LU2IN013 – Projet

Sujets de projets

Sujet 1 : Recommandation et analyse de sous-titres

- Encadrants : Nicolas Baskiotis, Vincent Guigue (`prenom.nom@lip6.fr`)
- Titre : Recommandation et analyse de sous-titres
- Nombre d'étudiants : 2 ou 3
- Description : Les algorithmes de recommandation sont au coeur de nombreux produits industriels : recommandation d'amis dans les réseaux sociaux, de produits dans les sites marchands, de vidéos dans les sites de partage ou de VOD ... Leur objectif est de dépasser les algorithmes usuels de recherche d'informations en proposant des suggestions directement à partir du profil d'un individu, sans forcément attendre une requête explicite sur le contenu. Les algorithmes de recommandation se divisent en deux familles : les algorithmes basés sur le contenu, qui recommandent des choses proches des items déjà visités et les algorithmes de filtrage collaboratif, basés sur la recommandation de produits appréciés par les personnes qui aiment les mêmes choses que la personne cible.
  Nous proposons d'explorer ces différentes stratégies sur la recommandation de séries TV, en exploitant une base de sous-titres en guise de contenu et un ensemble d'avis d'utilisateurs. Le projet se déroulera selon le plan suivant :
  - Prise en main des données textuelles et des outils de traitement de la langue
  - Prise en main des algorithmes de filtrage collaboratif
  - Comparaison des stratégies de recommandation
  - Scrapping de données pour l'enrichissement du contenu et des interactions, et l'évaluation des méthodes mises en oeuvre.

  Il est recommandé d'avoir une connaissance du langage python.

Sujet 2 : Délégation sécurisée du cryptosystème de Rabin

- Encadrant : Damien Vergnaud (`damien.vergnaud@lip6.fr`)
- Titre : Délégation sécurisée du cryptosystème de Rabin
- Nombre d'étudiants : 2 ou 3
- Description : Les primitives cryptographiques à clé publique sont souvent trop coûteuses (du point de vue algorithmique) pour des dispositifs de calcul à ressources limitées (cartes à puces, puces RFID, ...). Une approche classique consiste donc à déléguer de façon sécurisée des opérations cryptographiques d'un appareil ayant des capacités de calcul (relativement) faibles — le client — vers un appareil plus puissant mais *a priori* non fiable — le serveur. Un protocole de délégation doit généralement répondre à deux objectifs de sécurité : *la confidentialité* – les informations secrètes du client ne doivent pas être révélées au serveur — et *la vérifiabilité* — un serveur malveillant ne doit pas pouvoir faire accepter au client une valeur invalide comme résultat du calcul délégué. En 2020, un protocole de délégation du déchiffrement (et de la signature) de Rabin a été proposé par Hanlin Zhang, Jia Yu, Chengliang Tian, Guobin Xu, Peng Gao et Jie Lin (IEEE Internet Things J). Cette proposition n'atteint cependant pas les propriétés de sécurité attendues d'un tel protocole. Les objectifs de ce projet PIMA sont les suivants :

  1. comprendre les notions de base de l'arithmétique du cryptosystème de Rabin ;

2. exhiber les failles de sécurité du protocole de Zhang *et al.* en proposant et en implantant une ou plusieurs attaques ;
3. étudier les algorithmes connus d'extraction de racines carrées modulaires ;
4. adapter les idées de ces algorithmes pour proposer un nouveau protocole de délégation sécurisé pour le cryptosystème de Rabin ;
5. analyser et prouver la sécurité du protocole de délégation construit ;
6. implanter le protocole construit et étudier le gain d'efficacité obtenu.

Sujet 3 : Algorithmes de factorisation de polynôme

— Encadrant : Jérémy Berthomieu (`jeremy.berthomieu@lip6.fr`)
— Titre : PPIMA : Algorithmes de factorisation de polynôme
— Nombre d'étudiants : 2 ou 3.
— Description :

Factorisation sans carré.

En calcul scientifique, beaucoup de méthodes calculant des solutions exactes ou approchées d'un polynôme, comme l'algorithme de Sturm, la méthode de Newton ou l'algorithme de factorisation de Berlekamp, supposent que le polynôme est sans facteur carré. C'est-à-dire qu'elles supposent que le polynôme n'a pas de racines multiples. Dans le cas où le polynôme n'est pas sans facteur carré, ces algorithmes peuvent échouer ou renvoyer un résultat partiel.
La factorisation sans carré d'un polynôme P est la factorisation de P sous la forme $P = P_1\,P_2^2 \cdots P_r^r$ où les P_i sont premiers entre eux deux à deux et sont sans facteurs carrés. Autrement dit, les racines de P_i sont exactement les racines de P qui ont multiplicité i.
Un premier but de ce projet est de comprendre des algorithmes de factorisation sans carré, comme l'algorithme naïf ou l'algorithme de Yun, sur les rationnels ou sur un corps fini.
Une implémentation des ces algorithmes sera effectuée en `C`.

Factorisation irréductible.

Un polynôme irréductible est un polynôme non constant qui n'admet pas de facteurs non triviaux. Les polynômes irréductibles sont donc les analogues des nombres premiers dans l'anneau des polynômes.
L'algorithme de Berlekamp permet de calculer la factorisation irréductible d'un polynôme sans facteur carré à coefficients dans un corps fini. C'est-à-dire qu'il permet de trouver tous les facteurs irréductibles du polynôme passé en entrée.
Un second objectif de ce projet sera de comprendre cet algorithme et de l'implémenter en `C` afin d'obtenir un algorithme complet de factorisation irréductible d'un polynôme quelconque à coefficients dans un corps fini.

Sujet 4 : Génération automatisée de quiz d'algorithmique en ligne

— Encadrant : Olivier Spanjaard (`olivier.spanjaard@lip6.fr`)
— Titre : Génération automatisée de quiz d'algorithmique en ligne
— Nombre d'étudiants : 2 ou 3
— Description : Ce projet a pour objectif d'initier le développement d'un outil de génération automatisée de quiz en ligne pour l'UE d'algorithmique de L3. Le quiz mêlera des questions de type QCM, tirées d'une base de données de questions, classées par catégories, et des questions de mise en oeuvre d'algorithmes classiques sur des données générées aléatoirement (parcours en largeur et en profondeur, algorithmes de Dijkstra, de Prim, de Huffman, de Bellman, de Bellman-Ford, etc.). Chaque utilisateur se verra proposer un questionnaire différent, généré

de façon automatisée en « piochant », dans un ordre arbitraire, des questions de type QCM et des questions de mise en oeuvre, tout en respectant néanmoins un certain équilibre sur les sujets abordés. A l'issue de la session en ligne, les réponses seront soumis pour évaluation. Cette dernière se fera de façon automatique. Le projet se déroulera en plusieurs phases : dans une première phase, vous serez amenés à examiner les outils logiciels disponibles pour mettre en place une telle plate-forme ; dans une deuxième phase, une fois le dispositif logiciel choisi, vous concevrez une « architecture logicielle » pour la plate-forme ; enfin, la dernière phase correspondra bien sûr au développement proprement dit de la plate-forme. Le projet étant d'envergure, il est vraisemblable que le prototype développé ne sera pas entièrement finalisé au terme de l'année : il devra donc être codé de façon suffisamment claire et commentée pour pouvoir être poursuivi par d'autres ultérieurement. Ce projet mêle à la fois un aspect programmation et un aspect mathématique, lié à la compréhension des algorithmes qui feront l'objet des questionnaires (ce qui vous permettra de prendre un peu d'avance sur l'UE d'algorithmique de l'année prochaine).

Sujet 5 : Continuous-Time Bayesian Network (CTBN) : étude et implémentation

- Encadrant : Pierre-Henri WUILLEMIN (`pierre-henri.wuillemin@lip6.fr`)
- Titre : Continuous-Time Bayesian Network (CTBN) : étude et implémentation
- Nombre d'étudiants : 2 ou 3
- Les réseaux bayésiens (BNs) sont un modèle probabiliste qui s'appuie sur un graphe (orienté sans cycle) pour représenter une distribution jointe d'un grand nombre de variables aléatoires. Ce modèle à la fois numérique (distribution) et qualitatif (graphe) est un point de contact intéressant entre probabilités, statistiques et intelligence artificielle. Il permet d'implémenter des outils de raisonnement, de calcul de fiabilité, d'explications causales, mais aussi d'apprentissage statistiques et des outils de classification (machine learning, etc.). Le but de ce projet est d'étudier un modèle issus des BNs permettant la représentation de processus stochastique en temps continu. Les différentes tâches de ce projet seront : 1- état de l'art sur le domaine (UAI2002) 2- étude et implémentation du modèle CTBN dans la librairie aGrUM (http ://a-grum.org) qui propose un grand nombre des composantes de base pour l'implémentation d'un tel modèle en python 3 [optionel]- implémentation d'un algorithme d'apprentissage de CTBN (PGM2020)
- Bibliographie
  [UAI2002] U. Nodelman, C. R. Shelton, and D. Koller. *Continuous Time Bayesian Networks.* In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, pages 378–387, 2002.
  [PGM2020] Alessandro Bregoli, Marco Scutari and Fabio Stella *Constraint-Based Learning for Continuous-Time Bayesian Networks*, in PGM'20

Sujet 6 : Automatisation des cryptosystèmes classiques à l'aide d'algorithmes modernes

- Encadrante : Valérie Ménissier-Morain (`Valerie.Menissier@lip6.fr`)
- Titre : Automatisation des cryptosystèmes classiques à l'aide d'algorithmes modernes
- Nombre d'étudiants : 2 ou 3
- Description : Les cryptosystèmes classiques ont été utilisés de l'Antiquité à la seconde guerre mondiale pour protes affaires privées, déger les secrets militaires et civils, des affaires d'état ou

des puissants et des humbles : chiffrements par substitition, par transposition, homophoniques, de Vigenère, de Playfair, ADFGVX, etc. Le cassage d'un texte chiffré par un tel système de chiffrement s'effectuait donc manuellement avec une technique plus ou moins aboutie selon les époques.

Depuis une trentaine d'années un certain nombre de techniques modernes ont été utilisées pour essayer d'automatiser la cryptanalyse de ces cryptosystèmes. Qu'il s'agisse de recuit simulé, d'algorithmes génétiques ou plus récemment de *hill climbing* et autres méta-heuristiques propres à la fouille de donnnées, ces techniques combinent une progresssion plus ou moins guidée et plus ou moins aléatoire avec une fonction d'évaluation de la qualité de la solution (*fitness function*).

Le but de ce projet est de systématiser cette exploration d'une part en comparant plusieurs techniques sur un même cryptosystème et d'autre part d'essayer une même technique sur plusieurs systèmes de chiffrement. Le point de départ sera le *hill climbing* appliqué à la cryptanalyse des substitutions mono-alphabétiques en testant plusieurs fonctions d'évaluation et en faisant varier le nombre d'itérations. Ensuite l'exploration ira en s'élargissant en alternant les phases de familiarisation avec les cryptosystèmes et les méta-heuristiques, vous progresserez dans votre comparaison expérimentale de ces techniques sur ces systèmes de chiffrement. Cette seconde phase pourra être exécutée en parallèle selon les affinités de chacun.

Sujet 7 : Ordonnancement avec des données incertaines

- Encadrants : Evripidis Bampis et Konstantinos Dogeas (`evripidis.bampis@lip6.fr`)
- Titre : Ordonnancement avec des données incertaines
- Nombre d'étudiants : 2 ou 3
- Description :

  Dans ce projet, nous allons revisiter un problème d'ordonnancement des tâches dans lesquel les données que l'on dispose ne sont exacts. Ainsi, on supposera par exemple qu'au lieu de disposer la vrai valeur du travail à effectuer pour une tâche, on ne dispose que d'une borne supérieure de sa valeur. On peut connaitre la vraie valeur du travail à effectuer pour une tâche en effectuant une requête. Or, cette requête a un coût (connu à l'avance) que l'on doit prendre en compte. L'algorithme a donc le choix soit d'utiliser la borne supérieure sans faire la requête, soit de faire la requête en payant le coût et ainsi utiliser la vraie valeur du travail. Nous proposons de mettre en place certains algorithmes pour traiter ce problème et comparer leurs performances.

Sujet 8 : Calcul de pseudospectres

- Encadrant : Stef Graillat (`stef.graillat@lip6.fr`)
- Titre : Calcul de pseudospectres
- Nombre d'étudiants : 2 ou 3
- Description :

  Le ε-pseudospectre d'une matrice A est défini comme le sous-ensemble du plan complexe consistant en toutes les valeurs propres de toutes les matrices situées à une distance ε de A. C'est un outil très utilisé en théorie du contrôle et en automatique pour tester la robustesse de la stabilité d'un système.

  Considérons maintenant une matrice $A \in M_n(\mathbb{C})$. Nous notons par $\Lambda(A)$ son spectre. Étant donné $\varepsilon > 0$, le ε-pseudospectre de la matrice $A \in M_n(\mathbb{C})$ est l'ensemble $\Lambda_\varepsilon(A)$ défini par

$$\Lambda_\varepsilon(A) = \{z \in \mathbb{C} : z \in \Lambda(X) \text{ avec } X \in M_n(\mathbb{C}) \text{ et } \|X - A\|_2 \leq \varepsilon\}.$$

On peut montrer que

$$\Lambda_\varepsilon(A) = \{z \in \mathbb{C} : \sigma_{\min}(A - zI) \leq \varepsilon\},$$

où $\sigma_{\min}$ représente la plus petite valeur singulière. Cela donne un algorithme de calcul du pseudospectre connu sous le nom de `GRID`.
Cet algorithme est massivement parallèle. En effet, il revient à calculer de manière indépendante une SVD (décomposition en valeurs singulières) de $A - zI$ pour chaque point z de la grille. Un tel algorithme devrait donc pleinement tirer parti des architectures parallèles. Néanmoins, il nécessite beaucoup de calcul de valeurs singulières en des points qui ne font pas partie du pseudospectre. Une méthode basée sur un algorithme de prédiction-correction (suivi de trajectoire) a été proposée pour pallier ce problème.

Le travail pourra se dérouler de la manière suivante.

1. Étude théorique des pseudospectres et de la décomposition en valeur singulière (complexité, algorithme de calcul en particulier).
2. Implantation de l'algorithme `GRID` en Python. Proposer une version parallèle de cet algorithme.
3. Implantation en Python de l'algorithme de prédiction-correction. Comparaison en terme de performance et de parallélisation avec `GRID`.
4. Étendre ces algorithmes à la notion de pseudospectres par composante.

Sujet 9 : énumération de motifs dans les graphes

— Encadrant : Lionel Tabourier (`lionel.tabourier@lip6.fr`)
— Titre : énumération de motifs dans les graphes
— Nombre d'étudiants : 2 ou 3
— Description :
Les graphes représentant des réseaux réels, tels que des réseaux sociaux, des réseaux d'infrastructure ou encore des réseaux d'interactions biologiques, présentent une structure souvent vaste et complexe dont il est difficile de rendre compte à l'aide d'une visualisation. Pour décrire ce type d'objet, on a donc recours à des mesures structurelles qui permettent de quantifier les caractéristiques de ces graphes et notamment d'évaluer si les observations correspondent à ce que les modèles prédisent ou non.

Parmi ces mesures structurelles, les motifs dans les graphes, tels que les triades, les cycles, les cliques, etc. que l'on regroupe parfois sous la dénomination de *graphlets*, présentent un intérêt particulier car ils permettent de comprendre les processus dynamiques qui expliquent la structure des graphes réels observés.

Le projet propose de créer des algorithmes d'énumération de motifs, puis d'étudier le passage à l'échelle de ces algorithmes. On étudiera ensuite dans différents contextes (e.g. graphes de réseaux sociaux, graphes de réseaux biologiques) s'il existe des motifs sur- ou sous-représentés dans ces graphes lorsqu'on les compare aux valeurs attendues selon certains modèles simples.

Sujet 10 : Amélioration de la méthode de recherche de similarité structurale dans une banque Yakusa.

— Encadrant(e) : Mathilde Carpentier (`mathilde.carpentier@sorbonne-universite.fr`)

— Titre : Amélioration de la méthode de recherche de similarité structurale dans une banque Yakusa.
— Nombre d'étudiants : 2 ou 3
— Description : YAKUSA [1] est une méthode permettant de trouver très rapidement les similitudes structurales entre une structure dite requête et toutes les structures d'une banque. Elle est basée sur une description simple de la chaîne polypeptidique (les angles alpha [2]) et est inspirée de l'algorithme de Aho-Corasick [3]qui permet de rechercher en parallèle plusieurs motifs dans un texte. Cette méthode permet de trouver des blocs structuraux sans gap communs entre la protéine requête et les structures de la banque, de calculer un score selon la probabilité de trouver ces blocs dans la banque et d'ordonner les structures de la banque selon leur similitude avec la structure requête. Le service a été intégré dans la plateforme RPBS (Ressource parisienne en bioinformatique structurale) qui permet l'accès à plusieurs ressources originales en bioinformatique structurale [4]. Il est accessible via une interface web et téléchargeable à l'adresse http ://bioserv.rpbs.jussieu.fr/Yakusa.
La recherche de motifs avec l'algorithme de Aho-Corasick est très efficace mais un trop grand nombre de motifs sont identifiés. Il faut ensuite les sélectionner selon des critères de score et de compatibilité structurale. L'objectif de ce projet est de développer un algorithme pour cela, de l'intégrer dans le programme puis de tester les gains de performance.

1. CARPENTIER, M., BROUILLET, S. et POTHIER, J. (2005). Yakusa : a fast structural database scanning method. Proteins, 61(1) :137–51.
2. LEVITT, M. A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding. J. Mol. Biol. 104, 59-107 (1976).
3. AHO, A. et CORASICK, H. (1975). Efficient string matching : an aid to bibliographic search. Comm. ACM, 18(6) :333– 340.
4. ALLAND, C., MOREEWS, F., BOENS, D., CARPENTIER, M., CHIUSA, S., LONQUETY, M., RENAULT, N., WONG, Y., CANTALLOUBE, H., CHOMILIER, J., HOCHEZ, J., POTHIER, J., VILLOUTREIX, B. O., ZAGURY, J. F. et TUFFERY, P. (2005). RPBS : a web resource for structural bioinformatics. Nucleic Acids Res, 33 :W44–9.

Sujet 11 : Ordonnancements équitables

— Encadrant : Fanny Pascual (`fanny.pascual@lip6.fr`)
— Titre : Résolution et génération de puzzles Picross
— Nombre d'étudiants : 2 ou 3
— Description :
Les problèmes d'ordonnancement prennent comme données un ensemble de tâches, un ensemble de machines, et une fonction objectif (un critère à optimiser). Ils consistent à exécuter les tâches sur les machines de façon à optimiser la fonction objectif, sachant qu'une machine ne peut exécuter qu'une seule tâche à la fois. Par exemple, chaque tâche peut avoir une durée d'exécution et une date d'échéance (ou deadline) à laquelle elle doit être exécutée. Dans ce cas, le but peut être d'exécuter les tâches de manière à minimiser le nombre de tâches en retard (se terminant après leur date d'échéance), ou bien de minimiser le retard moyen d'une tâche. Ces problèmes ont été très étudiés depuis une cinquantaine d'années, du fait de leurs nombreuses applications pratiques. On s'intéresse cependant depuis assez peu de temps au cas où les tâches appartiennent à plusieurs utilisateurs, aux intérêts divergents. Le but de ce projet est de coder (et éventuellement de concevoir) des algorithmes d'ordonnancements en présence de plusieurs utilisateurs qui chacun possèdent des tâches à ordonnancer sur des machines partagées. On s'intéressera tout d'abord à coder certains algorithmes classiques d'ordonnancement, puis à

étendre ces algorithmes au cas où il y a plusieurs utilisateurs afin de retourner des ordonnancements équitables (plusieurs notions d'équité pourront être utilisées). Une interface graphique permetant de visualiser les ordonnancements obtenus pourra être proposée.