VIỆN QUỐC TẾ PHÁP NGỮ
Institut Francophone International

Travaux Personnels Encadrés (TPE)
Promotion 20, 2016-2017

Coordinateur: HO Tuong Vinh
Email: ho.tuong.vinh@ifi.edu.vn

Liste des sujets

Travaux Personnels Encadrés (TPE)
Promotion 20, 2016-2017

1. Sujet 1 : Diffusion d'opinions dans les réseaux sociaux : l'évacuation d'une foule

Encadrement

- Dominique LONGIN (IRIT, Toulouse, France)
- HO Tuong Vinh (IFI, Hanoi)

Contexte

Depuis quelques années, les réseaux sociaux sont très étudiés car ils sont générateurs d'un grand nombre d'interactions entre les personnes, et celles-ci ont pour effet une diffusion rapide des opinions, des idées, des goûts, des modes, *etc.*

Lorsqu'on se retrouve au sein d'une foule notamment, par exemple lors de l'évacuation d'un lieu en proie aux flammes, la direction qu'on choisit de prendre dépend de la direction des personnes (qu'on appelle nos « influenceurs ») se trouvant directement dans notre entourage et qui peuvent ne pas être toutes les mêmes à chaque instant (une certaine proportion de cet entourage change à chaque fois).

On se propose de modéliser mathématiquement ce mécanisme de diffusion à l'aide de la logique propositionnelle puis de l'implémenter.

Travaux théoriques

Ils concernent différents aspects :

- recherches bibliographiques (dont l'article mentionné ci-dessous peut être un point d'entrée pour trouver d'autres articles sur le sujet) dans le domaine de la diffusion d'opinion ;
- rafraichissement des connaissances en logique propositionnelle, notamment les tableaux de valeurs (modèles) associés aux formules
- modélisation d'une contrainte de type « on ne peut aller que dans une direction donnée » ;
- formalisation de l'opinion d'un agent et du mécanisme de mise à jour en passant de l'instant t à l'instant $t + 1$.
- On pourra par la suite réfléchir à d'autres modèles d'influence (voir la littérature à ce sujet).

Travaux pratiques

- Implémenter un modèle sous JAVA en faisant varier différents paramètres (nombre d'agents, nombres d'influenceurs, proportion des influenceurs renouvelés chaque fois, distribution initiale des opinions, *etc.* par exemple) et analyser les résultats (tout le monde fuit-il dans la même direction ? dans tous les cas ? dans le même laps de temps ? *etc.*)

- Implémenter le modèle précédent dans l'architecture GAMA avec visualisation spatiale des agents et passage d'un ensemble discret de direction à un ensemble continu.

Références

[1] U. Grandi, E. Lorini, L. Perrussel (2016). « Propositional opinion diffusion ». In Proceedings of international conference on Autonomous Agents and Multi-Agent Systems (AAMAS-2016).

2. Sujet 2: MODÉLISATION DE LA CROISSANCE TUMORALE A L'AIDE D'UN AUTOMATE HYBRIDE

Encadrement

- Alexandra Fronville (LATIM), Bernard Pottier (Lab-Sticc)
- Ho Tuong Vinh (IFI)

Contexte

Comprendre le fonctionnement et les lois de la division et de la croissance cellulaire est primordiale dans la cancérogenèse afin de déterminer le traitement le mieux adapté à chaque cancer. En effet, la complexité du système des cellules tumorales fait que la tumeur n'est plus considérée comme une entité homogène. On parle d'une hétérogénéité intra-tumorale [4]. Plusieurs hypothèses tentent d'expliquer ce phénomène [5]. Du fait de la présence de différentes zones tumorales, les tumeurs deviennent plus résistantes aux traitements délivrés. Il est donc essentiel, pour mieux cibler et adapter les chimiothérapies et la radiothérapie, de mieux comprendre les zones d'hétérogénéité des tumeurs [6].

Les automates cellulaires ou les systèmes multi-agents permettent d'expérimenter l'impact de changements de dynamique cellulaire sur deux états temporels d'une forme tumorale. Basé sur un modèle mathématique formalisant les mécanismes de la dynamique cellulaire, il permettent d'expérimenter des hypothèses de différenciation cellulaire basée sur les contraintes spatiales [2] et le rôle de l'environnement sur la différenciation cellulaire et la topologie de la tumeur.

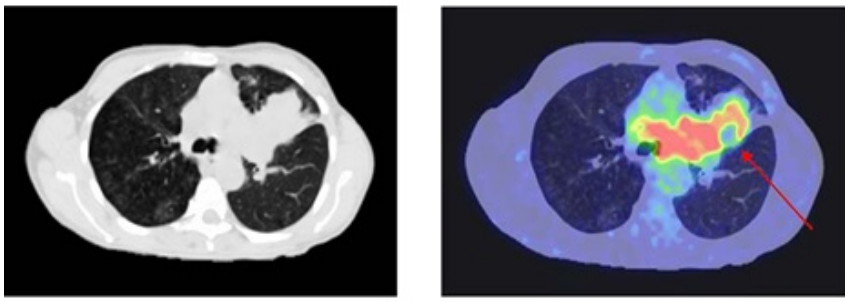

Figure 1 : Image CT et TEP 1

Ce sujet propose de simuler des croissances tumorales à partir d'images TEP et CT en utilisant les données physiologiques de l'hétérogénéité intra-tumorale. On se propose de surveiller l'impact des conditions modifiées (*e.g.* hypoxie) sur la dynamique cellulaire. Ces données fourniront des informations spatiales et temporelles 4D pour certaines lignées cellulaires et permettront de tester ces paramètres dans le simulateur de cellules saines et tumorales et ainsi mieux comprendre le lien entre dynamique cellulaire, différentiation cellulaire et forme tumorale.

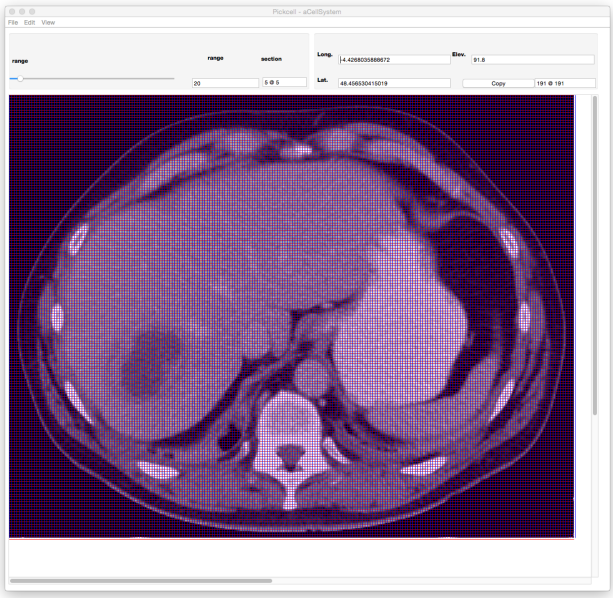

Figure 2 : Initialisation de l'automate 1

Travaux théoriques:

- Etude des automates cellulaires et les systèmes multi-agents

Travaux pratiques:

- Prise en main des outils PickCell/Netgen
- Découverte de la programmation des automates cellulaires produits par ces outils.
- Création d'AC adaptés au développement cellulaire hétérogène.

- Traces, historiques d'évolutions, représentation graphique.

Références

[1] Fronville, F. Harrouet, A. Desilles, P. Deloor, Simulation tool for morphological analysis, in: ESM 2010, 2010, pp. 127–132.

[2] Fronville, A. Sarr, P. Ballet and V. Rodin.Mutational analysis-inspired algorithms for cells self-organization towards a dynamic under viability constraints. SASO 2012, 6th IEEE International Conference on Self-Adaptive and Self-Organizing Systems, pages 181-186, Lyon (France), 10-14 September 2012.

[3] Sarr, A. Fronville, P. Ballet and V. Rodin. French flag tracking by morphogenetic simulation under developmental constraints. CIBB 2013, 10th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, Nice (France), 20-22 june, 2013. Accepted in the post-proceedings Springer series of Lecture Notes in Bioinformatics (LNBI), 2014.

[4] M. Shibata, M.M. Shen. The roots of cancer: stem cells and the basis for tumor heterogeneity. Bioessays. 2013, 35(3): 253-260.

[5] R.A Gatenby and R.J. Gillies. Why do cancers have high aerobic glycolysis? Nat Rev Cancer. 2004;4(11):891-899.

[6] T.S. Gerashchenko, E.V. Denisov, N.V. Litviakov, M.V. Zavyalova, S.V. Vtorushin, M.M. Tsyganov, V.M. Perelmuter, N.V. Cherdyntseva. Intratumor heterogeneity: nature and biological significance. Biochemistry. 2013; 78(11): 1201-1215.

3. Sujet 3: VietVoice

Encadrant:

- DO Phan Thuan
  Institut Polytechnique de Hanoï
  Email à: thuandp@soict.hust.edu.vn

Collaboration externe:

- DAO Manh Cuong, LEXRAY.COM, USA

Mots-clé :
In English: Machine Learning, Deep Learning, Algorithms, Vietnamese Language, Signal Processing, Recurrent
Neuron Network (RNN)

En français: Apprentissage artificiel, Apprentissage profond, Algorithmes, Langage Vietnamienne, Traitement du Signal, Réseau de Neurones Récurrents

Contexte

Les meilleurs systèmes de reconnaissance vocale reposent sur des pipelines complexes composées de plusieurs algorithmes et des étapes de traitement conçues à la main. Dans ce projet, on souhaite construire un système vocal de

bout en bout pour le langage vietnamien, appelé “Vietvoice”, où l’apprentissage profond remplace ces étapes de traitement. L’approche est inspiré du système “Deep Speech” en [1, 2]. Pour des modèles de langage quelconque, cette approche permet d’obtenir des performances plus élevées que les méthodes traditionnelles sur des tâches difficiles de reconnaissance vocale, tout en étant beaucoup plus simple. Les résultats seront rendus possibles par l’entraînement d’un réseau large de neurones récurrents (RNN) en utilisant plusieurs GPUs et des milliers d’heures de données. Parce que ce système apprend directement à partir des données, on ne demande pas de composants spécialisés pour l’adaptation du locuteur ou de filtrage de bruits.

Travaux théoriques :

- Réviser un cours l’apprentissage artificiel
- Étudier quelques pistes de l’apprentissage profond
- Étudier sérieusement les articles [1] et [2]

Travaux pratiques :

- Implémenter quelques fonctions basics du système proposé
- Collecter quelques données de audios/textes vietnamiens
- Analyser les résultats obtenus
- Faire un rapport détaillé
- (Facultatif) Si les résultats sont publiables, rédiger un article pour une conférence internationale

Remarque: Deep Speech était un de 10 meilleurs technologies innovatrices en 25e Février 2016 BaiduResearch

Références

[1] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. ArXiv e-prints, Dec. 2015.
[2] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep Speech: Scaling up end-to-end speech recognition. ArXiv e-prints, Dec. 2014.

4. Sujet 4: Détection/reconnaissance des objets et suivi d'objets fondé sur la transformée de Hough généralisée

Encadrant :

- Thanh Phuong NGUYEN, Maître de conférences, Laboratoire des Sciences de l’Information et des Systèmes (LSIS) - Université de Toulon, France
- Ho Tuong Vinh, IFI

Contexte et objectifs de l'étude

Dans le contexte d'un système de vision embarqué sur un porteur mobile (ex : véhicule, ...), il s'agit d'abord de concevoir un modèle représentant de façon conjointe un ensemble d'objets physiques (ex: voitures, piétons, signalisation,...), et la scène perçue dans son ensemble (ex: sol, bâtiments, bas-côté,...). Ce modèle sera associé à un algorithme de détection dont l'objectif est de reconnaître et localiser les objets connus en temps réel sur le porteur mobile. Le cadre scientifique de base retenu pour l'élaboration du modèle est une version spatialement variante des méthodes de Hough généralisées, où le vote dans l'espace des caractéristiques s'effectue de façon différente selon l'endroit où on se trouve dans la scène, ce qui implique une structuration - implicite ou explicite - de la scène observée, de nature géométrique et/ou sémantique.

Les transformées de Hough, que nous proposons de retenir comme base scientifique pour ce sujet, sont une des classes de techniques les plus anciennes en vision par ordinateur [Hough59]. Elles sont fondées sur la représentation d'une forme ou d'un objet sous la forme d'un ensemble de paramètres incluant la localisation spatiale, de sorte qu'un point dans l'espace des paramètres correspond à une instance de l'objet localisé dans l'image. La construction du modèle se fait à partir d'un ensemble de prototypes qu'on recueille dans une phase préliminaire qui s'assimile à un apprentissage. La détection, elle consiste à examiner tous ou une partie des points de l'image et à réaliser pour chacun des votes dans l'espace des paramètres, de telle sorte que les points ayant recueilli le plus de suffrages correspondent aux localisations les plus probables de l'objet. Il est remarquable que le succès de ces méthodes ne se soit jamais démenti et que des variations de toutes sortes aux méthodes de Hough généralisées aient été proposées encore récemment [LLS08, GYRGL11]. Il faut cependant noter que les techniques classiquement utilisées présentent deux défauts qui peuvent s'avérer déterminants dans notre contexte. Le premier est de nature computationnelle, de par le coût du calcul pour obtenir l'ensemble des points qui vont participer au vote, ainsi que le coût en mémoire de la représentation du prototype. Le second est une réserve de robustesse liée au caractère épars des votes qui peut devenir rédhibitoire lorsque la scène présente une structure plus pauvre en termes de contours ou de texture.

Dans une étude récente [MNL16], nous avons montré qu'il était possible d'obtenir de bonnes performances avec des transformées de Hough où le vote se réalise massivement pour tous les pixels sans transformation de l'image de type contours, points saillants ou régions, directement à partir des dérivées partielles estimées localement à l'ordre 1 (e.g. direction du gradient) et à l'ordre 2 (e.g. courbure de l'isophote). Cette propriété répond en grande partie aux problèmes évoqués plus haut, à la fois par la limitation du temps de calcul et sur le caractère dense du vote qui s'ensuit. Le caractère extrêmement régulier et parallèle des calculs se prête d'autre part à une implantation temps réel plus efficace.

Nous proposons dans le cadre de ce travail d'adapter ces techniques de Hough généralisées à une modélisation conjointe objet-scène, où la construction du prototype prend en compte l'aspect de l'objet de façon relative à sa position d'occurrence, de telle sorte que, pour la détection, le vote dans l'espace des

caractéristiques s'effectue de façon différente selon l'endroit où on se trouve dans la scène, ce qui implique une structuration - implicite ou explicite - de la scène observée, de nature géométrique et/ou sémantique. Cette structuration peut être rendue temporellement variable en fonction d'une identification du contexte, qui peut être locale, régionale ou globale.

Travaux théoriques
- Étudier les techniques de Hough généralisées
- Étudier la modélisation conjointe objet-scène

Travaux pratiques
- Construire un prototype

Personne à contacter
- Thanh Phuong NGUYEN, McF, Université de Toulon, France, Equipe SIIM, LSIS Email : thanh-phuong.nguyen@univ-tln.fr
- Site web : http://tpnguyen.univ-tln.fr/

Références
[Hough59] P.V.C. HOUGH : Machine analysis of bubble chamber pictures. In Int. Conf. High Energy Accelerators and Instrumentation, pages 554–556, 1959.
[GYRGL11] Juergen GALL, Angela YAO, Nima RAZAVI, Luc J. Van GOOL et Victor S. LEMPITSKY : Hough forests for object detection, tracking, and action recognition. pages 2188–2202, 2011.
[LLS08] Bastian LEIBE, Ales LEONARDIS et Bernt SCHIELE : Robust object detection with interleaved categorization and segmentation. International Journal of Computer Vision, 77(1- 3):259–289, 2008.
[MNL16] Antoine MANZANERA, Thanh Phuong NGUYEN et Xiao LEI : Evaluation of the one-to-one dense hough transforms for line and circle detection. Soumis à Pattern Recognition.

5. Sujet 5: PAQUET R POUR LES DÉCOMPOSITIONS EN ONDELETTES

Encadrement :
- Marc Choisy et Le Viet Thanh
  Collaboration externe : Institut de Recherche pour le Développement (IRD) et Oxford University Clinical Research Unit (OUCRU), Hanoi.
- Langues de communication : français (MC), anglais (MC, LVT) et vietnamien (LVT).

Contexte
R est un langage interprété libre créé en 1993 sur la base du langage S créé en 1976. Initialement développé comme un langage dédié aux statistiques, il est récemment devenu un langage généraliste en science comme peut l'être Python. Le succès de R tient en grande partie à la puissance de son organisation en paquets développés par les utilisateurs pour accomplir des tâches spécifiques.

De 40 paquets en 2000, le dépôt CRAN (cran.r- project.org) en compte aujourd'hui plus de 8000.
La décomposition en ondelettes est une méthode d'analyse de séries temporelles, particulièrement puissante pour l'analyse de série temporelles non-stationnaires. D'abord développée dans les années quatre-vingt pour des applications d'ingénierie (par exemple pour les algorithmes de compression), cette méthode d'analyse à récemment été appliquée à de nombreux domaines tels que la climatologie, l'océanographie, la biologie ou la médecine. Il existe à ce jour 6 paquets R permettant de faire des analyses en ondelettes mais (1) leur utilisation n'est pas simple, (2) aucun ne propose toutes ondelettes existantes, (3) ils sont généralement lents et (4) ils ne contiennent pas les derniers avancements, notamment en terme de test de significativité.

Travaux théoriques

- Faire une revue systématique des paquets faisant de la décomposition en ondelettes actuellement disponibles dans R, en comparant leurs fonctionnalités sous la forme d'un tableau ;
- Faire un état de l'art des avancées les plus récentes dans le domaine de l'analyse en ondelettes.

Travaux pratiques

- Développer un paquet R comblant les manques identifiés plus haut. Le paquet sera développé pour partie en C++ pour en augmenter la rapidité de calcul et sera interfacé avec R grâce au paquet Rcpp.

Références:

1 Cazelles, B., K. Cazelles, and M. Chavez. 2014. Wavelet analysis in ecology and epidemiology: impact of statistical tests. J R Soc Interface 11:20130585.
2 Cazelles, B., M. Chavez, D. Berteaux, F. Ménard, J. O. Vik, S. Jenouvrier, and N. C. Stenseth. 2008. Wavelet analysis of ecological time series. Oecologia 156:287–304.
3 Cazelles, B., M. Chavez, G. C. de Magny, J.-F. Guégan, and S. Hales. 2007. Time- dependent spectral analysis of epidemiological time-series with wavelets. J R Soc Interface 4:625–636.
4 Torrence, C., and G. P. Compo. 1998. A practical guide to wavelet analysis. Bulletin of the American Meteorological Society 79:61–78.
5 Wickham H. 2015. Advanced R, Chapman & Hall CRC.
6 Wickham H. 2015. R Packages, O'Reilly.

6. Sujet 6 : PAQUET R POUR L'ANALYSE DE DONNÉES SÉROLOGIQUES ET DE CONTACT SOCIAUX

Encadrement :

- Marc Choisy et Le Viet Thanh
  Collaboration externe : Institut de Recherche pour le Développement (IRD) et Oxford University Clinical Research Unit (OUCRU), Hanoi.

- Langues de communication : français (MC), anglais (MC, LVT) et vietnamien (LVT).

Contexte

R est un langage interprété libre créé en 1993 sur la base du langage S créé en 1976. Initialement développé comme un langage dédié aux statistiques, il est récemment devenu un langage généraliste en science comme peut l'être Python. Le succès de R tient en grande partie à la puissance de son organisation en paquets développés par les utilisateurs pour accomplir des tâches spécifiques. De 40 paquets en 2000, le dépôt CRAN (cran.r-project.org) en compte aujourd'hui plus de 8000.

Pour quiconque étudie l'épidémiologie des maladies infectieuses, il est crucial de connaître le statut immunitaire des populations dans lesquelles les maladies sont transmises. Le statut immunitaire des populations peut-être estimé à partir de données d'enquêtes de séro-prévalence stratifiées par âge, localité et toute autre information sociale. De nombreux modèles (paramétriques et non-paramétriques) ont été développés pour faire de l'inférence sur les processus de transmission passés et présents à partir de telles données. Ces modèles ont été publiés mais il n'existe à ce jour aucun paquet R implémentant ces méthodes.

Travaux théoriques

- Faire l'état de l'art des différentes approches proposées pour faire de l'inférence sur la transmission à partir de données d'enquêtes de séro-prévalence ;
- Proposer un cadre général pour implémenter ces modèles de façon intégrée.

Travaux pratiques

- Développer un paquet R implémentant le cadre proposé ci-dessus. Le paquet sera développé pour partie en C++ pour en augmenter la rapidité de calcul et sera interfacé avec R grâce au paquet Rcpp.

Références:

1 Hens, N., M. Aerts, C. Faes, Z. Shkedy, O. Lejeune, P. van Damme, and P. Beutels. 2010. Seventy-five years of estimating the force of infection from current status data. Epidemiology and Infection 138:802–812.

2 Hens, N., Z. Shkedy, M. Aerts, C. Faes, P. van Damme and V. Beutels. 2012. Modeling Infectious Diseases Parameters Based on Serological and Social Contact Data, Springer-Verlag.

3 Wickham H. 2015. Advanced R, Chapman & Hall CRC.

4 Wickham H. 2015. R Packages, O'Reilly.

7. Sujet 7 : Un systèmes d'aide à la décision pour opération le système de bus dans la ville de Hanoi avec la méthode multi-agents

Encadrement: HO Tuong Vinh, NGUYEN-NGOC Doanh, Nguyen Manh Hung

Collaboration externe : UMMISCO Vietnam et Université Thuyloi (TLU)

Contexte

Le développement des réseaux de transport public en bus est un besoin urgent pour les grandes villes afin de répondre aux besoins de déplacement des gens, d'éviter les embouteillages, la pollution environnementale et d'assurer la sécurité routière. Selon les informations de la société de transport Hanoi (TRANSECO), à Hanoi, les lignes de bus de la société réalisent en moyen quotidiennement près de 10.000 déplacements. Dans des cas particuliers tels que les congestions, les accidents, les inondations, etc., l'opération du système de bus est souvent perturbée. Il nécessite donc un système de gestion permettant aux operateurs de bus de réagir rapidement et efficacement aux cas particulier cités précédents.

Ce sujet de TPE a pour but de concevoir un système d'aide à la décision, à base de systèmes multi-agents, pour mieux opérer le système de bus de la ville de Hanoi.

Travaux théoriques

- Étudier la simulation des réseaux de transport à base de systèmes multi-agents
- Proposer un modèle multi-agents pour simuler les systèmes de bus de la ville de Hanoi
- Créer différents scénarios pour tester l'utilité du système dans la gestion des lignes de bus dans des cas particuliers (congestion, accident, inondation dans certaines rues de la ville).

Travaux pratiques

- Implémenter le modèle avec l'outil GAMA
- Expérimenter les scénarios crées et proposer des améliorations nécessaires.

Références

[1] GAMA platform:
http://vps226121.ovh.net/references#PlatformDocumentation
[2] Trafic model with GAMA:
http://vps226121.ovh.net/references#TrafficSimpletrafficmodel
[3] Rouhieh, B. and Alecsandru, C. (2014). "Optimizing route choice in multimodal transportation networks." *Canadian Journal of Civil Engineering*, 10.1139/cjce-2013-0331, 800-810.
[4] Nair, R., Coffey, C., Pinelli, F., and Calabrese, F. (2013). "Large-Scale Transit Schedule Coordination Based on Journey Planner Requests." *Transportation Research Record: Journal of the Transportation Research Board*, 10.3141/2351-08, 65-75.
[5] Jeremy J. Blum and Tom V. Mathew. (2011) "Intelligent Agent Optimization of Urban Bus Transit System Design, Journal of Computing in Civil Engineering, 25(5), 357-369.

8. Sujet 8 : Étude et implémentation d'un solveur CSP incrémental pour éditeurs graphiques

Encadrement

- Frédéric Lardeux - *(LERIA) Université d'Angers*
- Frédéric Jouault, Fabien Chhel (porteur - fabien.chhel@eseo.fr) - *(TRAME) ESEO Angers*
- Ho Tuong Vinh, IFI

Problématique

Dans le cadre d'éditeurs graphiques, il existe de nombreuses implémentations *ad-hoc* d'algorithmes permettant le placement de composants graphiques respectant des contraintes. Par exemple, la bibliothèque GraphViz [http://www.graphviz.org/] propose des algorithmes de placement et de routage pour des graphes. Elle est notamment utilisée par [http://plantuml.com/] pour des diagrammes UML. Hors, une approche déclarative basée sur les problèmes de satisfaction de contraintes (CSP) permet de séparer l'aspect modélisation du problème géométrique de l'aspect résolution. Dès lors, le concepteur ne doit recueillir que les contraintes du problème et le solveur calcule une solution de routage et/ou de placement. L'aspect dynamique du placement de composants graphiques (e.g., quand l'utilisateur déplace des éléments à la souris) implique que le solveur CSP est appelé plusieurs fois sur des ensembles de contraintes très proches. Toutefois, il n'est pas nécessaire de recalculer à chaque fois l'ensemble des contraintes du problème mais seulement celles qui sont impactées par les changements et d'identifier les propagations. L'utilisation d'un CSP incrémental est donc tout à fait appropriée pour traiter ce genre de problèmes.

Positionnement

Si, historiquement, la programmation par contraintes prend son essence dans la résolution de problèmes de visualisation [Macworth1973, Montanari1974], les solveurs se sont ensuite plutôt concentrés sur des applications plus classiques de l'optimisation combinatoires et de l'intelligence artificielle (ordonnancement, planning, etc.). On peut noter que des travaux ont été menés dans le domaine des contraintes continues en lien avec des problèmes géométriques [Benhamou2004]. Très récemment, l'utilisation de la programmation par contraintes au sein de langages dédiés a été envisagée pour la gestion de graphiques interactifs [Hosobe2016].

La résolution des contraintes doit être envisagée dans un contexte particulier puisque qu'il s'agira de vérifier ces contraintes de manière incrémentale et réactive en fonction des interactions de l'utilisateur. Ce type de problème est identifié de longue date en programmation par contraintes [Wallace1996] et constitue naturellement l'une des propriétés essentielles des approches de type métaheuristiques [Loudni2003].

Le choix de SVG se justifie par le fait que se standard est facilement manipulable et visualisable par n'importe quel navigateur.

Travaux théoriques

- Étudier un solveur CSP incrémental permettant de résoudre un ensemble de contraintes graphiques dans le contexte de SVG.

Travaux pratiques

- Implémenter (une partie) un solveur CSP incrémental permettant de résoudre un ensemble de contraintes graphiques dans le contexte de SVG.

Références

[Hosobe2016] Hiroshi Hosobe, Toward a new constraint imperative programming language for interactive graphics. 15th International Conference on Modularity, 2016: 34-35
[Macworth1973] A.K. Macworth, Interpreting pictures of polyhedral scenes, Artificial Intelligence, Volume 4, Issue 2, 1973, Pages 121-137
[Montanari1974] Ugo Montanari: Networks of constraints: Fundamental properties and applications to picture processing. Inf. Sci. 7: 95-132, 1974.
[Benhamou2004] Frédéric Benhamou, Frédéric Goualard, Eric Languénou, Marc Christie, Interval constraint solving for camera control and motion planning. ACM Trans. Comput. Log. 5(4): 732-767 (2004)
[Condotta2006] Jean-François Condotta, Mahmoud Saade, Gerard Ligozat: A Generic Toolkit for n-ary Qualitative Temporal and Spatial Calculi. TIME 2006: 78-86
[Varro2007]Dániel Varró, et al. Transformation of UML Models to CSP: A Case Study for Graph Transformation Tools. Applications of Graph Transformations with Industrial Relevance 2007: 540-565.
[Wallace1996]Richard J. Wallace, Eugene C. Freuder: Anytime Algorithms for Constraint Satisfaction and SAT Problems. SIGART Bulletin 7(2): 7-10 (1996)
[Loudni2003] Samir Loudni, Patrice Boizumault: Solving Constraint Optimization Problems in Anytime Contexts. IJCAI 2003: 251-256

9. Sujet 9 : SIMULATION POUR L'ORGANISATION DES MOYENS DE SECOURS DANS LE CAS DE TREMBLEMENT DE TERRE EN ZONE URBAINE

Encadrement : HO Tuong Vinh, Nguyen Manh Hung
Collaboration externe : NGUYEN Hong Phuong (IGP, VAST)

Contexte

Dans le cadre d'une collaboration entre l'équipe MSI et la l'institut de Géophysique de la VAST plusieurs travaux se sont développés autour du modèle et de la simulation pour la gestion de secours en cas de tremblement de terre en zone urbain. Grâce au simulateur ArcRisk, développé par l'IGP, avec les données SIG nous avons pu construire des scenarios de tremblement de terre et proposer des outils pour estimer les dommages faits aux infrastructures et les blessés et victimes de ces scénarios. En utilisant ces informations, nous avons pu développer un modèle à base de système multi-agents permettant simuler l'organisation des actions de secours. Des travaux préliminaires ont déjà été

conduits permettant expérimenter quelques stratégies de secours, mais beaucoup reste à faire pour que l'on puisse utiliser les résultats des simulations de manière effective.
L'objectif de ce TPE est d'évaluer le modèle existant et proposer des améliorations pour le rendre plus efficace.

Travaux théoriques
- Faire un survol sur les modèles de gestion de secours au cas de tremblement de terre en zone urbain
- Évaluer le modèle existant et proposer des améliorations
- Proposer quelques scénarios pour évaluer le modèle amélioré

Travaux pratiques
- Implémenter les améliorations proposées
- Évaluer le nouveau modèle en expérimentant quelques scénarios de gestion de secours

Références

1 Thanh-Quang Chu, Alexis Drogoul, Alain Boucher & Jean-Daniel Zucker. Interactive Learning of Independent Experts' Criteria for Rescue Simulations. Journal of Universal Computer Science, 15(13), pp. 2701-2725, 2009
2 Le Xuan Sang, ArcRisk2GAMA - Interfacer le simulateur de tremblements de terre ArcRisk et la plateforme GAMA, Rapport de TPE, IFI, 2012

10. Sujet 10 : CSCL – Apprentissage collaboratif assisté par ordinateur

Encadrement:
- NGUYEN Trong Khanh, Institut Poste-Telecom
- HO Tuong Vinh, IFI

Contexte
L'apprentissage collaboratif à distance dans des environnements médiatisés, plus connu sous l'abréviation anglo-saxonne CSCL (Computer Supported Collaborative Learning), est un domaine de recherche et d'application relativement récent. Le premier atelier dans ce domaine date de 1991 et la première conférence internationale s'est tenue en 1995 à Bloomington (Indiana).

Dans ce domaine, il s'agit justement d'étudier la manière par laquelle l'apprentissage collaboratif dans des environnements médiatisés peut faciliter les interactions entre apprenants et le travail en groupe. Il s'agit également de déterminer comment la collaboration et la technologie facilitent l'expression, le partage et l'échange d'informations, de connaissances et de compétences entre les membres d'une communauté. Le CSCL s'inspire des recherches sur le CSCW (Computer-Supported Cooperative Work) qui ont insisté sur la nature collaborative du travail assisté par un collecticiel (ou groupware).

Travaux théoriques

- Faire un état de l'art des travaux dans le domaine du l'apprentissage collaboratif assisté par ordinateur et classifier les plateformes, les outils en fonction des plateforms utilisés.
- Proposer un meta-model de l'outil permettant du codage collabratif pour servir des courses du program.

Travaux pratiques

- Implementer le model proposé pour une plateform à base de Web.

Références

1 Dutta, M., Sethi, K.K., Khatri, A., 2014. Web Based Integrated Development Environment. Int. J. Innov. Technol. Explor. Eng. 3, 56– 60.

2 Katerina Zourou, « *Computer Supported Collaborative Learning* (*CSCL*) et apprentissage des langues assisté par ordinateur (Alao) : un dialogue à ne pas manquer - Réflexions autour du colloque mondial CSCL 09 », *Alsic* [Online], Vol. 12 | 2009

3 Bravo, C., Marcelino, M.J., Gomes, A., Esteves, M., Mendes, A.J., 2005. Integrating Educational Tools for Collaborative. J. Univers. Comput. Sci. 11, 1505–1517.

4 Johann W. Sarmiento-Klapper, 2009. Sustaining Collaborative Knowledge Building: Continuity in Virtual Math Teams. Drexel Univ.

5 Ludvigsen, S.R., Mørch, A., 2010. Computer-Supported Collaborative Learning: Basic Concepts, Multiple Perspectives, and Emerging Trends. Int. Encycl. Educ. 3rd Ed. 290–296.

6 Popescu, E., 2014. Providing collaborative learning support with social media in an integrated environment. World Wide Web 17, 199–212. doi:10.1007/s11280-012-0172-6

7 Stahl, G., Koschmann, T., Suthers, D., 2006. Computer-supported collaborative learning: An historical perspective. Cambridge Handb. Learn. Sci. 409–426. doi:10.1145/1124772.1124855

8 Teague, D., Roe, P., 2008. Collaborative learning - towards a solution for novice programmers. Conf. Res. Pract. Inf. Technol. Ser. 78, 147– 153.

9 Tran, H.T., Dang, H.H., Do, K.N., Tran, T.D., Nguyen, V., 2013. An interactive Web-based IDE towards teaching and learning in programming courses. Proc. 2013 IEEE Int. Conf. Teaching, Assess. Learn. Eng. TALE 2013 439–444. doi:10.1109/TALE.2013.6654478

10 Zhang, Z., Sun, Y., Lu, Y., 2013. Proceedings of the 2012 International Conference on Information Technology and Software Engineering. Lect. Notes Electr. Eng. 212, 783–789. doi:10.1007/978-3-642-34531-9

11. SUJET 11 : Outil d'aide a la division de l'espace de simulation sur GAMA

Encadrement: Nguyen Hong Quang
Collaboration externe : Nguyen Manh Hung (PTIT), Alexis Drogul (IRD)

Contexte
GAMA (GIS Agent-based Modeling Architecture) [1] est un environnement de développement de modélisation et de simulation pour la construction des simulations à base d'agents spatialement explicites. GAMA a été développé par plusieurs équipes dans le cadre de l'unité de recherche internationale IRD / UPMC UMMISCO et a été utilisé dans plusieurs projets de recherche d'UMMISCO [2].
L'architecture de GAMA consiste en une combinaison des projets Eclipse fonctionnant sur un même PC. Par conséquent, le système a des difficultés face aux grandes simulations constituant des centaines milles, voire des millions d'agents. Des efforts de paralléliser la plate-forme ont été faits [3]. Cependant, les résultats obtenus sont encore assez modestes.
Nous proposons une autre approche qui consiste a paralléliser la simulation plutôt que la plate-forme. L'idée est de partitionner l'espace de simulation en régions géographiques séparées puis assigner chaque régions (y compris ses agents) a une machine GAMA pour l'exécution. Ainsi, chaque machine GAMA ne s'occupe qu'une partie de la simulation globale.
L'environnement de travail des agents GAMA est représenté par le Système d'Information Géographique (SIG ou GIS en anglais pour *Geographic Information System*) [4]. Ce TPE vise a créer un outil permettant de diviser l'environnement d'une simulation en régions séparées. On s'intéresse aussi a la proposition d'un langage permettant de designer la frontière des régions pour but d'automatiser la division.

Travaux théoriques
- Étudier le fondement des systèmes GIS et leur rôle dans les applications
- Étudier la plate-forme GAMA et l'utilisation du GIS dans les simulations
- Étudier le mécanisme de déterminer la position des agents GAMA

Travaux pratiques
- Proposer une méthode de diviser l'environnement d'une simulation en régions a la demande.
- Implémenter un outil d'aide a la division selon la méthode proposée.

Références
[1] https://github.com/gama-platform/gama/wiki
[2] https://github.com/gama-platform/gama/wiki/Projects
[3] https://github.com/gama-platform/gama/issues/738
[4] https://fr.wikipedia.org/wiki/Syst%C3%A8me_d%27information_g%C3%A9ographique

12. SUJET 12 : Étude du modèle de communication entre des simulations GAMA

Encadrement: Nguyen Hong Quang
Collaboration externe : Nguyen Manh Hung (PTIT), Alexis Drogul (IRD)

Contexte

GAMA (GIS Agent-based Modeling Architecture) [1] est un environnement de développement de modélisation et de simulation pour la construction des simulations à base d'agents spatialement explicites. GAMA a été développé par plusieurs équipes dans le cadre de l'unité de recherche internationale IRD / UPMC UMMISCO et a été utilisé dans plusieurs projets de recherche d'UMMISCO [2].

L'architecture de GAMA consiste en une combinaison des projets Eclipse fonctionnant sur un même PC. Par conséquent, le système a des difficultés face aux grandes simulations constituant des centaines milles, voire des millions d'agents. Des efforts de paralléliser la plate-forme ont été faits [3]. Cependant, les résultats obtenus sont encore assez modestes.

Nous proposons une autre approche qui consiste a paralléliser la simulation plutôt que la plate-forme. L'idée est de partitionner l'espace de simulation en régions géographiques séparées puis assigner chaque régions (y compris ses agents) a une machine GAMA pour l'exécution. Ainsi, chaque machine GAMA ne s'occupe qu'une partie de la simulation globale.

Un des problèmes essentiels a résoudre pour cette approche est comment les agents dans la zone frontière des deux régions voisines peuvent recevoir les informations l'une de l'autre . Ce TPE vise a trouver le modèle de communication entre deux simulations séparées mais connexes étroitement. Ainsi, réaliser l'implémentation d'un prototype permettant de valider le modèle dans quelques cas.

Travaux théoriques

- Étudier le fondement des systèmes multi-agents [4] et leurs modèles de communication entre agents
- Étudier la plate-forme GAMA et son modèle de communication entre ses agents dans les simulations
- Étudier le scope (distance) de communication des agents GAMA dans son environnement géographique sous-jacent.

Travaux pratiques

- Proposer une méthode de déterminer la zone « visible » d'un agent de l'environnement d'une simulation connexe.
- Proposer un modèle de communication pour « mettre au courant » sur les changements survenus dans la zone « visible » des deux simulations connexes
- Implémenter un prototype pour valider le modèle proposé.

Références

[1] https://github.com/gama-platform/gama/wiki
[2] https://github.com/gama-platform/gama/wiki/Projects
[3] https://github.com/gama-platform/gama/issues/738
[4] https://fr.wikipedia.org/wiki/Syst%C3%A8me_multi-agents

13. SUJET 13 : Solution d'affichage des simulations connexes sur GAMA

Encadrement: Nguyen Hong Quang
Collaboration externe : Nguyen Manh Hung (PTIT), Alexis Drogul (IRD)

Contexte
GAMA (GIS Agent-based Modeling Architecture) [1] est un environnement de développement de modélisation et de simulation pour la construction des simulations à base d'agents spatialement explicites. GAMA a été développé par plusieurs équipes dans le cadre de l'unité de recherche internationale IRD / UPMC UMMISCO et a été utilisé dans plusieurs projets de recherche d'UMMISCO [2].
L'architecture de GAMA consiste en une combinaison des projets Eclipse fonctionnant sur un même PC. Par conséquent, le système a des difficultés face aux grandes simulations constituant des centaines milles, voire des millions d'agents. Des efforts de paralléliser la plate-forme ont été faits [3]. Cependant, les résultats obtenus sont encore assez modestes.
Nous proposons une autre approche qui consiste a paralléliser la simulation plutôt que la plate-forme. L'idée est de partitionner l'espace de simulation en régions géographiques séparées puis assigner chaque régions (y compris ses agents) a une machine GAMA pour l'exécution. Ainsi, chaque machine GAMA ne s'occupe qu'une partie de la simulation globale.
On souhaite bien entendu de visualiser le résultat de toutes ces simulations dites connexes comme un entier. Ce TPE vise a créer un outil permettant de visualiser le résultat de toutes les simulation connexes comme dans une seule simulation. Cet outil jouera en même temps le rôle du « chef d'orchestre » de toutes ces simulation connexes.
Travaux théoriques
- Étudier la plate-forme GAMA et les mécanismes d'affichage des simulations
- Étudier la relation d'affichage entre l'environnement GIS [4] et l'état des agents dans GAMA.

Travaux pratiques
- Proposer une méthode de placer le résultat des simulations connexes sur un espace d'affichage.
- Implémenter un prototype permettant d'afficher des simulations connexes simples.

Références
[1] https://github.com/gama-platform/gama/wiki
[2] https://github.com/gama-platform/gama/wiki/Projects
[3] https://github.com/gama-platform/gama/issues/738
[4] https://fr.wikipedia.org/wiki/Syst%C3%A8me_d%27information_g%C3%A9ographique

14. SUJET 14 : Diffusion d'opinions dans les réseaux sociaux : gestion de la discrimination

Encadrement

- Dominique LONGIN (IRIT, Toulouse, France)
- HO Tuong Vinh (IFI, Hanoi)

Contexte

Depuis quelques années, les réseaux sociaux sont très étudiés car ils sont générateurs d'un grand nombre d'interactions entre les personnes, et celles-ci ont pour effet une diffusion rapide des opinions, des idées, des goûts, des modes, *etc.*

Une question qui commence à être posée, c'est comment les individus se comportent lorsqu'ils sont en présence de personnes indésirables. Le problème peut se modéliser de la façon suivante : on suppose que les agents sont partagés en deux « équipes ». Lorsque le nombre de personnes qui nous entourent et appartenant à l'équipe adverse devient trop important, nous fuyons ce groupe d'individus en partant dans la direction opposée. Ainsi, nous sommes soumis à un autre entourage d'individus qui lui-même a pu bouger et qui est susceptible de contenir à la fois des personnes qui vont de nouveau nous pousser à fuir, ou des personnes qui vont elles-mêmes fuir à notre approche.

On se propose de modéliser mathématiquement ce mécanisme de diffusion puis de l'implémenter.

Travaux théoriques

Ils concernent différents aspects :

- recherches bibliographiques (dont l'article mentionné ci-dessous peut être un point d'entrée pour trouver d'autres articles sur le sujet) dans le domaine de la diffusion d'opinion ;
- élaborer un modèle mathématique minimal pour modéliser ce problème. On se donnera toutes les fonctions élémentaires (ne pouvant être obtenues par composition des autres fonctions utilisées) nécessaires. Par exemple, la fonction $Pos : AGT \to \mathbb{N} \times \mathbb{N}$ retourne la position d'un agent sous forme de couple de coordonnées, la contrainte « on ne peut aller que dans une direction donnée », *etc.*
- formalisation de l'opinion d'un agent et du mécanisme de mise à jour en passant de l'instant t à l'instant $t+1$.
- On pourra par la suite réfléchir à d'autres modèles d'influence (voir la littérature à ce sujet).

Travaux pratiques

- implémenter un modèle sous JAVA en faisant varier différents paramètres (nombre d'agents dans chaque équipe, répartition géographique initiale, seuil à partir duquel un agent fuit, *etc.*) et analyser les résultats. Comment le système se stabilise ?
- implémenter le modèle précédent dans l'architecture GAMA avec visualisation spatiale des agents et passage d'un ensemble discret de directions à un ensemble continu.

Références

U. Grandi, E. Lorini, L. Perrussel (2016). « Propositional opinion diffusion ». *In Proceedings of international conference on Autonomous Agents and Multi-Agent Systems (AAMAS-2016).*

15. SUJET 15 : Un progiciel de R basé sur la recherche locale pour les problèmes de groupement avec plus contraintes

(A R pakage based on local search for clutering problems with additional constraints)

Encadrement

- BUI Quoc Trung (Institut de recherche FPT)
- HO Tuong Vinh (IFI, Hanoi)

Contexte

L'objective du problème de groupement général est d'assigner chaque élément à partir d'un ensemble à sous-ensemble tel que les éléments dans même sous-ensemble sont similaires autant que possible et les éléments dans les différents sous-ensemble sont très différents autant que possible. Le nombre de sous-ensembles dans quelques situations est disponible, dans d'autres situations il est inconnu. Ce sont considérés comme les problèmes les plus importants dans les problèmes d'apprentissage non-supervisés. Cependant, des problèmes de groupement avec plus contraintes considèrent plus de contraintes en comparaison avec les problèmes de groupement normales, par exemple, la contrainte de must-link (deux éléments doivent être dans même un sous-ensemble), la contrainte de balance (les sous-ensembles doivent avoir presque une même taille), …

Maintenant il existe beaucoup d'algorithmes pour tous les problèmes de groupement généraux et les problèmes de groupement avec plus contraintes. Presque tous les algorithmes pour les problèmes de groupement généraux sont disponibles dans des logiciels gratuits dans le domaine de fouille de données, par exemple R, Weka, Orange,... Toutefois, les algorithmes pour les problèmes de groupement avec plus contraintes sont rarement intégrés dans ces logiciels.

La recherche local est très efficace pour les problèmes réels de grandes échelles, sur tous les problèmes avec des contraintes compliquées. Elle peut donner de bonnes solutions dans un mottant de temps raison.

Travaux théoriques

- Faire un état de l'art des algorithmes pour les problèmes de groupement avec plus contraintes
- Apprendre la recherche locale générale, la recherche Tabu, la recherche locale it érée
- Proposér deux algorithmes de recherche locale pour deux problèmes de groupement avec plus contraintes

Travaux pratiques

- Reimplémenter avec le langage R au moins quatre algorithmes existantes pour deux problèmes de groupement avec plus contraintes
- Implémenter les algorithmes proposés avec le langage R

Réferences

1. Wagstaff, K., & Cardie, C. (2000). Clustering with Instance-level Constraints
2. Basu, S., Davidson, I., & Wagstaff, K. (2008). Constrained Clustering Advances in Algorithms, Theory, and Applications.
3. Davidson, I., & Ravi. (2005). Clustering With Constraints: Feasibility Issues and the k-Means Algorithm.
4. Kulis, B., Basu, S., Dhillon, I., & Mooney, R. (2005). Semi-supervised Graph Clustering: A Kernel Approach.
5. Pelleg, D., & Baras, D. (2007). K-means with Large and Noisy Constraint Set.
6. Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge.
7. Yuan, T. (n.d.). Retrieved from https://cran.r-project.org/web/packages/dml/

16. SUJET 16 : Une nouveau algorithme de metaheuristic pour le problème de Covering Salesman

(A NEW METAHEURISTIC FOR THE COVERING SALESMAN PROBLEM - CSP)

Encadrement

- BUI Quoc Trung (Institut de recherche FPT)
- HO Tuong Vinh (IFI, Hanoi)

Context

Récemment, Thibaut Vidal (https://w1.cirrelt.ca/~vidalt/en/publications-thibaut-vidal.html) a proposé un original algorithme qui domine tous les algorithmes existants pour le problème de routage du véhicule et ses variantes. Cet algorithme a capable d'adapter pour les problèmes de routage. Cependant, le problème CSP est très connu dans les problèmes de routage et il n'existe pas encore une recherche qui essaie d'adapter l'algorithme de Thibaut pour ce problème.

Travaux théoriques

- Comprendre bien l'algorithme de Thibaut
- Faire un état de l'art des algorithmes pour le problème CSP
- Adapter l'algorithme de Thibaut pour le problème CSP

Travaux practiques

- Implémenter l'algorithme proposé

Références

1 Kramer, R., Subramanian, A., Vidal, T., Cabral, L.A.F. **(2015)**. A matheuristic approach for the Pollution-Routing Problem. *European Journal of Operational Research* , 243(2), 523-539 WP : DOI : 10.1016/j.ejor.2014.12.009

2 Vidal, T., Battarra M., Subramanian A., Erdo?an, G. **(2015)**. Hybrid metaheuristics for the Clustered Vehicle Routing Problem. *Computers & Operations Research* , 58(1), 87-99 WP : DOI : 10.1016/j.cor.2014.10.019

17. SUJET 17: analyse de composition de séquence

Encadrement : Ho Bich Hai, Edi Prifti, Ho Tuong Vinh
Collaboration externe : ICAN (INSERM/UPMC) and UMMISCO (IRD), France

Contexte

La séquence, telle que l'ADN, l'ARN ou une protéine, est l'un des types de données primaires dans la biologie computationnelle. Prenons une séquence d'ADN comme exemple: ATCGGATTAAC. Elle est une liste séquentielle de 4 nucléotides possibles (A, C, T, G), au format texte. Composition d'une séquence, à savoir le motif et la fréquence, peut être informative sur son origine, la fonctionnalité, etc. Ainsi, de nombreuses analyses commencent par les études sur les caractéristiques de composition de séquence. Dans le contexte de la génomique / métagénomique, nous nous concentrons sur les séquences d'ADN de bactéries de l'environnement. Autrement dit, une échantillon est en fait un mélange de séquences à partir de nombreuses espèces bactériennes, comme une soupe.
Regroupement taxonomique (le regroupement de ces séquences dans des espèces provisoirs est fondé sur l'hypothèse selon laquelle les séquences d'une bactérie particulière sont différentes de celles de l'autre en terme de composition. Par conséquent, les caractéristiques de composition sont extraites et utilisées comme entrées pour un algorithme de groupement (clustering) pour réaliser cette tâche. Les caractéristiques peuvent être très différents, par exemple le pourcentage de GC, l'utilisation des codons, la fréquence du groupe de k nucléotides (kmer), etc. L'objectif de ce sujet TPE est d'explorer les caractéristiques de composition récents et tester leur efficacité en regroupement taxonomique des séquences bactériennes.

Travaux théoriques

- Étudier les connaissances de base en bio-informatique (format de séquence et pré-traitement)
- Faire un survol sur les caractéristiques de composition et statistique de séquence [1]-[3]
- Étudier le regourpement taxonomique par caractéristiques de composition de séquence [4]-[7].

Travaux pratiques

- Apprendre tutoriels sur DNA sequence preprocessing, sequence composition feature extraction by Bioconductor (in R, https://www.bioconductor.org/), PyCogent [8] (in Python) packages, Seqool (http://www.biossc.de/seqool/) and principal component analysis (PCA) etc.
- Analyse de regoupement taxonomique en utilisant ces caractéristiques: tester les caractéristiques (et sous ensembles des caractéristiques), expérimenter quelques algorithmes de clustering, et évaluer avec les données
- Implémentation d'un workflow (avec les étapes précédentes) et analyse avec des données réelles (s'il le temps nous permet).

Exgigence:

- Des connaissances avec Unix (HPC Unix-based system).
- Connaitre un des langages de programmation: Python, Perl, R, and C++.

Références

(Please contact hobichhai@gmail.com if you need these papers' files)

[1] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The European Molecular Biology Open Software Suite," Trends in Genetics, vol. 6, no. 16, pp. 276–277, May 2000.

[2] S. Sinha, "YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation," Nucleic Acids Research, vol. 31, no. 13, pp. 3586–3588, Jul. 2003.

[3] T. Lassmann, Y. Hayashizaki, and C. O. Daub, "SAMStat: monitoring biases in next generation sequencing data.," Bioinformatics, vol. 27, no. 1, pp. 130–131, Jan. 2011.

[4] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments.," Genome Biology, vol. 15, no. 3, p. R46, 2014.

[5] M. Imelfort, D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson, "GroopM: an automated tool for the recovery of population genomes from related metagenomes," PeerJ, vol. 2, pp. e603–16, 2014.

[6] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities," PeerJ, vol. 3, no. 8, pp. e1165–15, 2015.

[7] Y.-W. Wu, Y.-H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer, "MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.," Microbiome, vol. 2, no. 1, p. 26, 2014.

[8] R. Knight, P. Maxwell, A. Birmingham, J. Carnes, J. G. Caporaso, B. C. Easton, M. Eaton, M. Hamady, H. Lindsay, Z. Liu, C. Lozupone, D. McDonald, M. Robeson, R. Sammut, S. Smit, M. J. Wakefield, J. Widmann, S. Wikman, S. Wilson, H. Ying, and G. A. Huttley, "PyCogent: a toolkit for making sense from sequence," Genome Biology, vol. 8, no. 171, pp. 1–16, Oct. 2007.

18. SUJET 18 : analyse de diversité communautaire microbienne à base d'AMPLICON
(Amplicon-based microbial community diversity analysis)

Encadrement : Ho Bich Hai, Jean-Daniel Zucker, and Ho Tuong Vinh
Collaboration externe : ICAN (INSERM/UPMC) and UMMISCO (IRD), France

Contexte
La communauté microbienne récupérée directement à partir de l'environnement est d'intérêt pour ses diverses applications, par exemple pour la santé et l'agriculture. Pour comprendre cette communauté, l'une des premières questions à répondre est «qui», à savoir l'identification des présentes espèces microbiennes et leurs abondances. Cela peut donner un aperçu de la diversité d'une communauté et peut être utilisé pour comparer les communautés dans différentes conditions. D'une manière générale, les matériels génétiques de la communauté sont extraites et séquencés dans le laboratoire humide.
Nous allons analyser les séquences d'ADN, sous forme textuelle , contenant 4 caractères A, C , T , G , et éventuellement jusqu'à quelques centaines de caractères. Une approche, entre autres, est de regarder les gènes spéciaux qui contiennent des régions hypervariables , et qui peuvent ainsi différencier les espèces bactériennes . Pour les bactéries, les gènes de l'ARNr 16S sont généralement choisis pour cet objectif.Ceux-ci peuvent être obtenus par séquençage comme amplicon, d'où l'approche à base d'amplicon . En regroupant les gènes ARNr 16S trouvés dans une communauté, nous pouvons dire qui est là. En ce qui concerne la méthodologie, cette tâche peut se faire de trois façons: à bas de référence, de novo, ou l'hybride de ceux-ci. Autrement dit, ils sont basés sur une mesure de similarité entre les séquences ou avec référence (par alignement) et les techniques d'apprentissage de cluster / de classification. Les résultats sont des groupes de séquences, que l'on appelle l'unité taxonomique opérationnelle (OTU). Les indices de diversité à base de OTU caractérisent une communauté microbienne. L'objectif de ce sujet est de comprendre les méthodes récentes et acquérir de l'expérience par l'analyse des données séquencées.

Travaux théoriques
- Comprendre les connaissance de base de la bioinformatique (le contexte et les données): analyse de la communauté microbienne et traitement des données séquencées de l'amplicon
- Étudier les méthodes d'analyse OTU des données séquencées de l'amplicon [1] - [4] .
- Étudier l'analyse comparative sur les communautés microbiennes en termes d'indices de diversité [5] , [6].

Travaux pratiques
- Apprendre des tutoriels sur le prétraitement de séquence d'ADN , analyse OTU, et l'analyse comparative [7] - [9] .
- Participer à l'analyse d'un ensemble de données réelles et publication des résultats.

Exigence
- Être familier avec l'environnement Unix (toutes les analyses se feront dans les systèmes à base de Unix HPC) .

- Etre capable de programmer avec un des langages: Python , Perl , R et C ++ .

Références

(Please contact hobichhai@gmail.com if you need these papers' files)
[1] J. R. Rideout, Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell, S. M. Gibbons, J. Chase, D. McDonald, A. Gonzalez, A. Robbins-Pianka, J. C. Clemente, J. A. Gilbert, S. M. Huse, H.-W. Zhou, R. Knight, and J. G. Caporaso, "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences," PeerJ, vol. 2, no. 5, p. e545, 2014.
[2] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.," Appl. Environ. Microbiol., vol. 72, no. 7, pp. 5069–5072, Jul. 2006.
[3] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glockner, "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools," Nucleic Acids Research, vol. 41, no. 1, pp. D590–D596, Dec. 2012.
[4] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.," Appl. Environ. Microbiol., vol. 73, no. 16, pp. 5261–5267, Aug. 2007.
[5] D. Guttman, A. C. McHardy, and P. Schulze-Lefert, "Microbial genome-enabled insights into plant–microorganism interactions," Nature Publishing Group, vol. 15, no. 12, pp. 797–813, Sep. 2014.
[6] D. S. Lundberg, S. L. Lebeis, S. H. Paredes, S. Yourstone, J. Gehring, S. Malfatti, J. Tremblay, A. Engelbrektson, V. Kunin, T. G. D. Rio, R. C. Edgar, T. Eickhorst, R. E. Ley, P. Hugenholtz, S. G. Tringe, and J. L. Dangl, "Defining the core Arabidopsis thaliana root microbiome," Nature, vol. 488, no. 7409, pp. 86–90, Aug. 2012.
[7] B. CALLAHAN, D. PROCTOR, D. RELMAN, J. FUKUYAMA, and S. Holmes, "Reproducible research workflow in r for the analysis of personalized human microbiome data ," presented at the Proceedings of Pacific Symposium on Biocomputing, 2016, no. 21, pp. 183–194.
[8] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber, "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities," Appl. Environ. Microbiol., vol. 75, no. 23, pp. 7537–7541, Nov. 2009.
[9] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight, "QIIME allows analysis of high-throughput community sequencing data.," Nature Publishing Group, vol. 7, no. 5, pp. 335–336, May 2010.

19. SUJET 19 : ANNOTATION D'IMAGE SEMI-AUTOMATIQUE : APPLICATION AU projet ARCHIVES de l'usth

(Semi-automatic image annotation: application to usth archives project)

Encadrement

- NGHIEM Thi Phuong, TRAN Giang Son (USTH)
- HO Tuong Vinh (IFI, Hanoi)

Contexte:

Ce sujet se concentre sur l'annotation d'images où chaque image est étiqueté avec un ensemble d'étiquettes pertinentes. Traditionnellement, l'annotation est effectuée manuellement par l'homme où les résultats soulèvent certaines questions liées aux coûts humains et de la subjectivité.
Récemment, les chercheurs introduisent l'annotation automatique où les techniques d'apprentissage automatique sont utilisés pour construire automatiquement des modèles d'annotation. Bien que , les résultats d'annotation sont généralement bonnes par rapport à l'annotation manuelle, l'annotation automatique nécessite énorme quantité de données d'image étiquéteés pour l'apprentisage.
Pour cette raison, l'annotation automatique devient insuffisante pour les cas où la quantité de annotées données disponibles est limité. Ce problème est particulièrement vrai dans le cas du projet ARCHIVES de l'USTH où les documents historiques annotées des catastrophes naturelles passées à Hanoi est très limitée en nombre.
Le but de ce sujet est d'apprendre un modèle structuré à partir d'un ensemble limité d'images annotées manuellement puis chercher l'aide des utilisateurs finaux (ici les historiens) afin d'annoter de nouvelles images à venir. Deux étapes sont envisagées. Tout d'abord , certaines images étiquetées automatiques sont présentés aux utilisateurs finaux pour leur retour de pertinence. Ensuite, la réponse de l'utilisateur sur l'exactitude d'une étiquette donnée est utilisée pour introduire dans le modèle structuré pour corriger les scores de pertinence des autres étiquettes.

Travaux théoriques:

- Étudier les connaissances de base de l'apprentissage automatique
- Étudier les connaissances de base de l'annotation d'images

Travaux pratiques:

- Implementer le modèle d'annotation décrit dans Thomas Mensik's [1]
- Améliorer le modèle de Thomas
- Expérimenter le modèle avec les données du projet ARCHIVES

References:

[1] https://staff.fnwi.uva.nl/t.e.j.mensink/publications/mensink12phd.pdf

ANNEXE : SUJETS EN ANGLAIS

SUJET 17: ANALYSIS OF SEQUENCE COMPOSTION

Encadrement : Ho Bich Hai, Edi Prifti, Ho Tuong Vinh
Collaboration externe : ICAN (INSERM/UPMC) and UMMISCO (IRD), France

Contexte
Sequence, such as DNA, RNA or protein, is one of the primary data types in computational biology. Let's take a DNA sequence as an example: ATCGGATTAAC. It is a sequential list of 4 possible nucleotides (A,C,T,G), in text format. Composition of a sequence, i.e. the pattern and frequency, can be informative about its origin, functionality etc. Thus, many analyses begin with investigating sequence composition features. In the context of genomics/metagenomics, we focus on DNA sequences of bacteria from the environment. That is, a sample is actually a mixture of sequences from many bacterial species, like a soup. Taxonomic binning, grouping those sequences into tentative species, is based on the hypothesis that sequences of a specific bacterium is different from those of another one in term of composition. Hence, composition features are extracted and used as input for a clustering algorithm to achieve this task. Features can be quite various, for example percentage of GC, codon usage, frequency of group of *k* nucleotides (kmer), etc. The objective of this subject TPE is to explore the state-of-the-art composition features and test their efficiency in taxonomic binning of bacterial sequences.

Travaux théoriques

- Understand basic bioinformatics knowledge about the context and data: sequence format and preprocessing.
- Review the state-of-the-art composition features and sequence statistics [1]-[3]
- Review the related work taxonomic binning using sequence composition features [4]-[7].

Travaux pratiques

- Follow tutorials on DNA sequence preprocessing, sequence composition feature extraction by Bioconductor (in R, https://www.bioconductor.org/), PyCogent [8] (in Python) packages, Seqool (http://www.biossc.de/seqool/) and principal component analysis (PCA) etc.
- Taxonomic binning analysis using those features: testing features (and subsets of features), using a number of clustering algorithms, and evaluating on simulated data.
- Implementation of a workflow (containing the above steps) and analysis of a real dataset (if time allows).

Requirements

- Being motivated and ready to learn some domain knowledge (a few biology concepts and bioinformatics practices).
- Being familiar with Unix environment (all analysis will be done in HPC Unix-based system).

- Programming skill in at least one of Python, Perl, R, and C++.

Miscellaneous

- Working with the supervisors and an IFI Master 2 intern.
- Chance to participate in exchange program abroad or continue the topic as Master 2 internship.

Références

(Please contact hobichhai@gmail.com if you need these papers' files)

[1] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The European Molecular Biology Open Software Suite," *Trends in Genetics*, vol. 6, no. 16, pp. 276–277, May 2000.

[2] S. Sinha, "YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3586–3588, Jul. 2003.

[3] T. Lassmann, Y. Hayashizaki, and C. O. Daub, "SAMStat: monitoring biases in next generation sequencing data.," *Bioinformatics*, vol. 27, no. 1, pp. 130–131, Jan. 2011.

[4] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments.," *Genome Biology*, vol. 15, no. 3, p. R46, 2014.

[5] M. Imelfort, D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson, "GroopM: an automated tool for the recovery of population genomes from related metagenomes," *PeerJ*, vol. 2, pp. e603–16, 2014.

[6] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities," *PeerJ*, vol. 3, no. 8, pp. e1165–15, 2015.

[7] Y.-W. Wu, Y.-H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer, "MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.," *Microbiome*, vol. 2, no. 1, p. 26, 2014.

[8] R. Knight, P. Maxwell, A. Birmingham, J. Carnes, J. G. Caporaso, B. C. Easton, M. Eaton, M. Hamady, H. Lindsay, Z. Liu, C. Lozupone, D. McDonald, M. Robeson, R. Sammut, S. Smit, M. J. Wakefield, J. Widmann, S. Wikman, S. Wilson, H. Ying, and G. A. Huttley, "PyCogent: a toolkit for making sense from sequence," *Genome Biology*, vol. 8, no. 171, pp. 1–16, Oct. 2007.

SUJET 18 : AMPLICON-BASED MICROBIAL COMMUNITY DIVERSITY ANALYSIS

Encadrement : Ho Bich Hai, Jean-Daniel Zucker, and Ho Tuong Vinh
Collaboration externe : ICAN (INSERM/UPMC) and UMMISCO (IRD), France

Contexte

Microbial community retrieved directly from environment is of interest for its various applications, e.g. health and agriculture. To understand such community, one of the first questions to answer is "who", i.e. identifying the present microbial species and their abundances. This can give insights to the diversity of a community and can be used to compare communities from different conditions. Generally speaking, the genetic materials of the community are

extracted and sequenced in wet laboratory. We will analysis the resulting DNA sequences, i.e. in text format, containing 4 characters A, C, T, G, and possibly of up to a few hundreds character length. One approach, among others, is to look at special genes that contain hypervariable regions, thus can differentiate bacterial species. For bacteria, 16S rRNA genes are widely chosen for this purpose. These can be derived by as amplicon sequencing, hence the amplicon-based approach. By grouping the 16S rRNA genes found in a community, we can tell who is there. Regarding methodology, this task can be done in three ways: reference-based, *de novo*, or the hybrid of these. Simply put, they are based on a similarity measure among sequences or with reference (by alignment) and clustering/classification machine learning techniques. The results are groups of sequences, so-called Operational Taxonomic Unit (OTUs). The OTUs and OTU-based diversity indexes characterize a microbial community. The objective of this internship is to understand the state-of-the-art methods and gain hand-on experience by analysis sequencing data.

Travaux théoriques

- Understand basic bioinformatics knowledge about the context and data: microbial community analysis and amplicon sequencing data processing.
- Review the state-of-the-art methods in OTU analysis of amplicon sequencing data [1]-[4] .
- Review the comparative analysis on microbial communities in terms of diversity indexes [5], [6].

Travaux pratiques

- Follow tutorials on DNA sequence preprocessing, OTU analysis, and comparative analysis [7]-[9].
- Participate in analyzing a real in-house dataset and publishing the results.

Requirements

- Being motivated and ready to learn some domain knowledge (a few biology concepts and bioinformatics practices).
- Being familiar with Unix environment (all analysis will be done in HPC Unix-based system).
- Programming skill in at least one of Python, Perl, R, and C++.

Miscellaneous

- Working with the supervisors and one IFI Master 2 intern.
- Chance to participate in exchange program abroad or continue the topic as Master 2 internship.

Références

(Please contact hobichhai@gmail.com if you need these papers' files)

[1] J. R. Rideout, Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell, S. M. Gibbons, J. Chase, D. McDonald, A. Gonzalez, A. Robbins-Pianka, J. C. Clemente, J. A. Gilbert, S. M. Huse, H.-W. Zhou, R. Knight, and J. G. Caporaso, "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences," *PeerJ*, vol. 2, no. 5, p. e545, 2014.

[2] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.," *Appl. Environ. Microbiol.*, vol. 72, no. 7, pp. 5069–5072, Jul. 2006.

[3] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glockner, "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools," *Nucleic Acids Research*, vol. 41, no. 1, pp. D590–D596, Dec. 2012.
[4] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.," *Appl. Environ. Microbiol.*, vol. 73, no. 16, pp. 5261–5267, Aug. 2007.
[5] D. Guttman, A. C. McHardy, and P. Schulze-Lefert, "Microbial genome-enabled insights into plant–microorganism interactions," *Nature Publishing Group*, vol. 15, no. 12, pp. 797–813, Sep. 2014.
[6] D. S. Lundberg, S. L. Lebeis, S. H. Paredes, S. Yourstone, J. Gehring, S. Malfatti, J. Tremblay, A. Engelbrektson, V. Kunin, T. G. D. Rio, R. C. Edgar, T. Eickhorst, R. E. Ley, P. Hugenholtz, S. G. Tringe, and J. L. Dangl, "Defining the core Arabidopsis thaliana root microbiome," *Nature*, vol. 488, no. 7409, pp. 86–90, Aug. 2012.
[7] B. CALLAHAN, D. PROCTOR, D. RELMAN, J. FUKUYAMA, and S. Holmes, "Reproducible research workflow in r for the analysis of personalized human microbiome data ," presented at the Proceedings of Pacific Symposium on Biocomputing, 2016, no. 21, pp. 183–194.
[8] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber, "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities," *Appl. Environ. Microbiol.*, vol. 75, no. 23, pp. 7537–7541, Nov. 2009.
[9] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight, "QIIME allows analysis of high-throughput community sequencing data.," *Nature Publishing Group*, vol. 7, no. 5, pp. 335–336, May 2010.

SUJET 19 : SEMI-AUTOMATIC IMAGE ANNOTATION: APPLICATION TO USTH ARCHIVES PROJECT

Encadrement

- NGHIEM Thi Phuong, TRAN Giang Son (USTH)
- HO Tuong Vinh (IFI, Hanoi)

Description:
This internship focuses on image annotation where each image is tagged with a set of relevant labels. Traditionally, annotation is performed manually by humans where the results raises some issues linked to human cost and subjectivity. Recently, researchers introduce automatic annotation where machine learning techniques are used to automatically build annotation models. Although, annotation results are generally good compared to manual annotation,

automatic annotation requires huge amount of labeled image data for the training. Due to this, automatic annotation becomes insufficient for the cases when the amount of available annotated data is limited. This problem is particularly true for the case of USTH ARCHIVES project where annotated historical documents of past natural disasters in Hanoi is very limited in numbers.

Objectives:

The goal of this internship is to learn a structured model from a limited set of manually annotated images and then seek for the help of final users (here the historians) in order to annotate upcoming new images. Two steps are considered. First, some automatic labeled images are presented to the final users for their relevance feedback. Then the response from the user about the correctness of a given label is used to propagated in the structured model for correcting the relevance scores of other labels.

Expected outcomes:

- Implement the annotation model described in Thomas Mensik's [1]
- Improve the Thomas' model and propagation modalities
- Apply the model to the dataset of ARCHIVES project

Pre¬requisites:

- Have good programming skills
- Have good theoretical background in data mining, machine learning, image processing and image analysis

References:

[1] https://staff.fnwi.uva.nl/t.e.j.mensink/publications/mensink12phd.pdf