Rapport d'activité 2021

Équipe PASTIS

1 Composition et domaines de recherche du groupe

Le groupe de recherche PASTIS est composé de 10 titulaires, 4 doctorants et 4 ATER :

- Adrien Revault d'Allonnes (MCF)
- Anna Pappa (MCF)
- Farès Belhadj (MCF)
- Françoise Balmas (MCF-HDR)
- Jean-Jacques Bourdin (PR)
- Nicolas Jouandeau (PR - responsable de l'équipe)
- Pablo Rauzy (MCF)
- Revekka Kyriakoglou (MCF - installée en septembre 2021)
- Sylvia Chalençon (MCF)
- Vincent Boyer (MCF-HDR - en disponibilité)
- Jean-Pascal Palus (Doctorant 2021 - contrat doctoral ED CLI)
- Maroua Boudabous (Doctorante 2019 - contrat CIFRE avec la société NOVAGEN)
- Patrick Gikunda (Doctorant 2018 - financement Ministère des Affaires Etrangères et de l'Europe (MEAE) / Dedan Kimathi University, Nyeri, Kenya)
- Stephen Obonyo (Doctorant 2021 - financement Ministère des Affaires Etrangères et de l'Europe (MEAE) / Strathmore University, Nairobi, Kenya)
- Emna Chebbi (ATER 2021-2022)
- Hanane Zerdoum (ATER 2021-2022)
- Oumaima El Joubari (ATER 2021-2022)
- Syrine Saidi (ATER 2021-2022)
- Rahima Zaouche (ATER 2020–2021)
- Rémi Nollet (ATER 2020–2021)
- Sarah Zouinina (ATER 2020–2021)

Le groupe a réalisé des travaux en synthèse d'images expressives, en enseignement de l'informatique graphique, en résolution de jeux et apprentissage automatique, en représentation du raisonnement et logique non classique, en langage naturel et apprentissage automatique, en dynamique symbolique, et en sécurité et privacy.

Jean-Pascal Palus est sous la direction de Adrien Revault d'Allonnes et Nicolas Jouandeau. Son sujet de thèse est « Dynamique et évolution de la confiance : explication et formalisation de la mésinformation ».

Maroua Boudabous est sous la direction de Anna Pappa et Françoise Balmas. Son sujet de thèse est « Modélisation d'un système intelligent d'analyse prédictive des données textuelles massives pour l'aide à la décision ».

Patrick Gikunda était sous la direction de Nicolas Jouandeau et a soutenu sa thèse le 9 novembre 2021. Le titre de sa thèse est « Precipitation Forecasting with Deep Transfer Active

Learning for Agricultural Adaptation ». Il est aujourd'hui Lecturer à School of computer science and IT, University of Nairobi.

Stephen Obonyo est sous la direction de Nicolas Jouandeau. Son sujet de thèse est « Bioinformatic Computation With Active Learning ».

2 Bilan en chiffres

Pour l'année 2021, le groupe PASTIS compte :

- publications en conférence et revues : 4
- soutenances de thèse et HDR : 1
- organisations/participations à des évènements scientifiques : 4
- projets en cours : 4
- nouveaux projets déposés : 1
- implications dans les sociétés savantes, GdR et centres de recherche : 4

Le groupe est impliqué dans le Conseil d'administration de l'AFIG [1], dirige un chapitre de IEEE CIS France [2], est impliqué dans le centre de recherche Geopolitique de la datasphere (GEODE) [3], est impliqué dans le Conseil enseignement de EUROGRAPHICS [4].

3 Soutenances de thèse

Patrick Kinyua Gikunda

Sujet : *Precipitation Forecasting with Deep Transfer Active Learning for Agricultural Adaptation.*

Jury : Thomas Guyet (IRISA, Rapporteur), Tristan Cazenave (LAMSADE, Rapporteur), Frédéric Précioso (I3S, Examinateur), Jean-Noël Vittaut (LIP6, Examinateur), Anna Pappa (LIASD, Examinatrice), Nicolas Jouandeau (LIASD, Directeur de thèse).

Date : 9 novembre 2021.

4 Organisation et participation à des évènements scientifiques

- Jean-Jacques Bourdin : animation de la table ronde sur le thème « premiers enseignements de l'informatique graphique », 34èmes journées de l'AFIG 2021, conférence jFIG, INRIA de Sophia-Antipolis, 24/11/21.
- Farès Belhadj : présentation de l'approche pédagogique adoptée pour enseigner l'informatique graphique dans le cadre de la mineur jeux vidéo de la Licence Informatique de l'Université Paris 8, 34èmes journées de l'AFIG 2021, conférence jFIG, INRIA de Sophia-Antipolis, 24/11/21.
- Pappa Anna, Berenguer Rocio, Laborderie Arnaud : Litte_Bot, un agent conversationnel : une dramaturgie pour dialoguer avec Dom Juan. Journées de pré-conférence : Retours d'expériences pratiques et techniques, Les Futurs Fantastiques 2021 #FF21, présentation/démo (Université Paris Saclay le 8/12/21, cf. Section 5.2).

1. https://www.asso-afig.fr
2. http://ieee-ci.lip6.fr/
3. https://geode.science/
4. https://www.eg.org/

— Pappa Anna, Berenguer Rocio, Laborderie Arnaud : Litte_Bot, un agent conversationnel : une dramaturgie pour dialoguer avec Dom Juan. Journée de la Conférence : Les défis de l'IA dans les bibliothèques, archives et musée (BnF le 10/12/21, cf. Section 5.2).

5 Projets en cours

L'équipe participe a quatre projets, le projet PAMOJA financé par le MEAE, le projet Litte_Bot financé par l'EUR-ArTeC, le projet MALANTIN financé par l'appel à projet européen H2020 RISIS et le projet ReComp financé par l'ANR.

5.1 PHC PAMOJA « Deep Learning To Identify Fall Armyworm Pest »

Porteur : Nicolas Jouandeau
Partenaires : Université Paris 8 (France), University of Nairobi (Kenya) , KALRO (Kenya)
Date de début : 05/2020
Date de fin : 12/2021
Description : Ce projet a permis de proposer une nouvelle solution d'identification d'un insecte ravageur des cultures de maïs au Kenya. Il a été réalisé une étude des solutions de prise de décicion par apprentissage, une étude du comportement et des caractéristiques de l'insecte, la construction d'un dataset de référence pour ce problème, la définition d'un diagnostic d'infestation lié au comportement de l'insecte et une solution permetttant de produire un diagnostic à moindre coût computationnel.

5.2 AAP EUR-ArTeC « Litte_Bot »

Porteuse pour la partie informatique : Anna Pappa
Partenaires : Université Paris 8 (LIASD), Bibliothèque nationale de France (BnF), Rocio Berenguer (auteure, metteuse en scène, chorégraphe)
Date de début : 01/2019
Date de fin : 12/2021
URL : `https://eur-artec.fr/projets/litte_bot/`
Description : Le projet Litte_Bot est un dispositif numérique collaboratif autour des agents conversationnels, chatbot, pour dialoguer avec des personnages du théâtre du 17e siècle. Nous donnons l'accent au personnage de Don Juan, puisque le projet s'inscrit dans la perspective du quatre centième anniversaire de la naissance de Molière en 2022. Je travaille sur l'apprentissage automatique pour créer les dialogues entre l'humain et le bot. Le chatbot donne la réplique en respectant la langue et le style du 17e siècle (Financement EUR ArTeC AAP 20-21 40k € plus financement AAP UPL). Dans le cadre du master ArTeC, un module pédagogique de 10 heures a été mis en place pour les étudiants de M2 Technologies et médiations humaines, comprenant la présentation des étapes de création d'un dataset, l'identification des éléments de dialogue entre les personnages dans les textes de Molière, la présentation des étapes d'apprentissage automatique d'un modèle Seq2Seq et la réalisation d'échanges et analyse des répliques du chatbot Litte_bot. Ce module intitulé « Dialoguer avec le chatbot Don Juan inspiré de Molière » a été réalisé dans le cadre d'une journée intitulée « Journée d'étude sur la conception technique et dramaturgique d'un chatbot incarnant un personnage du théâtre » à la BnF, le 22/10/2021.

5.3 Projet MALANTIN soutenu par l'appel à projet européen H2020 RISIS - « MAchine Learning for Analysing Non Technological Innovation »

Porteuse : Anna Pappa

Partenaires : LISIS (Université Paris-Est Marne-la-Vallée), LIGM (Université Paris-Est Marne-la-Vallée) et LIASD (Université Paris 8)

Date de début : 01/2019

Date de fin : 12/2021

Description : Il s'agit d'explorer la possibilité de caractériser l'innovation non technologique de grands acteurs mondiaux en procédant à une analyse textuelle de documents (en ligne et/ou pdf) produits par et sur ces entreprises. La partie réalisée par le LIASD concerne la création de corpus à partir de données extraites du web, et l'apprentissage automatique pour l'annotation automatique sémantique en utilisant un modèle BERT (Financement 132 600 €).

5.4 ANR ReComp « Research on Realtime Compliance Mechanism for AI »

Porteur (à P8) : Pablo Rauzy

Partenaires : Université Paris 8 (France), Sorbonne Université (France), Institut Fur Angewandte Informatik (Allemagne), National Institute of Informatics (Japon).

Date de début : 03/2021

Date de fin : 02/2024

URL : `http://research.nii.ac.jp/RECOMP/`

Description : Le projet ReComp réunit des partenaires de trois pays : LIASD et LIP6 en France, IAI (Institut Fur Angewandte Informatik) en Allemagne, et NII (National Institute of Informatics) au Japon. Il s'agit d'un projet commun co-financé par l'ANR (France), la DFG (Allemagne), et la JST (Japon).

Côté France, le projet ANR (ANR-20-IADJ-0004) a démarré en mars 2021 et s'étale sur 36 mois, son budget total est de 251 200 €, dont 53 200 € à Paris 8, ce qui permet le financement d'un an de postdoc pour travailler à l'intégration du modèle Capacity (développé par Pablo Rauzy) de contrôle des données personnelles dans les cadres d'éthiques modulaires développés par Gauvain Bourgne et Jean-Gabriel Ganascia au LIP6. Il s'agira ensuite d'établir des spécifications éthiques de bonnes pratiques de conception vis-à-vis du traitement des données personnelles et d'aider à la traduction dans un cadre formel de loi spécifique comme le RGPD.

6 Nouveaux projets déposés

6.1 Nouveaux projets académiques

Algorithmes parallèles de preuve pour les problèmes d'informations incomplètes sur les GPU

Projet de financement de thèse (Programme VINCI, Université Franco Italienne, non accepté).

Porteur : Nicolas Jouandeau

Responsable prévu pour la cotutelle : Paolo Ciancarini (Université de Bologne)

Description : Ce projet de thèse a pour objectif de proposer et de mettre en œuvre de nouveaux algorithmes d'apprentissage par renforcement sur GPU.

6.2 Nouveaux projets industriels

Pas de nouveaux projet industriels pour cette année.

7 Activité du groupe par domaine

En synthèse d'images expressives, une réflexion fédérant l'ensemble ou une large partie de l'équipe a été initié autour de la question du « moi numérique ou digital ». L'idée est de pouvoir créer une instance virtuelle de soi nous représentant visuellement. Sans obligation de fidélité, il s'agit de définir plusieurs degrés d'abstractions (fidélité photoréaliste, caricatures et autres abstractions...), ou à l'opposé, de définir un modèle différent du sujet cible (soit par esthétisme, soit par volonté d'anonymisation). Reproduction des mouvements du corps et de la bouche du sujet cible en temps réel, animation à l'image du sujet cible par apprentissage automatique, identification et protection de la vie privée sont à étudier. Dans le cadre d'une collaboration avec le service maxillo-facial du CHU de la Pitié Salepétrière, l'équipe travaille sur l'utilisation d'un casque de réalité augmentée (Microsoft Hololens) pour de l'aide à la chirurgie. L'idée est de superposer l'image virtuelle des limites occultes de la lésion du patient sur le champ opératoire afin de permettre au chirurgien de mieux respecter les marges nécessaires à l'exérèse carcinologique de la tumeur.

En enseignement de l'informatique graphique, nous avons proposé un ensemble d'outils permettant d'aborder progressivement l'apprentissage de la programmation graphique. Notre bibliothèque open-source GL4Dummies développée depuis 2008 permet d'accéder aux ressources graphiques avec des primitives bas-niveau (le pixel) afin d'appréhender les concepts algorithmiques de l'analyse discrète différentielle, permet d'initier à OpenGL/GLSL sur GPU et d'approfondir les connaissances en modélisation et en algorithmie parallèle sur GPU. En termes de développement, les contributions pour cette année se résument à des corrections de bug effectuées par des étudiants dans le cadre d'un cours de développement de logiciels libres et à l'ajout de nouveaux exemples illustrant les notions abordées en programmation graphique.

En résolution de jeux et apprentissage automatique, nous avons proposé une solution d'apprentissage actif pour les séries temporelles et testé une solution d'apprentissage par renforcement pour la résolution du problème du morpion solitaire. Pour réduire les données nécessaires à l'apprentissage profond, nous avons proposé dans [1] une métrique d'évaluation des données utilisées par un apprentissage profond. Cette métrique combinant la pertinence et la représentativité de l'information dans les séries temporelles permet de réaliser un apprentissage actif obtenant des précisions similaires tout en nécessitant 20% de données en moins. Des tests sur quatre datasets de séries temporelles de l'état de l'art [2] ont présenté les courbes de gain d'apprentissage en utilisant 200 à 100k données, avec des gains toujours supérieurs aux meilleurs solutions existantes.

En représentation du raisonnement et logique non classique, nous travaillons sur la représentation et la modélisation de la croyance. Certains acquis formels de la tâche, en particulier les modèles standards de la logique modale, sont questionnés et des solutions alternatives proposées. Ces travaux s'appuient sur l'étude pluridisciplinaire menée, par Jean-Pascal Palus, doctorant ayant débuté ses travaux le 1er janvier 2021. Son étude confronte les points de vue d'informaticiens, de logiciens, de philosophes et psychologues traitant de la question de la croyance. À défaut d'établir un modèle consensuel, l'intégration pluridisciplinaire dans le domaine, permettra la confrontation d'avis et la validation par l'expérience des propositions. Ces travaux s'intègrent à la détermination, clarification et modélisation de la mésinformation (fake-news) et, dans leur aspect dynamique, des rumeurs, buzzes et autres mouvements informationnels.

En langage naturel, nous travaillons sur la création des corpus [3] et l'apprentissage automatique. Pour la création des corpus nous développons des outils (crawlers et scrapers) permettant

d'extraire des données textuelles à partir du web pour créer nos propres datasets pour l'apprentissage. Concernant l'apprentissage automatique nous travaillons sur trois axes :

— la création automatique d'un lexique de termes spécifiques à un domaine, en formalisant la notion d'« information non technologique », en reconnaissant automatiquement les segments correspondants et en particulier leurs emplacements dans les corpus (projet MALANTIN présenté en Section 5.2)
— la génération des dialogues pour des agents conversationnels littéraires (projet Litte_Bot présenté en Section 5.2)
— l'extraction des aspects explicites (Aspect Term Extraction (ATE)) sur des retours d'opinion, pour la prédiction des tendances (thèse cifre avec NOVAGEN correspondant aux travaux de thèse de Maroua Boudabous)

La création des lexiques spécialisés permet une annotation partielle des caractéristiques sémantiques sur des corpus construits automatiquement à partir des données du web. Plus précisément, les termes sont issus de données textuelles brutes multilingues qui sont extraites du web à l'aide d'un scraper web et d'un web crawler. Une première approche pour la création du lexique a été faite en utilisant n-grams et tf-idf. Une deuxième approche est également étudiée en utilisant des techniques d'apprentissage profond. Nous entraînerons un modèle d'apprentissage profond (modèle BERT) pour l'annotation et l'extension d'un lexique « seed »(construit par les experts). Pour la génération des dialogues des agents conversationnels nous utilisons des modèles génératifs de Transformers (Seq2Seq) et actuellement on fait des tests avec des modèles GPT pré-entrainés. Pour l'extraction des aspects explicites nous définissons un pseudo-labeler CRF en utilisant l'apprentissage inter-domaines. Ensuite, nous procédons au marquage des séquences avec BiLSTM-CNN-CRF, en profondeur et enfin, nous utilisons l'apprentissage actif pour faire face à l'absence d'étiquettes.

En dynamique symbolique, nous travaillons sur la notion de reconnaissabilité des morphismes. Nous étudions plusieurs questions liées à la notion de reconnaissable morphisme et à la notion de la dendricité. La principale question que nous étudions consiste à savoir si les langages créés par des morphismes reconnaissables sont des ensembles dendriques.

En sécurité et privacy nous avons développé une bibliothèque permettant la vérification de l'intégrité d'un calcul délégué à un tiers à qui on ne fait pas confiance, typiquement dans les cas d'utilisation de cryptographie homomorphique (qui permet de chiffrer les données envoyées au tiers avant de lui demander d'effectuer des opérations dessus), en mobilisant des propriétés algébriques déjà utilisées pour vérifier l'intégrité de calcul dans le contexte de la sécurité des systèmes cryptographiques embarqués face aux attaques physiques par injection de faute. Ce travail a été publié à la fois dans une conférence académique [4] et comme bibliothèque open source sur le Python Package Index (`https://pypi.org/project/thc`).

8 Présentations et réunions de travail réalisées dans le cadre du séminaire SOIF

Afin de faire émerger de nouvelles collaborations à l'intersection des différents domaines de recherche de l'équipe et d'échanger sur les travaux en cours des membres de l'équipe, il a été décidé de créer un Séminaire d'Ouverture aux Informatiques et de Formation (SOIF).

Patrick Gikunda
Date : 27 avril 2021
Titre : Precipitation Forecasting with Deep Transfer Active Learning for Agricultural Adaptation

Résumé : Weather events are defined by high dimensional data, interacting on many different spatio-temporal and chaotic dynamics. This makes weather prediction a complex and challenging task even when using state of the art numerical weather models. Many statistical models for weather prediction are either built upon human expertise in defining weather events or subjective thresholds of relevant physical variables which are not sufficient for many real world applications. Weather and climate datasets are series of data points indexed in time order with more than one time-dependent variable. Although some semi-supervised learning methods are proposed for univariate time series prediction, there are few deep learning works on multivariate time series prediction. Despite impressive performance of deep learning in many predictive tasks, training a deep learning model is highly dependent on availability of adequate labeled training data. On normal settings, it is expensive to collect and label adequate weather and climate data. In an effort to mitigate the requirement of large labeled dataset, we propose a Transfer Active Learning (TAL) method to emulate the dynamics of a general weather model that's provides forecast of relative short-range time series scale for adaptive agricultural management. Experiments using the proposed method on rainfall and several UCR multivariate datasets achieves a higher prediction accuracy than existing methods, using less training data.

Nicolas Jouandeau
Date : 12 mai 2021
Titre : Résolution des Jeux à Information Imparfaite avec reconnaissance
Résumé : Initiée notamment par des expérimentations d'apprentissage par renforcement profond sur des jeux vidéo ATARI, l'utilisation combinée des algorithmes MCTS et des algorithmes d'apprentissage marque un progrès important de l'IA dans la conception de programmes dits intelligents pour la décision dans les jeux à information complète. Les résultats obtenus nécessitent cependant des ressources de calculs très importantes pour des problèmes de jeux avec des conditions idéales de modélisation. Dans l'optique de l'utilisation de ce type de solution pour des problèmes avec des conditions plus réelles, il s'agit de trouver des solutions d'apprentissage actif et de considérer les jeux à information imparfaite avec reconnaissance.

Pablo Rauzy
Date : 3 juin 2021
Titre : Vérification pratique et efficace de calcul délégué
Résumé : La cryptographie homomorphique est utilisée lorsqu'un calcul est délégué à un tiers non fiable. On fait cependant implicitement l'hypothèse que ce tiers effectuera les calculs demandés, malgré sa supposée non-fiabilité. Cela pose des problèmes de confiance, notamment lorsque les calculs portent sur des données personnelles. Nous proposons un moyen pratique et efficace de vérifier que le calcul délégué à un tiers correspond à la séquence d'opérations attendues, ce qui permet de réduire drastiquement le niveau de confiance nécessaire. Notre approche se base sur la technique bien connue et étudiée de l'extension modulaire. Elle n'est donc pas liée à un cryptosystème homomorphique en particulier, et n'introduit pas nouvelle construction cryptographique qui n'aurait pas encore passé l'épreuve du temps. Nous présentons également une implémentation nommée THC (pour trustable homomorphic computation) que nous utilisons pour analyser les niveaux de sécurité et de performance en pratique. Pour illustrer sa simplicité d'utilisation, nous l'appliquons ensuite dans un système jouet de vote électronique.

Anna Pappa
Date : 17 juin 2021
Titre : Construction automatique d'un lexique spécifique à un domaine à l'aide de N-grammes

Résumé : L'enrichissement des corpus avec des caractéristiques sémantiques aide l'apprentissage automatique à évoluer vers une analyse plus profonde des données, comme la reconnaissance de concepts. Dans cet exposé, nous présentons une méthode pour créer automatiquement un lexique de termes spécifiques à un domaine. Il permet à étiqueter partiellement avec des caractéristiques sémantiques un corpus récolté sur le Web, qui servira de dataset pour d'autres tâches de ML semi-supervisées. Les termes sont issus de données textuelles brutes multilingues, utilisant un modèle probabiliste n-gramme et une mesure tf-idf. Étant donné un ensemble de quatre concepts de base comme la recherche, le développement, l'innovation et la conception, appliqués à vingt-sept catégories industrielles différentes, nous créons d'abord un corpus multilingue composé de descriptions et de rapports de sites Web d'entreprises. Ensuite, nous générons un lexique de termes, selon un concept sémantique spécifique ;le cas de cette étude est "l'innovation". Les tests d'évaluation montrent une grande similitude avec un lexique construit par un expert humain.

Emna Chebbi
Date : 25 octobre 2021
Titre : Classification et détection des attaques dans les réseaux ad-hoc de véhicule suite à une évaluation de protocoles de routage
Résumé : L'évolution des transports vers les véhicules autonomes nécessite des protocoles robustes offrant des garanties sur certaines de leurs propriétés. Les approches formelles permettent de fournir la preuve automatique de certaines propriétés, mais pour d'autres il est nécessaire de recourir à une preuve interactive impliquant le savoir d'un Expert. Mes travaux poursuivent l'objectif d'élaborer, dans le formalisme DEVS (Discrete Event System Specification), des modèles d'un ITS (Intelligent Transportation System) dont la simulation permettrait d'observer les propriétés, éventuellement vérifiées par une approche formelle, dans un scénario plus large et de générer sur les modèles des données susceptibles d'alimenter une boucle de preuve interactive au lieu d'un Expert. Prenant pour cible le protocole CBL-OLSR (Chain-Branch-Leaf in Optimized Link State Routing), l'approche montre comment un modèle DEVS et un modèle formel Event-B équivalents peuvent être construits à partir de la même spécification fonctionnelle d'un réseau ad-hoc où les noeuds utilisent ce protocole. Des propriétés relatives à la sûreté et à la sécurité sont introduites dans le modèle formel Event-B afin d'être vérifiées, puis une méthodologie est proposée afin de les transférer dans un modèle DEVS équivalent sous forme de contraintes, de choix ou d'observables selon des critères proposés. Les résultats de la simulation et de la modélisation exigent un mécanisme avancé permettant la détection des attaques dans le réseaux de communication de véhicules autonome. L'utilisation des algorithmes d'apprentissage automatique permet de détecter et de classifier ces attaques pour réaliser des actions preventives dans le but d'avoir un réseau véhiculaire autonome fiable.

Philippe Guillot
Date : 8 novembre 2021
Titre : Pour un renouveau de VLisp
Résumé : Dès sa création en 1969, l'Université de Vincennes comportait un département d'informatique où une communauté très active, autour de Patrick Greussay, Harald Wertz, Daniel Goossens, et bien d'autres, a conçu et développé une famille interprètes Lisp très réputés. Je présenterai les caractéristiques d'une maquette qui reprend les principes des interprètes originaux tout en prenant en compte plusieurs innovations dans la continuité de MetaVlisp d'Emmanuel Saint-James. Les problèmes posés par la liaison dynamique seront exposés ainsi que leur résolution par un mécanisme complet et correct de fermetures telles qu'exposées par Briot et al. dans un article datant de 1986. L'objectif de ce travail est qu'il soit repris afin que ne soit pas perdu un travail qui a fait l'originalité et la réputation de l'informatique à l'Université Paris 8.

Hanane Zerdoum

Date : 22 novembre 2021

Titre : Problèmes de suites à somme nulle sur les groupes abéliens finis : une approche explicite

Résumé : Les problèmes de suites à somme nulle présentent un thème de la théorie additive des nombres, aussi appelé la combinatoire additive. Parmi les quantités populaires dans la littérature de ces 25 dernières années, il y la constante de Davenport, la constante d'Erdös-Ginzburg-Ziv, la constante de Gao et la constante de Harborth d'un groupe fini. Les résultats les plus aboutis à ce sujet concernent les groupes commutatifs (structure équipée d'une loi interne commutative (l'addition), d'inverse, et d'un élément neutre pour cette opération). Je présenterai des algorithmes performants que nous avons implémenté et qui ont permis de déterminer les valeur de plusieurs constantes dans de nombreux cas auparavant ouverts. En outre, on a pu généraliser un résultat de J.J. Zhuang et W. Schmid sur la constante d'Erdös-Ginzburg-Ziv pour les groupes de la forme Cp+1p $\oplus$ Cp2 où p est un nombre premier. En effet, l'article original excluait le cas p = 2.

Maroua Boudabous

Date : 29 novembre 2021

Titre : WebT-IDC : Un outil pour la création intelligente de datasets à partir du web

Résumé : On présente WebT-IDC, un outil Web conçu pour la création intelligente de datasets, capable de construire des corpus "sans bruit" de commentaires utilisateurs portant sur différents sujets dans différentes langues, à partir de forums Web et de blogs. La méthode est basée sur un modèle d'extraction unique qui se base sur l'élément de pagination, totalement indépendant de la structure DOM. WebT-IDC est un outil holistique couvrant toutes les étapes, de la requête de l'utilisateur, l'exploration de pages Web et l'extraction de données pertinentes, à la création de corpus de texte, sans bruits, utile pour les tâches d'apprentissage statistique. WebT-IDC génère un dataset partiellement étiqueté qui reflète une vision sur les retours d'expérience des utilisateurs. Ce dataset a été utilisé pour entraîner un modèle de type BERT afin de montrer sa pertinence pour une utilisation immédiate dans la tâche d'apprentissage statistique. Les résultats montrent une grande précision dans la catégorisation de la polarité, de la langue et de la reconnaissance du produit. Le système est évalué en termes d'efficacité de filtrage du bruit et de temps de calcul ainsi que la précision et le rappel.

Revekka Kyriakoglou

Date : 13 décembre 2021

Titre : Recognizable morphisms and a decision algorithm for substitutive languages

Résumé : The concept of recognizability of morphisms originates in the paper of Martin 1973 under the term : determinization. This term was first used by Host in his paper on the Ergodic theory of Dynamical Systems. The notion of recognizability came in full bloom after the interest shown by many scientists due to its numerous theoretical applications in various topics, from combinatorics on words to symbolic dynamics. A similar notion is that of circularity. The two terms are often, but not always used as synonymous. This lack of consistency in the literature could lead to confusion. In this seminar, I will present my work on the different notions of recognizability, with the main goal of proving the equivalences and indicating the differences that exist between the different definitions. In the second part of this seminar, I will present an algorithm that allows us to describe all bispecial words of a substitutive language of a recognizable morphism, together with the set of their left and right extensions. More precisely, given a set of words S, one can associate with every word w $\in$ S its extension graph which describes the possible left and right extensions of w in S. Families of sets can be defined from the properties of

the extension graph of their elements : acyclic sets, dendric sets, neutral sets, etc. In the specific case of the set of words of a substitutive language of a recognizable morphism, we show that it is decidable whether these properties are verified or not.

9 Liste des publications pour l'année 2021

[1] P. Gikunda and N. Jouandeau. Homogeneous Transfer Active Learning for Time Series Classification. In *20th International Conference on Machine Learning and Applications*, pages 1–12, Pasadena, California, IEEE, (ICMLA–2021).

[2] P. Gikunda and N. Jouandeau. Sample-Label View Transfer Active Learning for Time Series Classification. In *30th International Conference on Artificial Neural Networks*, pages 1–12, Bratislava, Springer–LNCS, (ICANN–2021).

[3] M. Boudabous and A. Pappa. WebT-IDC : A WebTool for Intelligent Dataset Creation, a use case for forums and blogs. In *25th International Conference Knowledge-Based and Intelligent Information & Engineering Systems*, pages 1051–1060, volume 192, Elsevier, (KES–2021).

[4] A. Nehme and P. Rauzy. THC : Practical and Cost-Effective Verification of Delegated Computation. In *20th International Conference on Cryptology and Network Security*, pages 1–18, Springer–LNCS, (CANS–2021).