant eu lieu au Muséum d'Histoire Naturelle de La Rochelle en 2017. Ce jeu sérieux permettait de faire découvrir une exposition au public qui

devait alors répondre à de multiples questions en rapport avec l'exposition, posées par un robot humanoïde Nao (l'automate des possibles est présent en annexe, figure 1).

Les joueurs étaient tous identifiés à l'aide de symboles reconnaissables par le robot, afin de bien associer un événement effectué au bon joueur (répondre à une question par exemple). Cette activité a entraîné la création de nombreux logs (environ 6500 traces) pour chaque action qui a été effectuée par les participants. Plusieurs questions sont soulevées à la suite de cette expérimentation :

- L'exploitation des logs peut-elle conduire à l'amélioration future de cette activité ?
- Existe t-il des catégories de joueurs qui se dégagent dans les logs ?

Ces deux questions correspondent bien à l'esprit de la fouille de processus dans son objectif. La première question est intéressante pour observer les différents problèmes qu'il y a pu avoir lors de l'activité et éviter certaines erreurs ou abandons des participants.

En effet, lors du déroulement de l'activité, des traces indiquent la présence de multiples erreurs de reconnaissance vocale ou des problèmes d'identification du joueur, ce qui a engendré, à terme, des abandons sans doute dûs à une frustration de la part du joueur. Cela a pu donner une piste de résolution pour un emploi futur de cette activité mais aussi d'autres activités employant le même mode opératoire avec le robot.

La difficulté de la deuxième question réside dans le fait que chaque activité possède des paramètres bien distincts associés aux paramètres du robot et de l'activité (mouvements qui doivent être effectués par le robot, rapidité d'élocution du texte, identification de la question et texte à prononcer, réponse du joueur, etc.). Ces paramètres pourraient être bien évidemment supprimés de la trace mais une grande partie de la sémantique serait alors perdue. Du fait de la richesse combinée de toutes ces traces, l'exploitation du graphe de tous les joueurs n'est pas possible et donne lieu à un graphe incompréhensible doté de toute sorte de noeuds et d'arêtes entrelacés.

Afin de pouvoir extraire des informations des graphes des joueurs, il faut procéder à l'extraction du graphe du joueur demandé et visualiser manuellement ce graphe. Bien que les comportements communs se reflètent assez bien entre deux joueurs lorsque l'on se penche individuellement sur ces graphiques (voir figure 2 en annexe, on remarque qu'ils ont tous les deux donnés les mêmes réponses, donnant ainsi un comportement commun. Les graphiques ont été extraits à l'aide de la bibliothèque Python *pm4py*), effectuer cette tâche pour un jeu de données important est fastidieux. Il nous faut donc proposer des améliorations afin de pouvoir exploiter ces logs sans retirer d'informations, en conservant la sémantique des données.

4 Piste exploratoire

Une piste à explorer est l'utilisation de la fouille de séquences. En effet, un journal d'événements peut être assimilé à une séquence d'événements survenue avec un ordre précis. Mais le fait est que, comme nous l'avons vu dans la section 2, ce journal d'événements est composé d'un ensemble de données hétérogènes qui influent grandement sur la sémantique des données et que d'autres champs peuvent être ajoutés aux trois principaux. Une analyse avec un outil spécialisé peut être intéressant.

Dans le cadre de ses travaux de recherche, le Laboratoire Informatique, Image et Intéraction (L3i), de La Rochelle Université, propose un outil, centré sur l'analyste de données, capable de calculer un treillis de concept à partir de données complexes (type séquences) et hétérogènes, nommé **GALACTIC** (**GA**lois **LA**ttices, **C**oncept"'e **T**heory, **I**mplicational systems and **C**losures). Cet outil utilise l'algorithme NextPriorityConcept [1] et permet de calculer des concepts pour des données hétérogènes et complexes en entrée.

On obtient alors un treillis de concept qui représente les données sous forme de clusters hiérarchiques et ce de manière non supervisée. Chaque concept est décrit par un ensemble de prédicats, chaque prédicat étant spécifique à chaque caractéristique. Par exemple, une caractéristique numérique peut être décrite sous la forme d'un prédicat de la forme "*est plus petit/grand que n*", n étant une valeur numérique. Une séquence peut, elle, être décrite sous forme de prédicat de la manière suivante : "*correspond à la sous-séquence s*".

L'utilisation de cet outil permettrait donc d'effectuer une autre approche et d'obtenir un modèle de données plus compréhensible pour la même masse de données.

5 Conclusion

La fouille de processus est un enjeu très intéressant pour l'amélioration des jeux sérieux et l'extraction de connaissances associées à une activité. Mes expérimentations à court terme vont consister à me servir de l'outil Galactic afin d'extraire la connaissance des logs et comparer les approches évoquées dans ce papier.

Remerciements

Nous remercions le projet ANR SmartFCA Grant ANR-21-CE23-0023 de l'Agence Nationale Française de Recherche qui permet le financement de ce papier.

References

1. Christophe Demko, Karell Bertet, Cyril Faucher, Jean-François Viaud, and Sergei O. Kuznetsov. NextPriorityConcept: A new and generic algorithm computing concepts from complex and heterogeneous data. *Theoretical Computer Science*, 845:1–20, 2020.
2. Adam Godziński, Andrzei Stroinski, Wojciech Piątek, and Aleksander Stroiński. Pattern recognition in games using process mining. In *2022 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE)*, pages 42–45, 2022.
3. Juan Antonio Caballero Hernández, Manuel Palomo Duarte, and Juan Manuel Dodero. An architecture for skill assessment in serious games based on event sequence analysis. In *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM 2017, New York, NY, USA, 2017. Association for Computing Machinery.
4. Sander JJ Leemans, Dirk Fahland, and Wil MP Van Der Aalst. Discovering block-structured process models from event logs containing infrequent behaviour. In *Business Process Management Workshops: BPM 2013 International Workshops, Beijing, China, August 26, 2013, Revised Papers 11*, pages 66–78. Springer, 2014.
5. R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. van der Aalst, and P. J. M. Bakker. Application of process mining in healthcare – a case study in a dutch hospital. In Ana Fred, Joaquim Filipe, and Hugo Gamboa, editors, *Biomedical Engineering Systems and Technologies*, pages 425–438, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
6. T. Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580, 1989.
7. W. van der Aalst, T. Weijters, and L. Maruster. Workflow mining: discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142, 2004.
8. Wil M. P. Van der Aalst. Verification of workflow nets. In Pierre Azéma and Gianfranco Balbo, editors, *Application and Theory of Petri Nets 1997*, pages 407–426, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.
9. Wil MP Van Der Aalst, Kees M Van Hee, Arthur HM Ter Hofstede, Natalia Sidorova, HMW Verbeek, Marc Voorhoeve, and Moe Thandar Wynn. Soundness of workflow nets: classification, decidability, and analysis. *Formal aspects of computing*, 23:333–363, 2011.
10. W.M.P. van der Aalst, H.A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, and H.M.W. Verbeek. Business process mining: An industrial application. *Information Systems*, 32(5):713–732, 2007.
11. AJMM Weijters and Joel Tiago S Ribeiro. Flexible heuristics miner (fhm). In *2011 IEEE symposium on computational intelligence and data mining (CIDM)*, pages 310–317. IEEE, 2011.
12. AJMM Weijters, Wil MP van Der Aalst, and AK Alves De Medeiros. Process mining with the heuristics miner-algorithm. *Technische Universiteit Eindhoven, Tech. Rep. WP*, 166(July 2017):1–34, 2006.

Annexes

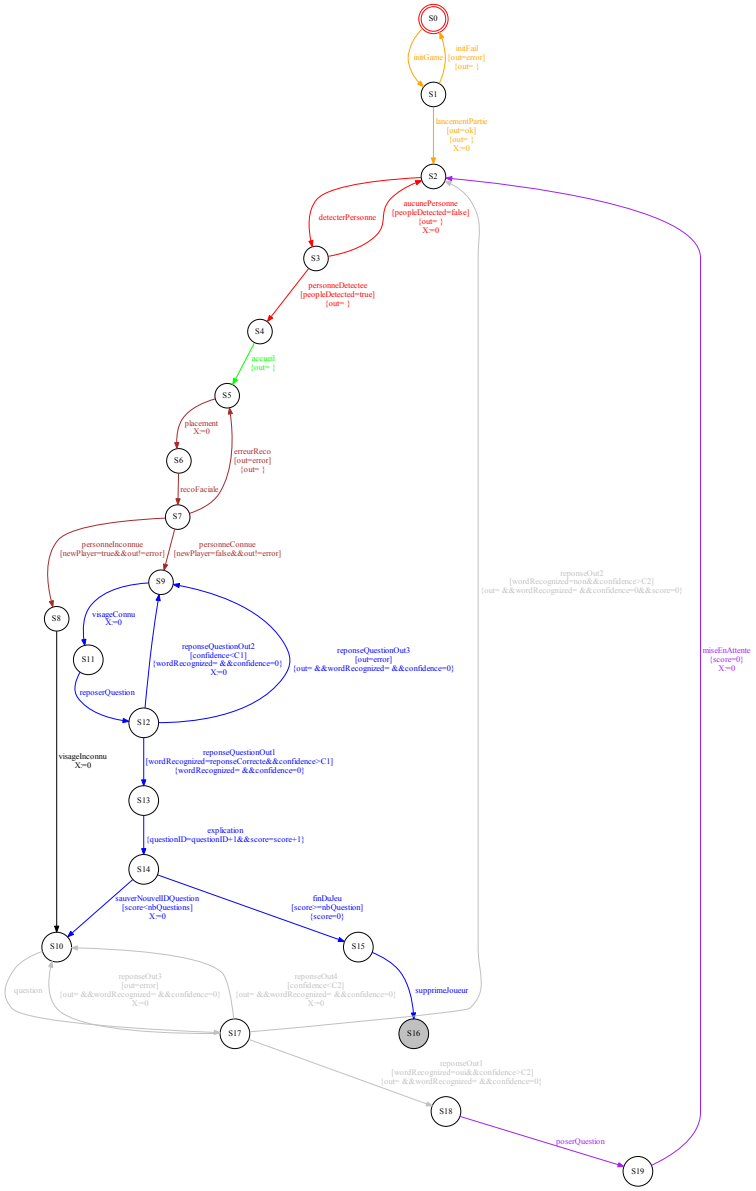

Fig. 1. Graphe des possibles de l'expérimentation du Muséum d'Histoire Naturelle

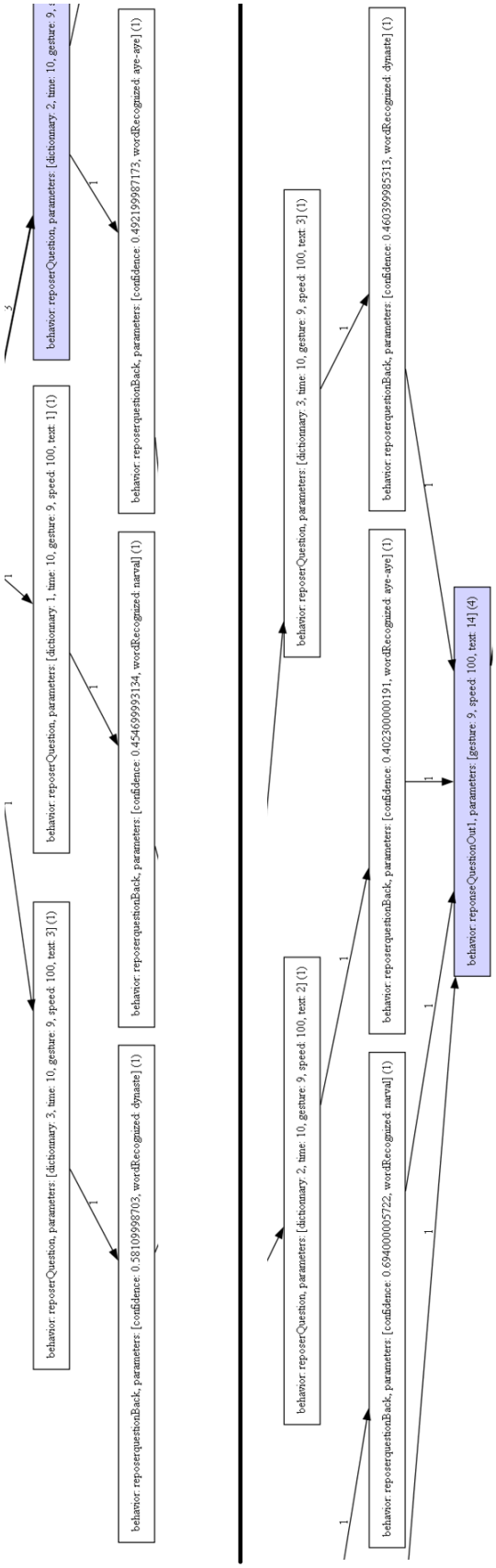

Fig. 2. Comparaison entre deux joueurs

Transcription de séries temporelles en séquences temporelles via conservation des caractéristiques de variation

Guillaume Savarit, Karell Bertet, et Christophe Demko

Laboratoire Informatique, Image et Interaction, La Rochelle Université, La Rochelle, France
guillaume.savarit@univ-lr.fr, karell.bertet@univ-lr.fr, christophe.demko@univ-lr.fr

Résumé : Dans cet article, nous nous intéressons à la possibilité de transcrire les séries temporelles en séquences temporelles en conservant des caractéristiques issue de la série, de manière à réduire la volumétrie des données et profiter de l'analyse multiséquentielle et hétérogène apportée par des outils déjà existant de l'analyse de séquence. Nous utilisons en exemple d'application les données issues des capteurs de niveau de la marée du Port Atlantique de La Rochelle.

Mots clés : Série temporelle · Séquence temporelle · Analyse de données

1 Introduction

Toutes données collectées avec un capteur au cours du temps vont former une série temporelle, sous la forme d'une association entre une valeur x_i et une unité de temps t_i.

$$y =< (x_i, t_i) > \tag{1}$$

Bien que de fait extrêmement commune, l'analyse des séries temporelles brutes soulève de nombreuses problématiques lors de l'analyse de celle-ci. Leurs tailles dépendant de la fréquence et de la durée de captation, il s'agit de données pouvant très vite atteindre un volume conséquent.

Pour réduire ce volume, il est possible de résumer ces séries en identifiant leurs caractéristiques. Selon l'objet de l'analyse, il est intéressant de détecter les outliers, les valeurs aberrantes que prendraient la série ponctuellement, ou de détecter les périodes de celle-ci. Pour cela, des algorithmes de changements de points ou de détection d'erreurs sont utilisés. Mais dans ce cas, le temps devient une indication des instants où les événements analysés ont eu lieu.

Un autre objet d'analyse des séries consiste à faire de la prédiction de nouvelles valeurs en fonction de celles captées. Dans ce cas, les méthodes de régression, linéaire ou statistique, sont utilisées. Il est ainsi possible d'approcher la série « idéale », un modèle représentant nos valeurs. Mais dans ce cas, les informations portées par les valeurs ne sont pas expliquées.

Pour tenter malgré tout d'expliquer les séries tout en conservant la notion temporelle, et en réduisant le volume des données, nous allons proposer ici une approche des séries temporelles sous forme de séquences conservant les caractéristiques de la série et pouvant permettre une analyse multiséquentielle hétérogène via la plateforme Galactic [3].

Nous allons tout d'abord définir les séries temporelles et les séquences, puis expliquer la chaine de traitement généralisé permettant de construire une séquence à partir d'une série, et enfin mettre en perspective les futurs traitements.

2 État de l'art

2.1 Séries temporelles

La notion de série temporelle est assez ancienne, réellement apparue dans la littérature dans les années 20. L'une des premières définitions est celle de Warren M. Persons en statistique [9] qui définit une série temporelle comme « un nombre agrégé ou moyenné ou relatif appliqué à un interval défini ou un temps défini », et donne trois conditions : même unité de temps (par exemple le jour, l'année, le mois), consécutifs dans le temps, construit par un critère fixe ou un standard. Cette définition a pour objectif de rapprocher une série temporelle d'une fonction en plus restreinte.

C'est en économie que Harold T. Davis [5] va poser une définition plus générale, qui est plus proche de celle actuellement utilisée : « une série de données observées successivement dans le temps ». Il exprime mathématiquement celles-ci comme une séquence tel que :

$$y = y_t \text{ avec } t = 1, 2, 3, \ldots, n \tag{2}$$

Contrairement à la définition précédente qui considérait les séries temporelles comme des fonctions, la définition de Davis implique la discontinuité des séries temporelles. Néanmoins il considère que pour des intervalle de temps suffisamment court nous pouvons considérer que la variable t est continue et écrire :

$$y = f(t) \tag{3}$$

Davis décrit aussi la différence entre les séries temporelles utilisées en astronomie et en économie. Il introduit ainsi différents types de séries temporelles, celles reposant sur des variables peu nombreuses et connues qu'il est possible de mettre en équation (astronomie) et celles reposant sur de multiples variables et des tendances (économie).

Dans les années 80 va émerger une précision sur la définition des séries temporelles. Considérant que toutes les séries reposent sur des fonctions π, il est possible d'écrire :

$$y_t = \sum_{k=1}^{\infty} \pi_k y_{t-k} + a_t \tag{4}$$

Où k est le lag et a le bruit.

Il existe plusieurs classification des séries temporelles. Comme de nombreux autres types de données on distingue les séries univariées (reposant sur une seule variable) et multivariées (reposant sur plusieurs variables). Il existe aussi une classification temporelle, entre les séries à haute fréquence (petit interval de temps) et basse fréquence (grand interval de temps). Cette distinction n'entre en compte que dans la comparaison de plusieurs séries temporelles.

L'analyse de séries temporelles se fait majoritairement en utilisant des modèles. Ainsi, dans leur article de 1972, qui sera ensuite utilisé pour poser les bases du modèle ARIMA (*AutoRegressive Integrated Moving Average*, un modèle statistique utilisant les valeurs précédentes pour prédire les valeurs suivantes), Amemiya et Wu [2] expliquent qu'un modèle d'agrégation fonctionnant à haute fréquence fonctionnera aussi à basse fréquence. Brewer en 1973 [4] va aussi soulever le problème de la fréquence, et le résoudre en généralisant le modèle.

En 1987, Lüktepohl [7] montre que les modèles ARIMA entres autres ne fonctionnent que sur des données univariées. Lorsqu'appliqué sur des données multivariées, la question de la fréquence redevient importante, notamment dû à l'hétéroscédasticité. En effet un modèle régressif ne fonctionne que si la variance de l'erreur des variables est faible (homoscédasticité). Dans le cas inverse, les erreurs de prédictions du modèle seront grandes.

2.2 Séquences temporelles

Les séquences sont un type de données très utilisé pour representer notamment les déplacements, les phrases et autres. Une séquence temporelle A est définie par un ensemble ordonné d'événements. Chaque événement associe un symbole s_i à un interval non nul $[\underline{t_i}, \overline{t_i}]$ où le symbole peut être une valeur symbolique issue d'un dictionnaire, ou une valeur numérique.

$$A =< (s_i, [\underline{t_i}, \overline{t_i}]) > \tag{5}$$

La fouille de motifs sequentiels repose originellement sur des algorithmes comme GSP d'Agrewal et Srikant [1], qui travaille sur les séquences sans notion d'intervalle. Guyet et Quiniou en 2008 [6] avec *QTempIntMiner* et Nakagaito en 2009 [8] avec *QTPSpan* s'intéressent à la fouille de séquences d'intervalles.

En 2020, Boukhetta [3] utilise l'analyse formelle de concept via le framework Galactic pour analyser les séquences temporelles. Cette approche permet aussi l'hétérogénéité des données analysées ainsi qu'une analyse multiséquence.

3 Chaîne de traitement

3.1 Général

Notre méthode de traitement consiste à transformer les séries temporelles en séquences temporelles en utilisant la composante sémantique pour décomposer les séries en différents épisodes de valeurs consecutives de même sémantique. Les séquences ainsi obtenues résument les séries temporelles tout en réduisant leur volumétrie, et il est possible d'utiliser des outils existants d'analyse de séquences.

L'analyste choisit une sémantique applicable à la série temporelle en construisant un dictionnaire représentant cette sémantique. Dans la suite nous nous intéresserons à un cas général, celui de l'analyse par la variation de la courbe, en étudiant le signe de la dérivée première de celle-ci. Le dictionnaire sera ainsi composé de 3 éléments au moins :

$$\Sigma = \{Croissant, Decroissant, Stable\} \tag{6}$$

Toujours dans le cas général, la stabilité peut être scindée en deux types. Soit le pic d'une courbe, entre deux intervalles croissant puis décroissant, soit la vallée, entre deux intervalles décroissant puis croissant.

Un algorithme dédié de changements de points, qui découvre les points d'inflexions de la série, sépare ensuite les valeurs de la série en différents épisodes consécutifs afin d'obtenir une séquence temporelle.

Enfin, les séquences sont annotées grâce au dictionnaire choisi. Nous obtenons ainsi à partir d'une série temporelle donnée la transformation suivante :

$$< (x_i, t_i) > \rightarrow < (s_j, [\underline{t_j}, \overline{t_j}]) > \text{avec} \quad s_j \in \{C, D, S\} \tag{7}$$

La chaîne de traitement est donc celle-ci ainsi :

Choix de la sémantique L'analyste choisit un dictionnaire qui représente la sémantique de ses données.
Découpage en séquence La série est transformée en séquence respectant le dictionnaire.
Annotation Le dictionnaire est appliqué à la séquence.

3.2 Application

Dans la suite nous allons montrer l'application sur un cas concret. Il s'agit d'une série temporelle issue des données de capteurs de niveaux de la marée dans le bassin à flot du Port Atlantique de La Rochelle. Ces capteurs sont placés autour du bassin à flot et retransmettent les niveaux d'eau de l'océan, du sas et du bassin à flot au cours du temps. Chaque mesure est prise toutes les 60 secondes en moyenne, et réprésente les données de mars 2021 à mars 2022. Cela représente quelques 981 070 mesures.

Nous nous intéressons en particulier aux mesures du coté de l'océan et à la question de prédiction des marées, nous allons utiliser une sémantique représentant ce que nous connaissons de la marée avec un dictionnaire à 4 éléments :

$$\Sigma = \{Montante, Descendante, Haute, Basse\} \tag{8}$$

Par rapport à notre dictionnaire général, les notions de marée *Montante* et *Descendante* sont équivalentes à la croissance et la décroissance. Les marées *Haute* et *Basse* quant à elles sont une séparation de la stabilité selon les symboles précédent et suivant de la séquence. Une marée stable après une marée montante et suivie par une marée descendante sera une marée haute.

Une fois la série traduite en séquence, nous obtenons :

$$< (x_i, t_i) > \rightarrow < (s_j, [\underline{t_j}, \overline{t_j}]) > \text{avec} \quad s_j \in \{M, D, H, B\} \tag{9}$$

Une fois le traitement terminé, nous obtenons 5000 séquences pour l'année. Ceci réduit drastiquement le volume de données pour l'analyse tout en résumant l'information portée par la série temporelle.

4 Perspective

Pour contrebalancer la perte d'information induite par la transformation en séquence, il est possible par la suite d'extraire d'autres caractéristiques internes aux séries, tel que les extremums, les moyennes, voire les valeurs elle-même. En tirant partie de la possibilité d'analyser des séquences avec des données hétérogènes offerte par GALACTIC, il est ainsi possible d'analyser les séquences avec plus d'information.

Il serait ainsi possible d'analyser la séquence suivante, prenant en compte les extremums :

$$< (x_i, t_i) > \rightarrow < (s_j, [x_{min}, x_{max}], [\underline{t_j}, \overline{t_j}]) > \text{avec} \quad s_j \in \{C, D, S\} \tag{10}$$

Ceci afin de conserver des informations des séries temporelles tout en réduisant le volume des données en tirant partie de l'analyse multiséquentielle.

En utilisant une telle approche, nous allons dans le futur mettre en place la possibilité d'expliquer les séries temporelles via les outils de l'analyse formel de concept, notamment GALACTIC, en conservant les caractéristiques de la série sous forme de séquence, réduisant le volume de donnée et permettant le traitement de celle-ci.

Remerciements

Nous remercions le projet ANR SmartFCA Grant ANR-21-CE23-0023 de l'Agence Nationale Française de Recherche qui permet le financement de ce papier.

References

1. R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, March 1995.
2. T. Amemiya and R. Y. Wu. The Effect of Aggregation on Prediction in the Autoregressive Model. *Journal of the American Statistical Association*, 67(339):628–632, September 1972. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1972.10481264.
3. S. E. Boukhetta, C. Demko, K. Bertet, J. Richard, and C. Cayèré. Temporal Sequence Mining Using FCA and GALACTIC. In Tanya Braun, Marcel Gehrke, Tom Hanika, and Nathalie Hernandez, editors, *Graph-Based Representation and Reasoning*, Lecture Notes in Computer Science, pages 185–199, Cham, 2021. Springer International Publishing.
4. K. R. W. Brewer. Some consequences of temporal aggregation and systematic sampling for ARMA and ARMAX models. *Journal of Econometrics*, 1(2):133–154, June 1973.
5. H. T. Davis. The analysis of economic time series. *Journal of the Royal Statistical Society*, 105(2):125–127, 03 1942.
6. T. Guyet and R. Quiniou. Mining Temporal Patterns with Quantitative Intervals. In *2008 IEEE International Conference on Data Mining Workshops*, pages 218–227, December 2008. ISSN: 2375-9259.
7. H. Lütkepohl. Forecasting Aggregated Vector ARMA Processes, 1987.
8. F. Nakagaito, T. Ozaki, and T. Ohkawa. Discovery of Quantitative Sequential Patterns from Event Sequences. In *2009 IEEE International Conference on Data Mining Workshops*, pages 31–36, December 2009. ISSN: 2375-9259.
9. W. M. Persons. Correlation of Time Series. *Journal of the American Statistical Association*, 18(142):713–726, June 1923. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1923.10502103.