5 Novembre 2015

8ème Journée des Thèses

14h00 -17h00 Journée des Thèses
Exposition de posters de doctorants
Présentation des équipes du LIRIS

17h30 -18h10 Séminaire
"Les acteurs de la recherche en informatique dans le bassin Lyonnais", Mohand-Saïd Hacid

18h10 -19h10 Buffet de clôture

Université Lyon1
Campus La Doua
Bâtiment Nautibus
Salles C1-C2

www.liris.cnrs.fr/jdt

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Lyon 1

LIRIS

cnrs

ÉCOLE CENTRALE

UNIVERSITÉ LUMIÈRE LYON 2

Le mot de la Présidente

La 8ème édition de la Journée des Thèses (JdT) aura lieu le jeudi 5 novembre 2015 de 14h à 17h, dans les salles C1 et C2 du bâtiment Nautibus de la Doua. Il faut signaler que cette édition est la 2ème de l'année 2015, après un travail conjoint de la direction et des doctorants du laboratoire pour faire évoluer son format. Grâce à ce travail d'autres événements, notamment l'accueil des nouveaux doctorants et la présentation des sujets de stage Master2 auront lieu dans la même journée.

La JdT du LIRIS est un événement organisé entièrement par les doctorants, qui leur permet de se rassembler autour de leur travail de recherche, de se connaître, et de se reconnaître comme partie active du laboratoire. C'est une opportunité exceptionnelle pour présenter leurs recherches au grand public, pour susciter des vocations de recherche entre les étudiants de Licence ou Master, et pour des échanges qui peuvent initier des projets inter-équipes. Pour cette édition, nous aurons l'opportunité de découvrir le travail de recherche de 20 doctorants en 2ème et 3ème année de thèse. Les doctorants participants ont préparé soigneusement leur poster pour les exposer et répondre aux questions des curieux. Ces posters couvrent les différentes thématiques de recherche des pôles du laboratoire : *Computer Vision Pattern Recognition ; Géométrie et Modélisation ; Data Science ; Simulation, Virtualité et Sciences Computationnelles ; Services, Systèmes distribuées et Sécurité ; et Interactions et Cognition.*

Comme il s'agit de la 2ème édition en 2015, je tiens à remarquer l'intérêt des doctorants à participer à cette journée et je les en remercie vivement. Je remercie fortement Mohand-Saïd Hacid et toute la direction du LIRIS pour le support d'organisation et financier, Dominique Barrière pour la gestion d'impression des posters, et les responsables des équipes et des formations pour permettre aux étudiants d'assister à cet événement durant les horaires de formation. Je tiens à remercier de manière spéciale chacun des membres de l'équipe organisatrice de la journée pour leur disponibilité et leur implication. J'ai bien appris de votre expérience.

Le présent document constitue les actes de la JdT 2015 bis. Il contient les travaux qui seront présentés, comme un aperçu avant l'événement et il constitue un support après. Vous trouverez les résumés des travaux ainsi qu'une copie des posters exposés le 5 novembre 2015.

Au nom de toute l'équipe organisatrice je vous souhaite de passer une belle Journée des Thèses 2015 bis. En espérant de vous voir dans les prochaines éditions,

Rubiela CARRILLO ROZO
Présidente de la JdT 2015 bis
https://liris.cnrs.fr/jdt/

Exposition de travaux de recherche des doctorants du LIRIS
Présentation des équipes du LIRIS

Université de Lyon 1, Campus de la Doua

Bâtiment Nautibus - Salles C1 et C2

Bryan Kong Win Chang
Assistance à l'utilisateur pour l'élicitation de connaissances du domaine lors de la génération d'exercices

Mehdi Terdjimi
Multi-level context adaptation in the Web of Things

Mahdi Bennara
Intégration des services liés dans le Web sémantique

Tarek Sayah
Contrôle d'accès aux données RDF distribuées

Alexandre Foncelle
Modeling the signaling pathway implicated in STDP: the role of endocannabinoid and dopamine signaling

Azzouz Hamdi-Cherif
Super-résolution basée sur l'auto-similarité pour les surfaces échantillonnées

Yazid Touileb
Patient-specific 4D dose calculation and treatment verification based on adaptative tetrahedral meshes

Fairouz Beggas
Étude de quelques paramètres de graphes : Décompositions et Dominations

Hayam Mousa
Management of selfish and malicious behavior in distributed collaborative systems

Natalia Neverova
Hand pose estimation by deep transductive learning

Bastien Moysset
Détection, localisation et typage de texte dans des images de documents hétérogènes par réseaux de neurones profonds

Mehdi Ayadi
Réalité Augmentée sur mobile en contexte urbain

Ouadie Gharroudi
Multi-label learning with Ensemble paradigm

Van Tinh Tran
Statistical Learning under Selection Bias

Gavin Kemp
Aggregating and Managing Big Realtime Data (AMBED) in the Cloud: application to intelligent Transport for Smart Cities

Sergio Peignier
Bio inspired data mining algorithms taking advantage of evolution of evolution

Charles Rocabert
Toward an Integrated Evolutionary Model to study Evolution of Evolution

Maximilien Guislain
Traitement joint de nuage de points et d'images pour l'analyse et la visualisation des formes 3D

Guillaume Bosc
Formalisation et mise en oeuvre de méthodes heuristiques de fouille de données massives et hétérogènes

Olivier Cavadenti
Fouille de traces unitaires de produits manufacturés

Assistance à l'utilisateur pour l'élicitation de connaissances du domaine lors de la génération d'exercices

Bryan Kong Win Chang, Nathalie Guin, Marie Lefevre

2ème année de thèse, Financement Ministère de la Recherche, Equipe Tweak
bkongwin@liris.cnrs.fr - http://liris.cnrs.fr/membres?idn=bkongwin

Résumé de la thèse

Au sein du LIRIS, des modèles et outils ont été développés pour proposer une génération semi-automatique d'exercices d'auto-évaluation. Ces générateurs s'appuient sur des modèles d'exercices, contenant les contraintes permettant de générer des exercices d'un type donné et portant sur des contenus propres à chaque discipline. Dans le cadre de cette thèse, on s'intéresse à la problématique de l'élicitation des connaissances du domaine par les créateurs des exercices.

Introduction

La base de travail utilisée dans le cadre de cette thèse est l'outil auteur ASKER (Authoring tool for aSsessing Knowledge through Evaluation exerRcices) [CABLE13]. Cet outil permet à des utilisateurs auteurs de créer des modèles d'exercices. Ces modèles sont utilisables par l'outil afin de générer un ensemble d'exercices et leurs corrections. Ces exercices peuvent être ensuite utilisés via l'interface d'ASKER par des apprenants dans le cadre d'une auto-évaluation. Disposer de connaissances du domaine serait intéressant pour d'une part faciliter la création de ces modèles et d'autre part fournir à l'apprenant un retour suite à ses erreurs (diagnostic). Pour atteindre ce but, un certain nombre d'étapes sont nécessaires. Nous avons tout d'abord identifié les connaissances potentiellement utiles que nous avons ensuite formalisées afin de proposer une acquisition fiable de ces connaissances.

Identifier les connaissances

La première étape de la thèse a consisté à faire l'inventaire des connaissances pouvant être nécessaires lors de la création des modèles d'exercices. Par exemple, les formules mathématiques sont des connaissances numériques utilisables pour la génération d'exercices dans le domaine scientifique. Un autre type de connaissance est par exemple une liste de mots sémantiquement liés, comme un ensemble de verbes utilisables pour générer des textes à trous, ou encore des règles permettant d'identifier ces verbes dans un texte. Ces connaissances étant très variées, nous avons défini un modèle général permettant de les formaliser.

Formaliser les connaissances

La formalisation proposée regroupe les connaissances autour de 3 concepts : les variables et constantes, les types et enfin les fonctions. Les variables et constantes sont des éléments ayant un nom et un ou plusieurs types. Un exemple de variable est une variable de nom « verbe » et dont le type est « verbe du premier groupe ». L'appartenance à ce type lui donne accès à une ou plusieurs fonctions d'instanciation et de manipulation. Une fonction d'instanciation va, à partir d'une variable, récupérer une instance ou un ensemble d'instances de la variable. Dans notre cas, une instance de notre variable « verbe » pourrait être le verbe « manger ». Une fonction de manipulation va récupérer un certain nombre de variables, instances de variables ou constantes pour produire un résultat. Par exemple, une fonction peut prendre la variable « verbe » et un mot et renvoyer oui si le mot passé en paramètre est une instance de « verbe » ou non. Cette même fonction peut être utilisée pour détecter puis enlever dans un texte donné l'ensemble des mots qui sont des verbes du premier groupe, créant par là même un exercice de type texte à trou.

Acquérir les connaissances

A partir de cette formalisation, la question qui se pose ensuite est celle de l'acquisition de ces connaissances. Une première méthode consiste à proposer un moyen pour l'utilisateur de définir des connaissances. Dans notre cas, la tâche est d'autant plus complexe que le domaine des connaissances n'est pas imposé et qu'on ne présuppose pas que notre utilisateur ait des connaissances en programmation. Ensuite, il nous faut trouver un moyen pour que ces connaissances puissent être partagées entre les différents utilisateurs, c'est-à-dire capitaliser ces connaissances. Une deuxième méthode pour acquérir ces connaissances consiste pour le système à observer l'activité de l'auteur qui définit des modèles d'exercices, afin de découvrir des connaissances qui seront proposées à l'utilisateur pour validation.

Références

[CABLE13] Baptiste CABLE, Nathalie GUIN, Marie LEFEVRE. An authoring tool for semi-automatic generation of self-assessment exercises. Conférence AIED 2013, Memphis, USA, pp.679-682.

Assistance à l'utilisateur pour l'élicitation de connaissances du domaine lors de la génération d'exercices

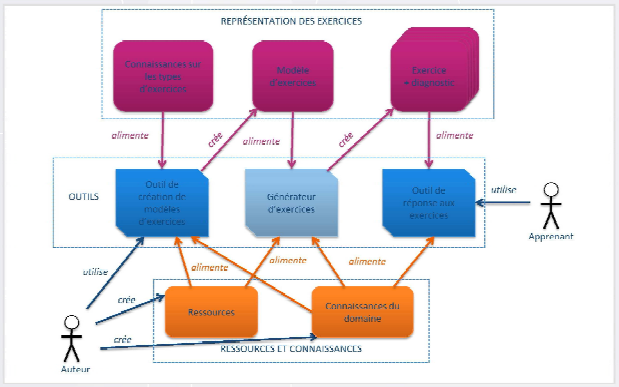
Bryan Kong Win Chang, Equipe TWEAK

Laboratoire d'InfoRmatique en Image et Systèmes d'information
LIRIS UMR 5205 CNRS / INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / Ecole Centrale de Lyon

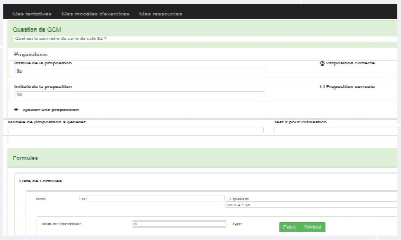
ASKER : Authoring tool for aSsessing Knowledge genErating exeRcises

Un outil semi-automatique de génération d'exercices :

- Des modèles d'exercices
- Des ressources diverses
- Des exercices et leurs corrections



Des connaissances dans les exercices



Un modèle pour plusieurs exercices :

- Un modèle décrit comment générer plusieurs exercices
- Des exercices générés à la demande, avec correction

Avantages d'avoir des connaissances :

- Générer les réponses au lieu de les lister
- Réutiliser ces connaissances dans d'autres modèles sans avoir à les redéfinir

Quelles connaissances sont utiles pour ASKER ? Comment définir ces connaissances ?

Comment permettre à l'auteur utilisant ASKER de définir et utiliser des connaissances du domaine pour créer ses modèles d'exercices ?

FORMALISATION DES CONNAISSANCES

VARIABLES ET CONSTANTES

Une variable nommée correspond à un ensemble de valeurs possibles qui sont ses instances. Une constante est une valeur ayant des propriété particulières.

Variable a
Variable b
Variable c
∏=3,14
Auxiliaires : être, avoir
Variable verbe

TYPES

Les types sont utilisés pour apporter des informations sur les variables et constantes.

Type Réel
Type Entier
Type Mot
Type Verbe du premier groupe

FONCTIONS

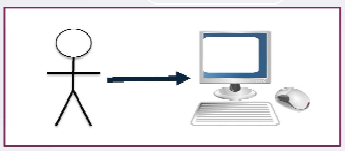
Une fonction prend en entrée des variables et valeurs et propose un résultat parmi les types existants.

$P = a * x^2 + b * x + c *$

Fonction RecoVerbe1erGroupe: Reconnait si un mot en entrée est un verbe du premier groupe

Variables → Types → Fonctions

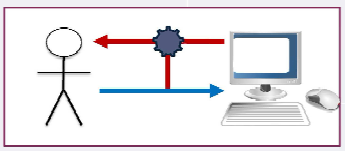
ELICITATER DES CONNAISSANCES



Pour des connaissances :

- Provenant de domaines variés
- Pour plusieurs types d'exercices différents

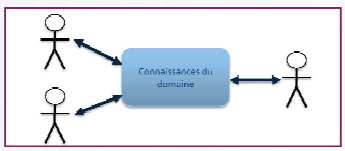
DECOUVRIR DES CONNAISSANCES



Découvrir des connaissances :

- Le système analyse les traces de l'utilisateur
- Le système propose des connaissances
- L'utilisateur modifie et/ou valide les propositions

PARTAGER DES CONNAISSANCES



Proposer un partage :

- Capitaliser les connaissances
- Mise en commun de connaissances générales

Multi-level context adaptation in the Web of Things

Mehdi Terdjimi, Lionel Médini, Michael Mrissa

1ère année de thèse, Financement ANR, Équipes SOC et TWEAK
mehdi.terdjimi@liris.cnrs.fr – https://liris.cnrs.fr/membres?idn=mterdjim

Résumé de la thèse

Le Web des Objets étend l'Internet des Objets afin de s'affranchir des silos propriétaires, reposant sur les technologies du Web. Cette thèse, qui s'inscrit dans le cadre du projet ASAWoO, a pour objectif de rendre les objets du Web sensibles au contexte en leur permettant de s'adapter à divers changements. De nouveaux challenges en résultent : de quelle manière peut-on raisoner sur des fonctionnalités d'objets ? Comment garantir la sécurité et la vie privée de l'utilisateur ? Comment assurer passage à l'échelle compte tenu du nombre important d'appareils connectés et de leur hétérogénéité ?

Introduction

Ma thèse s'inscrit dans le cadre du projet ANR ASAWoO [1]. Ce projet a pour objectif de porter des objets connectés sur le Web en tirant parti de ses technologies et standards : c'est Web des Objets. Dans notre cas, nous appellons "l'avatar" d'un objet sa représentation dans le monde virtuel. L'avatar d'un objet permet de l'augmenter en composant des fonctionnalités basiques en fonctionalités plus complexes, qui ne sont pas mises à disposition nativement. Ces fonctionnalités sont annotés sémantiquement, ce qui permet de raisonner dessus afin d'adapter leur comportement : par exemple si un composant lié à cet objet est défectueux, ou si la sécurité ne peut pas être assurée dans un certain environment, la fonctionnalité n'est pas réalisable et n'est donc pas exposée sur l'avatar.

Dans l'informatique pervasive (et donc le Web des Objets), les objets sont soumis à un nombre important de phénomènes, qu'ils soient environementaux, temporels, géographiques, etc. Comment adapter le comportement de ces objets au contexte ? Comment modéliser le contexte d'une manière suffisamment générique afin qu'il s'adapte à tout type de domaine et de besoin ?

Contexte multi-niveaux

Un travail de fond concernant l'état de l'art sur la modélisation du contexte a permis d'identifier de nombreux modèles et dimensions contextuelles, telles que la Localisation, le Temps, l'Environement, ou les politiques de Sécurité. Les modèles utilisés dans la littérature sont d'autant plus variés qu'ils s'appliquent à des domaines bien particuliers : l'architecture d'une application, la communication, le social, etc. L'idée n'est donc pas d'ajouter des dimensions aux modèles préexistants ou de totalement les remettre en cause. Notre contribution est donc de fournir des aspects (ou points de vue) à ces dimensions [1], ce qui permet une extension moins contraignante du modèle de contexte et permet la réeutilisation de dimensions à plusieurs niveaux.

1. https ://liris.cnrs.fr/asawoo/doku.php

Adaptation multi-niveaux

L'adaptation au contexte peut donc se faire sur différents aspects. Dans notre cas de figure, nous envisageons : 1) une adaptation niveau métier (domaine, fonctionnalités), et 2) une adaptation niveau raisonnement (processus, capacités) qui permet aux objets de raisonner via migration de code. La question est donc de savoir à quel endroit migrer et exécuter le code du raisonneur. Pour cela, nous développons le prototype HyLAR (pour Hybrid-Location Agnostic Reasoning) [2]. C'est une architecture qui sépare les étapes de raisonnement sur des ontologies : la classification, le parsing et la réponse aux requêtes SPARQL sur les ontologies classifiées. Ecrit en JavaScript, HyLAR permet d'exécuter des processus de raisonnement côté client [2].

Perspectives

Les perspectives suivantes sont envisagées pour la poursuite de cette thèse :

1. Proposer des règles d'adaptation pour la prise de décision (localisation du code, exécution, disponibilité d'une fonctionnalité...),
2. Gérer les aspects sécurité et vie privée,
3. Intégrer du raisonnement incrémental afin d'assurer le passage à l'échelle.

Références

[1] Mehdi Terdjimi. Multi-level context adaptation in the web of things. In *ISWC Doctoral Consortium 2015*, 2015.

[2] Mehdi Terdjimi, Lionel Médini, and Michael Mrissa. Hylar : Hybrid location-agnostic reasoning. In *ESWC Developers Workshop 2015*, page 1, 2015.

2. https ://github.com/ucbl/HyLAR

HyLAR: Hybrid Location-Agnostic Architecture
for Context-Aware Adaptive Reasoning

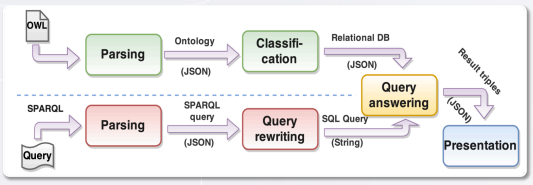
Laboratoire d'InfoRmatique en Image et Systèmes d'information
LIRIS UMR 5205 CNRS / INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / Ecole Centrale de Lyon

Motivations

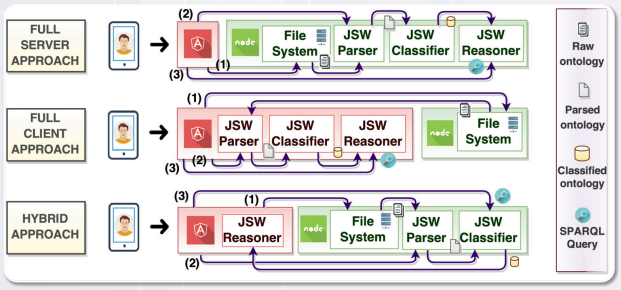
- Scalablility & availability of SPARQL endpoints
- Bridge the gap between the web and the semantic web *(Phil Archer, 2014)*
- Exploit client resources

Proposition

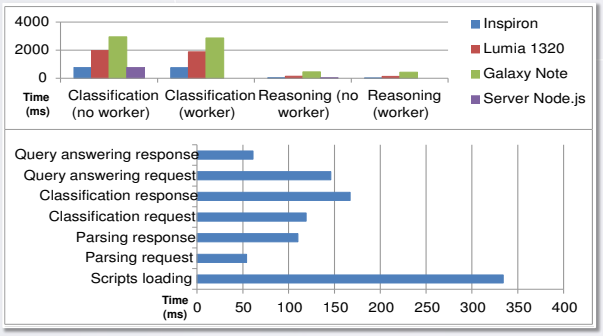
- Modularization of the reasoning steps
- Client-side code execution
- Chosen reasoner: OWLReasoner [1] OWL 2 EL, full JavaScript
- 3 architectures: full-client, full-server and hybrid [2]



Preliminary Evaluation



- *3 configurations x 3 devices, with and without web worker*
- *(1) Ontology Loading, (2) Ontology Classification, (3) Query Answering*



Time calculations depend on various parameters

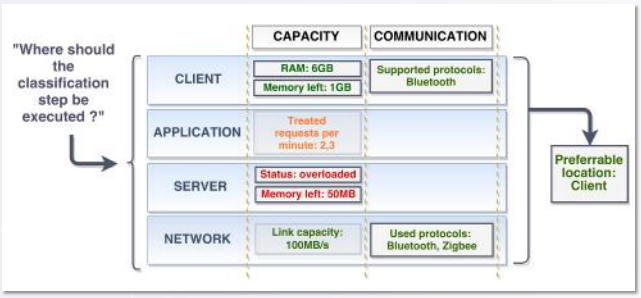
- for M users sending N queries,
- with Px, Qx, Rx respectively the processing, requesting and response times for a step (x) :

Server-side: $P2_{server} + M \times N \times (Q3 + P3_{server} + R3)$
Client-side: $M \times (Q0 + R0 + Q1 + R1) + P2_{client} + N \times P3_{client}$
Hybrid: $P2_{server} + M \times Q0 + R0 + Q2 + R2) + N \times P3_{client}$

Adapt code execution location

Contribution: Context-aware adaptive reasoning

- Evaluation parameters described as context dimensions
  - *The application dynamically deals with dimensions describing server and client capabilities, network status, server load and number of requests.*
- Abstraction of raw data as context states
  - *Client perspective, application architecture, network status, description of the physical world...*
- Context ontology: context states as named graphs
  - SPARQL querying on graphs
  - Response: change code execution location according to context state and available choices
    - *The example below shows that client-side classification is the best option as the server is currently overloaded and has no sufficient memory left, in comparison to the client.*



Execute code on the client side accordingly

Perspectives

- Complete SELECT, INSERT and DELETE support
- Incremental reasoning in view of dynamic Rule-based choices for the adaptation process
- Classification process optimization

References

[1] OWLReasoner, JavaScript Semantic Web Toolkit - https://code.google.com/p/owlreasoner/
[2] Terdjimi, M., Médini, L., & Mrissa, M. (2015, May). HyLAR: Hybrid Location-Agnostic Reasoning. *In ESWC Developers Workshop 2015.*

Intégration des services liés dans le Web sémantique

Mahdi BENNARA, Michaël MRISSA, Youssef AMGHAR

3ème année de thèse, Financement PALSE, Équipe SOC
mahdi.bennara@liris.cnrs.fr – http://liris.cnrs.fr/membres?idn=mbennara

Résumé de la thèse

Au cours des vingt dernières années, l'informatique orientée-services a favorisé l'interopérabilité entre les systèmes distribués. Malgré leur succès dans le monde de l'entreprise, les SOA posent encore des problèmes tels que le manque d'interopérabilité et d'évolutivité au niveau sémantique, ce qui a entravé leur adoption à grande échelle et sur le Web en particulier, ouvrant ainsi de nombreux défis pour la communauté de recherche. L'apparition du style architectural REST dans les dernières années a changé la donne. Le monde de l'entreprise s'intéresse de plus en plus aux services Web RESTful vu leur évolutivité et facilité d'adaptation au contexte large échelle. La problématique à laquelle on s'intéresse dans cette thèse concerne la description, la découverte ainsi que la composition de services REST dans le contexte ouvert et à large échelle.

Linked Web Services

Le Linked Data ou le Web des données est une approche visant à exposer et décrire des données structurées dans l'environnement du Web sémantique. Les Linked Web Services ou les services Web liés sont des resources Web qui utilisent les données sous forme de Linked Data, aussi bien pour la description que pour les échanges de données.

Le style architectural REST

REST est un ensemble de contraintes que les APIs des ressources Web doivent respecter afin d'avoir un comportement optimal qui assure l'indépendance entre le client et le serveur, la facilité de maintenance pour les applications ainsi qu'une haute tolérance aux pannes. REST [1] a été proposé par Roy Fielding dans sa dissertation en 2000.

Challenges de la thèse

Les principaux axes de recherche traités dans le cadre de cette thèse peuvent être résumés dans ce qui suit :

- Comment découvrir et sélectionner les services liés ?
- Comment connecter et composer ces services afin de créer des applications à valeur ajoutée ?
- Comment gérer les préoccupations de données liées à ces environnements de services (notamment : qualité, fiabilité et protection des données) ?

1 - Description

Afin de résoudre le problème de description nous proposons la notion de descripteur de ressource. Il s'agit d'une structure de données qui décrit et annote sémantiquement les opérations permises sur une ressource ainsi que l'ensemble de liens vers des ressources liées. Un descripteur est aussi une ressource, le lien entre ce dernier et la ressource qu'il décrit est établit à l'aide du LINK header. Un descripteur peut être partagé par plusieurs ressources qui ont le même comportement.

2 - Découverte

Le problème de découverte est traité en utilisant les algorithmes de parcours de graphes. Nous utilisons l'algorithms BFS pour le parcours du Web à la recherche des ressources nécessaires pour répondre à une requête client. Nous proposons aussi une variante qui se base sur les annotations sémantiques présentes sur les descripteurs pour obtenir une meilleure performance.

3 - Composition

Nous proposons la notion de Composition Directory pour répondre au problème de composition. Il s'agit de répertoires distribués sur le Web regroupant chacun un ensemble de workflows d'une composition de ressources Web. Chaque répertoire est associé à un utilisateur, un utilisateur peut accéder aux répertoires des autres utilisateurs s'il a le lien.

4 - Preoccupations de données

Nous proposons un modèle minimal pour la qualité de services afin d'explorer les attributs de QoS pour l'optimisation de l'algorithme de découverte. Les résultats de cet algorithme répondront aux besoins de l'utilisateur en termes de qualité.

Références

[1] Fielding, R.T. : Architectural styles and the design of network-based software architectures. Ph.D. thesis, University of California, Irvine (2000)

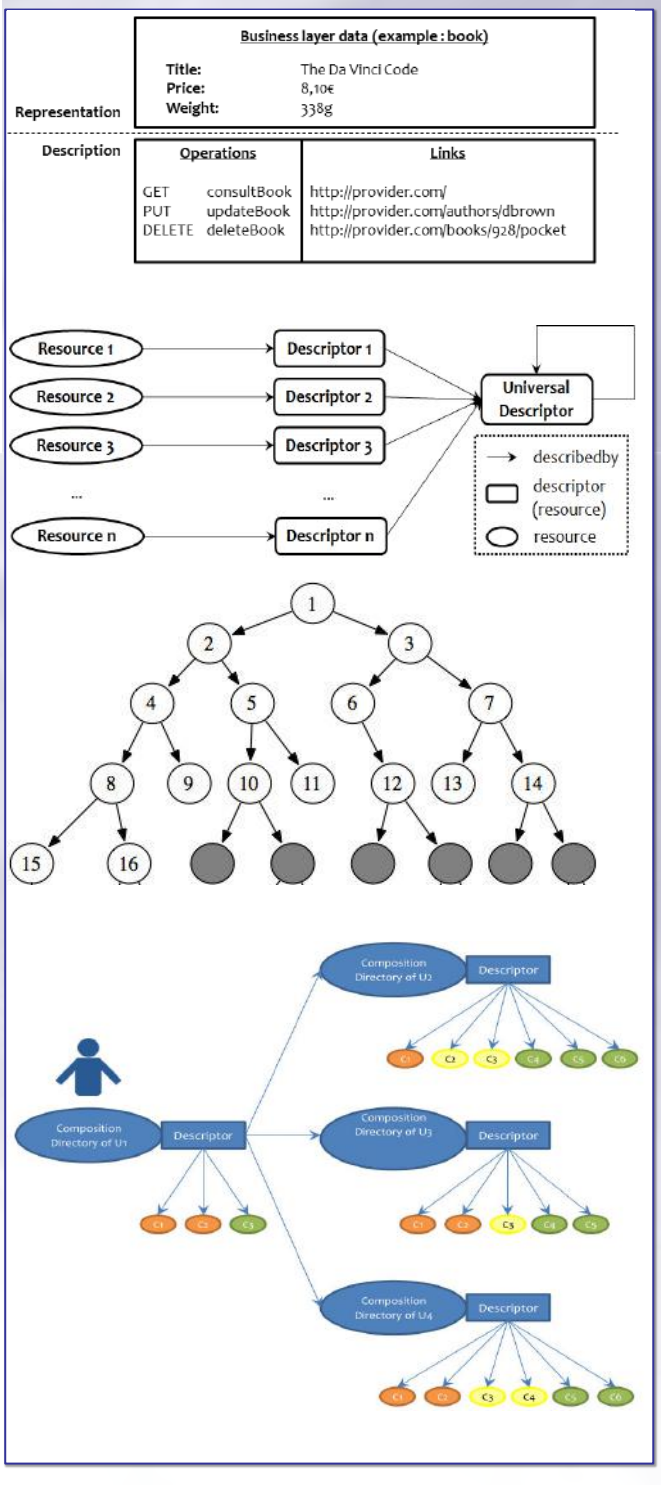
Intégration des services liés dans le Web sémantique

Challenges

- Comment découvrir et sélectionner les services liés ?
- Comment connecter et composer ces services afin de créer des applications à valeur ajoutée ?
- Comment gérer les préoccupations de données liées à ces environnements de services (notamment : qualité, fiabilité et protection des données) ?

Contexte

- Web des ressources
- Linked Data
- Linked Web Services
- RESTful Web Services
- Web Sémantique



Solutions

- Description
  - Notion de Descripteur
    - Description des opérations permises sur la ressource
    - Description des liens vers des ressources liées
    - Annotations sémantique
- Découverte
  - Algorithme BFS
    - Parcours en largeur du Web
    - Recherche des ressources qui offrent les opérations nécessaires
    - Ressources liées directement ont plus de chances de participer à la composition pour répondre à la requête
  - Variante sémantique
    - Utilisation des annotations
    - Ignorer les ressources similaires
    - Prendre en compte les ressources complémentaires
    - Explorer les liens complémentaires sur les ressources similaires
  - Variante sur la base de QoS
    - Exploration des attributs QoS de la ressource
    - Profil client
    - Contraintes soft
    - Contraintes hard
- Composition
  - Composition Directory
    - Un CD par utilisateur
    - Contient des compositions de l'utilisateur
    - Composition = workflow
    - Partage de CD
    - Partage de Composition
    - Edition de composition
- Préoccupations des données
  - Qualité des services
    - Modèle minimal pour l'algorithme de découverte
    - Les attributs les plus communs

Contact : Mahdi BENNARA mahdi.bennara@liris.cnrs.fr

Contrôle d'accès aux données RDF distribuées

Tarek Sayah, Emmanuel Coquery, Romuald Thion, Mohand-Saïd Hacid

3ème année de thèse, Financement Autre, Équipe BD
tarek.sayah@liris.cnrs.fr – http://liris.cnrs.fr/membres?idn=tsayah

Résumé de la thèse

RDF (Resource Description Framework) est devenu le format standard pour capturer les relations sémantiques entre les ressources web et représenter ces données dans une forme compréhensible par la machine. Les applications qui partagent et échangent des données RDF, potentiellement sensibles, sur le web augmentent de plus en plus dans de nombreux domaines comme la bioinformatique et le e-gouvernement. La problématique de l'exposition sélective des contenus RDF est devenue très importante en particulier dans le cotexte de l'open-data. L'objectif de cette thèse est de poser les bases d'un contrôle d'accès aux données RDF afin de répondre à cette problématique.

Introduction

De part le rôle central occupé aujourd'hui par le modèle de données RDF dans le Web sémantique, la problématique du contrôle d'accès à ces données prend de plus en plus d'importance. Une réponse est particulièrement attendue dans le contexte de l'open-data, car elle encouragerait les fournisseurs de données à mettre à disposition du public des données selon leurs propres termes et de façon sélective.

Modèle de contrôle d'accès

En premier lieu, nous avons commencé par définir un langage de contrôle d'accès pour les entrepôts RDF [1]. Ce langage permet de définir des autorisations à grain fin, c'est-à-dire au niveau d'un triplet, pour un utilisateur fixé. Une modélisation souple et expressive des stratégies de résolution des conflits entre autorisations est proposée. Elle permet entre autre de représenter les stratégies connues dans la littérature du contrôle d'accès telles que "Denials Take Precedence", "Permissions Take Precedence" ou encore "Most Specific Takes Precedence" où l'autorisation la plus spécifique est appliquée. La sémantique formelle du langage est définie en sous-graphe autorisé partir du graphe de base, les requêtes étant exécutées sur ce sous-graphe plutôt que sur le graphe originel. De ce point de vue notre mécanisme de contrôle d'accès est indépendant du langage de requête.

Fuites dues à l'inférence

L'utilisation de systèmes de déductions est commune dans le Web sémantique, comme par exemple les règles d'inférences liées au standard RDFS. Une des particularités du Web sémantique est la capacité à inférer de nouveaux faits. En effet, l'inférence dans le web sémantique permet de découvrir de nouvelles relations pour améliorer l'intégration dans le web, analyser automatiquement le contenu des données, gérer des connaissances dans le web ou détecter des inconsistances dans les données intégrées. Cependant, ce type d'inférence peut être utilisé par un utilisateur malicieux pour déduire des données confidentielles à partir de données publiques. Nous appelons ce problème "Fuite d'inférence". Nous avons formalisé ce problème et énoncé une propriété de *cohérence des autorisations vis-à-vis des règles d'inférence*. Cette dernière doit être respectée par la politique d'autorisation pour éviter les fuites d'inférence. Nous proposons un algorithme qui permet de vérifier statiquement la cohérence d'une politique vis-à-vis de règles d'inférence [1].

Travaux en cours

Actuellement nous travaillons à étendre ce travail à des politiques multi-utilisateurs :

- Via un langage de haut-niveau pour attribuer à chaque utilisateur un ensemble d'autorisations.
- Via une factorisation du calcul des graphes autorisés pour limiter le surcoût du contrôle d'accès.

Nous envisageons aussi d'étudier l'impact des mises à jour des données RDF, en effet, de nouveaux problèmes découlent des mises à jour, par exemple un utilisateur peut être autorisé à insérer un triplet, mais elle/il peut ne pas être autorisé à insérer certaines de ses conséquences qui peuvent être déduites.

Références

[1] Tarek Sayah, Emmanuel Coquery, Romuald Thion, and Mohand-Said Hacid. Inference leakage detection for authorization policies over RDF data. In *Data and Applications Security and Privacy XXIX*, pages 346–361, 2015.

Contrôle d'accès aux données RDF distribuées

Tarek Sayah, Emmanuel Coquery, Romuald Thion, Mohand-Saïd Hacid

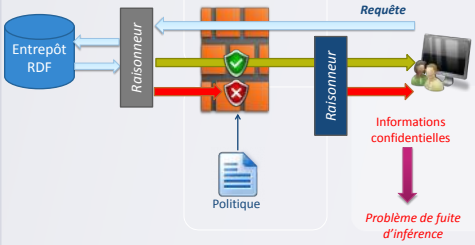
Laboratoire d'InfoRmatique en Image et Systèmes d'information
LIRIS UMR 5205 CNRS / INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / Ecole Centrale de Lyon

Contexte

- Adoption rapide de RDF (Resource Description Framework)
- Augmentation des applications qui publient des données RDF **sensibles** : bio-informatique, e-gouvernement ...
- **Problématique de contrôle d'accès au contenu RDF**
- **Encourager les fournisseurs de données à publier leurs données**

Objectif

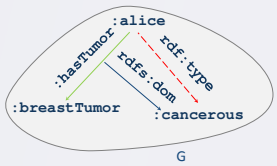
- **Définir un cadre flexible et expressif pour contrôler l'accès aux données RDF en présence de l'inférence**
  - Autorisations à grain fin (plus expressives)
  - Stratégies de résolution de conflits plus expressives que le « Denials Take Precedence », « Permissions Take Precedence » et le « Most Specfic Takes Precedence »
  - Prise en charge des triplets inférés
    - Détection de fuite d'inférence



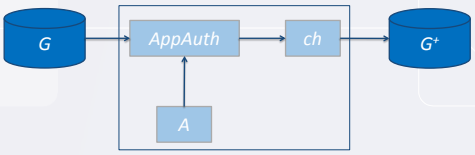
Autorisations et règles d'inférence

$\mathcal{A}_1$ = GRANT(?p; :hasTumor; ?t)

$$\frac{(?p; rdfs:dom; ?d) \quad (?x; ?p; ?y)}{(?x; rdfs:type; ?d)}$$



Modèle de contrôle d'accès [1]



G : Le graphe de base
A : L'ensemble des autorisations
AppAuth : Autorisations applicables
ch : Fonction de choix pour la résolution de conflits
G+ : Le sous-graphe autorisé

Fuite d'inférence

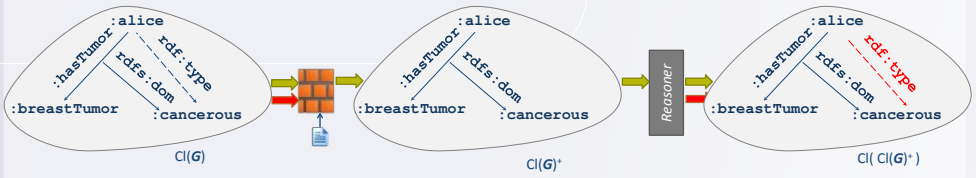
Propriété de consistance

$\mathcal{A}_1$ = GRANT(?p; :hasTumor; ?t)

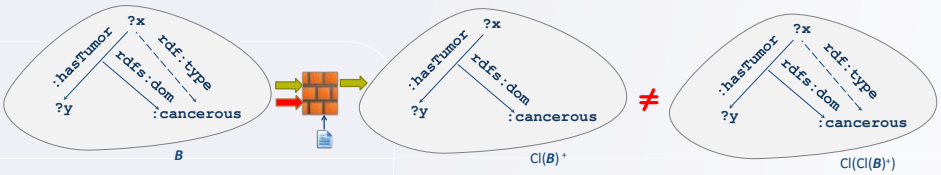
$\mathcal{A}_2$ = DENY(?p; rdf:type; :cancerous)

$\mathcal{A}_3$ = GRANT(?p; rdfs:dom; ?s)

$$\mathrm{RDom} = \frac{(?p; rdfs:dom; ?d) \quad (?x; ?p; ?y)}{(?x; rdfs:type; ?d)}$$



Algorithme de génération de patrons de graphes sources de conflits



Travaux en cours

Actuellement nous travaillons à étendre ce travail à des politiques multi-utilisateurs:

- Via un langage de haut-niveau pour attribuer à chaque utilisateur un ensemble d'autorisations.
- Via une factorisation du calcul des graphes autorisés pour limiter le surcoût du contrôle d'accès.
- Via le traitement des requêtes de mises à jours du graphe RDF.

[1] Tarek Sayah, Emmanuel Coquery, Romuald Thion, and Mohand-Said Hacid. Inference leakage detection for authorization Policies over RDF data. In *Data and Applications Security and Privacy XXIX - 29th Annual IFIP WG 11.3 Working Conference, DBSec 2015*, pages 346–361, 2015.

Modeling the signaling pathway implicated in STDP: the role of endocannabinoid and dopamine signaling

Alexandre Foncelle, Hugues Berry

2nd year of PhD, Funded by ANR, BEAGLETeam
alexandre.foncelle@liris.cnrs.fr – http://liris.cnrs.fr/membres?idn=afoncell

Abstract

This PhD project consists in developing computational models to investigate the bidirectional interactions between dopamine and intracellular calcium dynamics. By accounting for endocannabinoid production and interactions with dopamine (via PKA), we aim to develop a predictive function for plasticity direction and magnitude from stimulation parameters, and set the stage for unraveling the dynamical dopaminergic modulation of plasticity in physiological and pathological conditions.

Introduction

The influence of a neuronal cell on another neuron to which it is connected varies with time in a process called "synaptic plasticity". This process, the main cellular mechanism underlying learning and memory, has thoroughly been studied by neurobiologists at the cellular level (electrical activity of the cells). However, albeit experimental results in the past decade have identified a large set of signal transduction proteins involved in synaptic plasticity, we still do not understand its operation at the molecular level nor how to explain cellular responses on the basis of the molecular levels. For this purpose, one needs to build computational and mathematical models of the implied signal transduction networks thus effectively developing computational systems biology of synaptic plasticity. We are working on three independant projects at different scale of modelisation.

Molecular scale

This is the main project of my PhD. In collaboration with the experimental biology lab led by Laurent Venance at College de France, Paris (CNRS/ UMR 7241 - INSERM U1050) and with Kim Avrama Blackwell's lab at George Mason University (Fairfax, USA) for the details of the modeling of dopamine signaling [Lindskog2006], we are studying the molecular bases of synaptic plasticity (spike-timing dependent plasticity, STDP) in a part of the brain called the basal ganglia (involved in procedural learning).

Using a joint experimental-modeling approach, we have developed a computational model of the implicated molecular networks at play (glutamate receptors, CaMKII, endocannabinoids, PKA...) [Graupner2007], that precisely allows understanding cellular responses based on the molecular levels for a restricted set of experimental conditions.

Cell scale

As a side-project we collaborate with Yihui Cui from the experimental biology lab of Zhejiang University, China on the effect of dopamine in depression disease. The habenula region send dopamine to basal ganglia region and is known to be involved in depression disease. We are studying a neural-network spread between these different regions of the brain to show that differences in activity of habenula's neurons lead to dopamine perturbations, by using voltage recordings of neurons provided by Yihui Cui.

Brain region scale

Finally, with the experimental biology lab led by L. Venance at College de France, we work to analyse differences between neural responses of control rats and Parkinsonian rats. They came up with a protocol where they stimulate the cortex and they record the response of one neuron in the striatum, in each group of rat. By stimuling the cortex, they can observe a caracteristic tri-phasic response in the striatum, resulting of three pathway crossing different regions in the brain.
To caracterize this response and what differs with Parkinsonian condition, we build a network of each brain region implied and we hope to get caracteristic parameters for each group.

Bibliography

[Lindskog2006] Lindskog et .al, *Transient Calcium and Dopamine Increase PKA Activity and DARPP-32 Phosphorylation*, PLoS Comput Biol, 2006

[Graupner2007] Graupner and Brunel, *STDP in a Bistable Synapse Model Based on CaMKII and Associated Signaling Pathways*, Public Library of Science, 2007

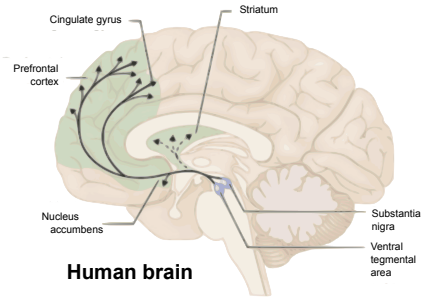
Modeling the triphasic response in the SNr

Alexandre Foncelle, Hugues Berry
BEAGLE team

alexandre.foncelle@liris.cnrs.fr
2nd year PhD, ANR financing
JdT 2015, Lyon

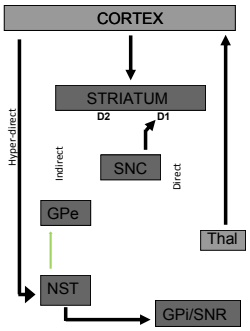
Biological background

Basal ganglia

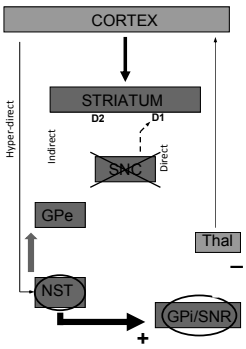


Human brain

Control condition



Parkinsonian condition



SNc : Subtancia nigra pars compacta
SNr : Substancia nigra pars reticulata
GPi : Internal part of the globus pallidus
Gpe : External part of the globus pallidus
Thal : Thalamus
NST : Sub-thalamic nucleus

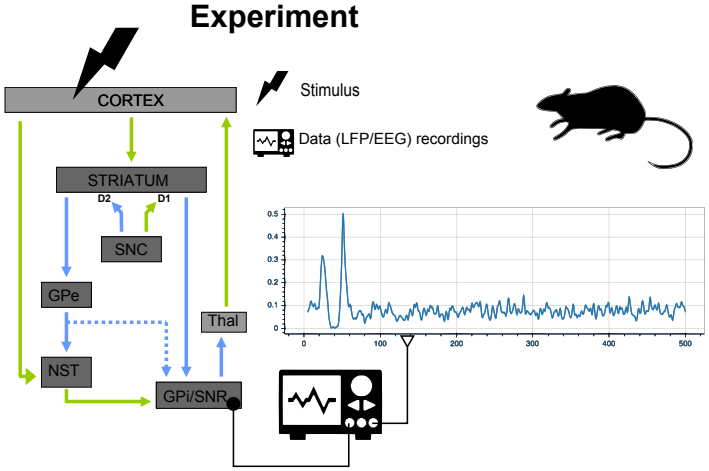
D1 : Dopaminergic receptors of type 1
D2 : Dopaminergic receptors of type 2

Excitatory pathway
Inhibitory pathway

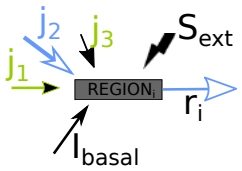
Hyperactivity

Material and methods

Experiment



Model

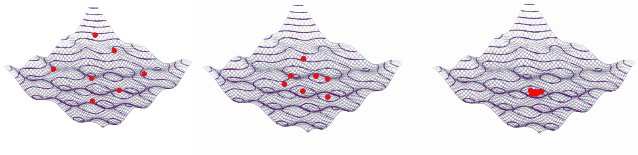


$$\tau_i \frac{dr_i(t)}{dt} = -r_i(t) + F(I_{basal_i} + S_{ext_i} + \sum_{j=1}^{n} r_j(t - t_{delay_{(i,j)}}) \times w_{(i,j)})$$

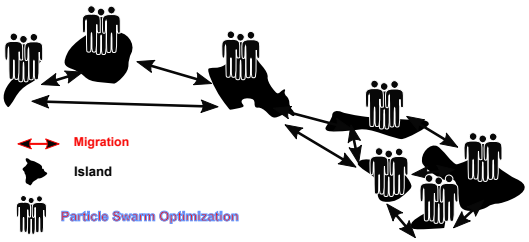
$$F(x) = \begin{cases} x \text{ if } x >= 0 \\ 0 \text{ if } x < 0 \end{cases}$$

Fitting algorithm

Particle Swarm Optimization (PSO)

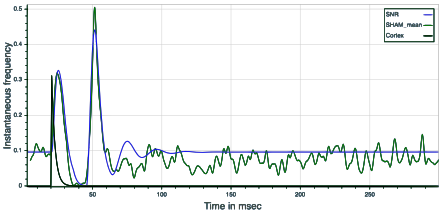


Parallelization

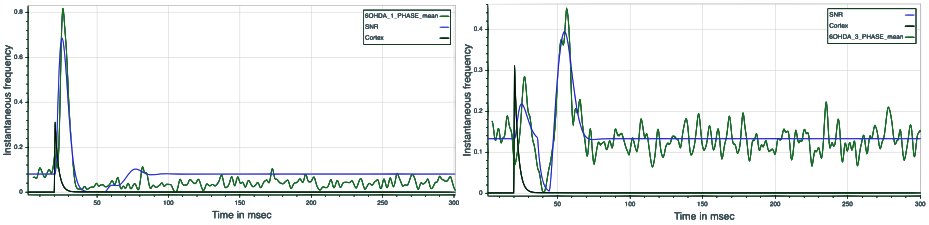


Results

Control condition



Parkinsonian condition



INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON
LIRIS
Inria INVENTEURS DU MONDE NUMÉRIQUE

Super-résolution basée sur l'auto-similarité pour les surfaces échantillonnées

Azzouz Hamdi-Cherif, Julie Digne, Raphaelle Chaine

2ème année de thèse, Financement Bourse ministérielle, Équipe GeoMod
azzouz.hamdi-cherif@liris.cnrs.fr – https://liris.cnrs.fr/membres?idn=ahamdich

Résumé de la thèse

La démocratisation des appareils d'acquisitions 3D grand public soulève de nouveaux défis algorithmiques. En effet, ces appareils proposent une numérisation avec un faible niveau de détail, d'où le besoin de solutions logicielles permettant un gain en qualité. Cependant, les formes numérisées présentent souvent une forte auto-similarité jusqu'alors inexploitée. Nous nous proposons d'utiliser cette auto-similarité pour augmenter le niveau de détails.

Introduction

Les appareils d'acquisition grand public, peu onéreux, proposent encore un niveau de détail très bas. Il est alors intéressant d'essayer de travailler sur l'amélioration logicielle de ces données. Il s'agit du problème de super-résolution qui consiste à améliorer la résolution d'un appareil de mesure. Ce problème bien que très étudié en image [1], est très peu exploré pour les surfaces [2].

Dans ce travail, l'approche adoptée consiste à s'appuyer sur l'*auto-similarité* des formes. En effet, deux endroits d'une même structure ont de fortes chances d'avoir des propriétés statistiques semblables, par exemple, l'écorce en deux endroit distincts d'un même tronc d'arbre est très semblable. L'auto-similarité est un atout pour résoudre le problème de la super-résolution. En effet, une seule acquisition d'un objet comportant des zones similaires peut être interprêtée comme plusieurs acquisitions d'une même zone. On se rapproche alors du problème de la super-résolution à partir d'images multiples.

Le but de ces travaux est donc de parvenir à capturer numériquement cette auto-similarité et de s'en servir pour dépasser la résolution physique limite de l'appareil d'acquisition à l'origine de la numérisation.

Méthode proposée

La super-résolution pouvant se résumer par l'ajout d'information (géométrique ou de texture), il est nécessaire de développer une méthode exploitant le maximum de la similarité disponible. Nous proposons d'utiliser un descripteur caractérisant la surface sous-jacente par un échantillonnage régulier et intrinsèque afin de mettre en évidence les similarités locales de la surface échantillonnée.

Dans le contexte de super-résolution, un descripteur local doit décrire localement l'information géométrique, c'est à dire, la surface sous-jacente au nuage de points. Il doit également être intrinsèque, en particulier, invariant aux transformations rigides (rotations et translations). Il doit résister, dans une certaine mesure, au bruit et aux variations d'échantillonnage souvent présentes car inhérentes aux appareils d'acquisitions les plus utilisés (scanner laser). Ces caractéristiques ont pour objectif de permettre la comparaison simple et efficace de deux régions d'une surface à travers leurs descripteurs respectif.

La première étape de notre méthode consiste à recouvrir l'entièreté de la surface par nos descripteurs locaux, soit D l'ensemble des descripteurs crées. Pour chaque descripteur $d \in D$, on identifie les descripteurs qui lui sont similaires S_d. Enfin, fort des informations supplémentaires contenues dans S_d on crée une version super-résolue de d. L'augmentation de la résolution de d est ici justifiée car on a accès à plus d'information.

Nos résultats actuels (fig. 1) surpassent le gain proposé par des méthodes récentes d'interpolation, validant ainsi l'augmentation d'information significative extraite par notre méthode permettant de dépasser les limites physiques d'un appareil d'acquisition 3D.

FIGURE 1 – (Gauche) Nuage de points en entrée, (droite) nuage de points produit par notre super-résolution. (reconstruits par *screened poisson*).

Références

[1] Michael Elad and Yacov Hel-Or. A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *IEEE Transactions on Image Processing*, 10(8) :1187–1193, 2001.

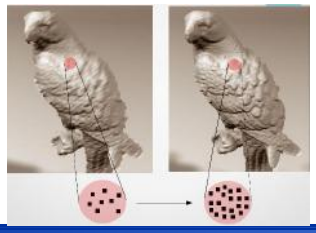
[2] Yong Joo Kil, Boris Mederos, and Nina Amenta. Laser scanner super-resolution. In *Proceedings of the 3rd Eurographics / IEEE VGTC Conference on Point-Based Graphics*, SPBG'06, pages 9–16, 2006.

Super-résolution basée sur l'auto-similarité pour les surfaces échantillonnées

Azzouz Hamdi-Cherif, Julie Digne,
Raphaëlle Chaine : équipe GeoMod

Résumé

La démocratisation des appareils d'acquisitions 3D grand public soulève de nouveaux défis algorithmiques. En effet, ces appareils proposent une numérisation avec un faible niveau de détail, d'où le besoin de solutions logicielles permettant un gain en qualité. Cependant, les formes numérisées présentent souvent une forte auto-similarité jusqu'alors inexploitée. Nous nous proposons d'utiliser cette auto-similarité pour augmenter le niveau de détails.

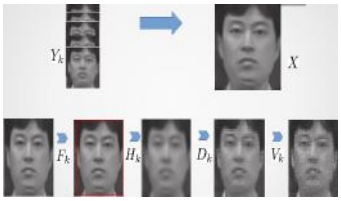


Contributions

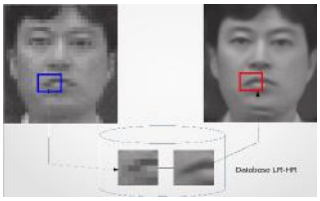
- Un seul scan requis
- Aucune contrainte sur l'appareil d'acquisition
- Sans étape de recalage
- Extraction de l'auto-similarité
- Descripteur local basé sur la quadrique

État de l'art

Super-résolution image : deux grandes familles

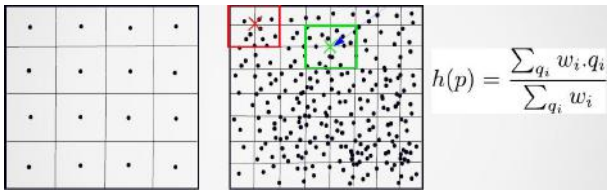


SR Classique [1]



SR basée exemple [2]

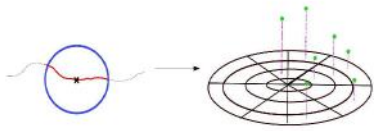
Super-résolution 3D :



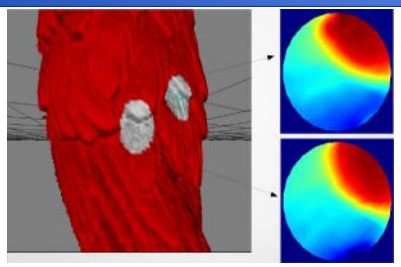
Adaptation de la SR Classique aux surfaces échantillonnée [3]

Méthode proposée

1 - Caractérisation

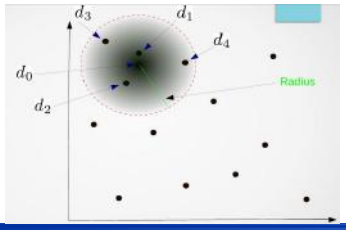


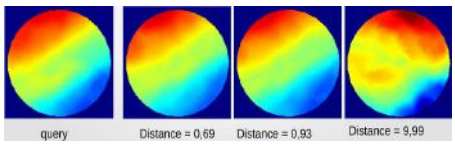
Échantillonnage + repère consistent régulier → représentation intrinsèque



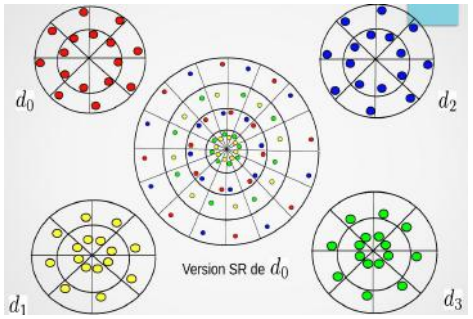
Comparaison facilitée

2 - Extraction



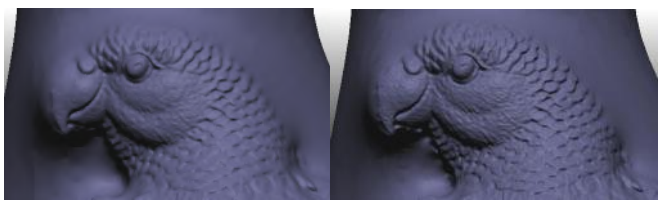


3 - Agrégation



$$h_{sr}(b) = \sum_{d_j} \left[w_{sim}(d_j, d_0) \sum_{q_i \in d_j} w(q_i, b) h(q_i) \right]$$

Résultats



Conclusion et perspectives

- Amélioration des positions des descripteurs
- Descripteur basé sur l'information fréquentielle
- Introduction d'une approche basée dictionnaire

Références

1. Elad & Hel-Or. A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *IEEE Transactions on Image Processing*, 2001.
2. Freeman & al. Example-Based Super-Resolution. *IEEE Comput. Graph. Appl.*, 2002.
3. Kil & al. Laser Scanner Super-resolution. *IEEE VGTC Conference on Point-Based Graphics*, 2006.

LIRIS — INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON — cnrs — Lyon 1

Patient-specific 4D dose calculation and treatment verification based on adaptative tetrahedral meshes

Yazid Touileb, Hamid Ladjal, Michael Beuve, Behzad Shariat

2nd year of PhD, Funded by Labex Primes, SAARATeam
yazid.touileb@liris.cnrs.fr – http://liris.cnrs.fr/membres?idn=ytouileb

Abstract

The estimation of the distribution pattern of energy and dose in respiratory-induced organ motion constitutes a big challenge in hadron therapy treatment planning and dosimetry, notably for lung cancer in which many difficulties arose, like tissue densities variation and the tumor position shifting during respiration. All these parameters affect the ranges of protons or ions used in treatment when passing through the matter and can easily induce to unexpected dose distribution. Our work consists of calculating the dose distributions of moving organs by means of Monte Carlo simulations. The dose distributions are calculated using a time-dependent tetrahedral density map describing the internal anatomy and respecting the principle of mass conservation. Unlike methods based on deformable image registration, the deposited energy is accumulated inside each deforming tetrahedron of the meshes, thus overcoming the issues related to dose interpolation. The objective of this thesis is the construction of an adaptive tetrahedral deformable model that can be used in the field of particle matter simulations and also for treatment verification with positron emission tomography or gamma prompt imaging. Besides, technical challenges have to be addressed to optimize this structure, including the improvement of simulation time and the validation of our approach on a real patient case. Furthermore, the validation of the tetrahedral model has to be performed using an anthropomorphic breathing phantom named LuCa incorporating a lung tumor model and a typical thoracic anatomy.

Beam shaping

In order to take into account the shape of the tumor in hadrontherapy beam line simulation and to improve the efficiency of treatment delivery, we have developed an algorithm that construct a patient-specific range compensator(RC) and multileaf collimator(MLC) from CT-images. Theses two devices are used in passive beam scattering to shape the beam and to minimize the delivery dose in organs at risk (OAR).

Tetrahedral model construction

We have developed a multiresolution tetrahedral model that takes into account the patient geometry and the volumetric mass density of different tissues. For the sake of making the model patient-specific, we have defined a pipeline that constructs the model from scratch and only by using 3D computed tomography images. This proceeding combines a set of algorithms that build the tetrahedral meshes of all the organs and embed densities issued from the voxel images in their nodes.

Geant4 implementation

A new layer was added to Geant4 platform to integrate this model and to perform Monte Carlo simulations on it using a passive scattering beam line and all the information related to the tumor shape and position. The energy and dose deposited in the tissues are accumulated in the elements of the meshes in each step of the breathing simulation. Since the model is multiresolution we can embed other information rather than densities or doses, and it can be used to improve 4D in-beam PET image reconstructions for treatment verification.

Evaluation and results

A comparison of the tetrahedral model and the conventional voxel-based structure based on CT-images was performed to evaluate the accuracy of dose distributions. These two structures were constructed based on a real patient anatomy, then, the movement derived from deformable image registration algorithm (DIR) of the set of CT-images was added to simulate human breathing. Final results show that dose distributions for both representations are in a good agreement, and dose homogeneity is about the same. However, motion-induced dose accumulations are more intuitive using tetrahedral model since they do not introduce additional uncertainties with image re-sampling and interpolation methods, and also for the fact that they respect the principle of mass conservation.

Conclusion

A unified model of 4D radiotherapy respiratory effects was developed where motion is coupled with dose calculations. Promising results demonstrate that this approach has significant potential for the treatment for moving tumors.

Dose calculations and motion compensated PET image reconstruction using deformable mesh models

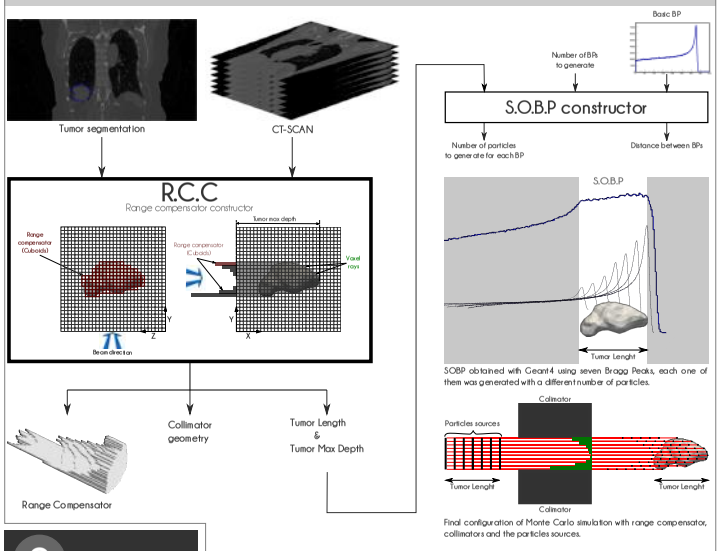
Yazid Touileb, Petru Manescu, Hamid Ladjal, Joseph Azencot, Michæl Beuve, Behzad Shariat

Laboratoire d'Informatique en Image et Système d'Information (LIRIS) CNRS UMR 5205, Université Claude Bernard Lyon 1, France
The Institute of Nuclear Physics of Lyon (INPL), UMR 5822 Université Claude Bernard Lyon 1, France

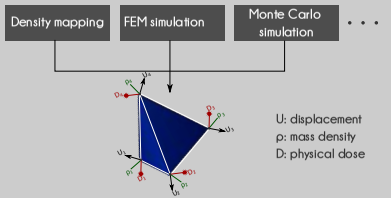
INTRODUCTION

Respiratory-induced organ motion is a technical challenge to nuclear imaging and to charged particle therapy dose calculations for lung cancer treatment in particular. Internal organ tissue displacements and deformations induced by breathing need to be taken into account when calculating Monte Carlo dose distributions as well as when performing tomographic reconstructions for PET imaging.
This project presents a methodological contribution to physical dose calculations and PET-based treatment verification for hadron therapy in the case of moving tumours. Contrary to DIR-based methods where motion is described by relative voxel displacement, each organ is represented as a deformable grid of tetrahedra where internal motion is described by mesh vertex transformations calculated using continuum mechanics. First, we proposes a method to calculate four dimensional dose distribution over tetrahedral meshes, which are deformed using biomechanical modeling based on Finite Element Analysis (FEA). Then, we focuses on motion compensation for PET image reconstruction using deformable tetrahedral meshes.

1 SOBP AND RANGE COMPENSATOR CONSTRUCTION

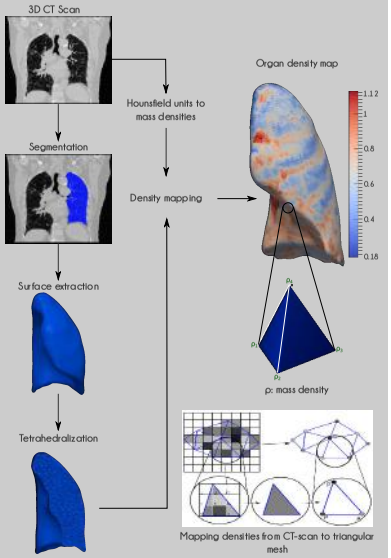


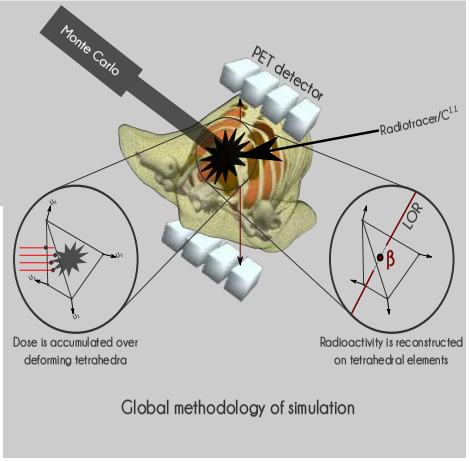
2 MULTI PHYSICS TETRAHEDRAL MESH



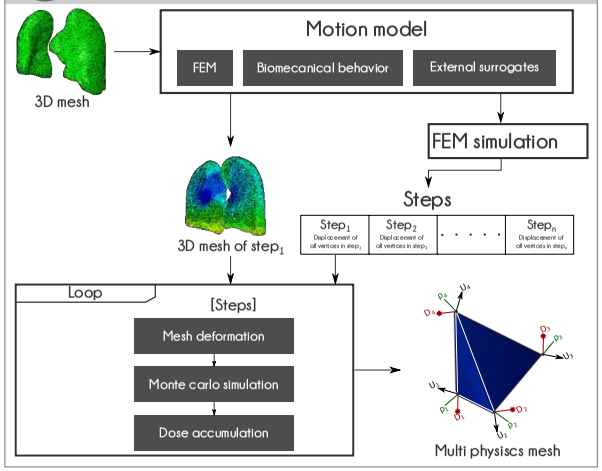
3 DENSITY MAPPING

Based on the principle of mass conservation we propose a new method to map material density information issued from CT-images to deformable tetrahedral meshes.

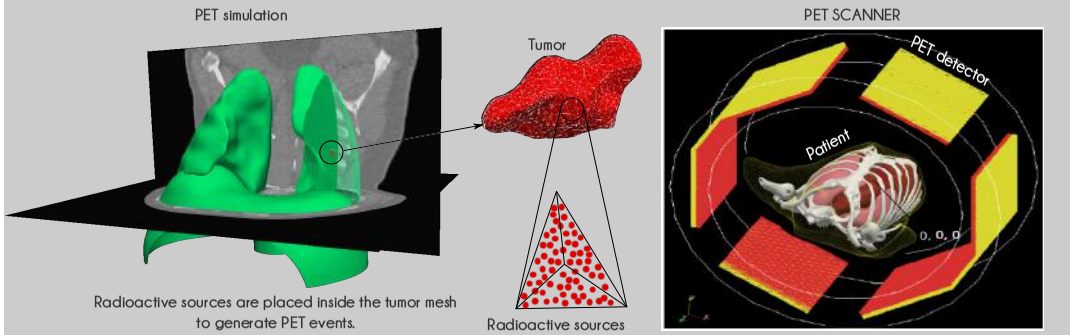




4 4D MONTE CARLO SIMULATION

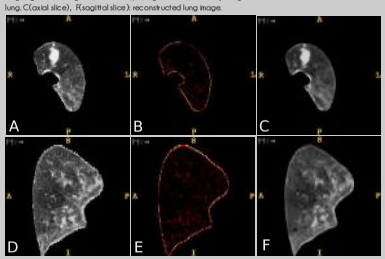


5 PET IMAGE RECONSTRUCTION

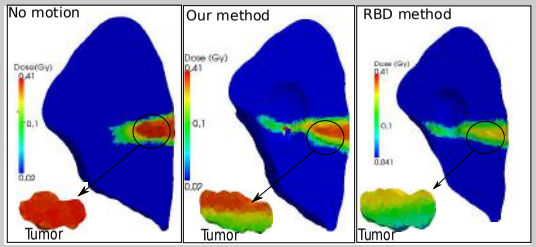


6 RESULTS

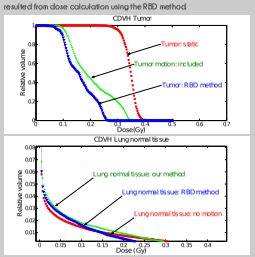
Density image reconstruction error map. A(axial slice), D(sagittal slice) : original lung segmented image. B(axial slice), E(sagittal slice) : density image error for the lung. C(axial slice), F(sagittal slice): reconstructed lung image.



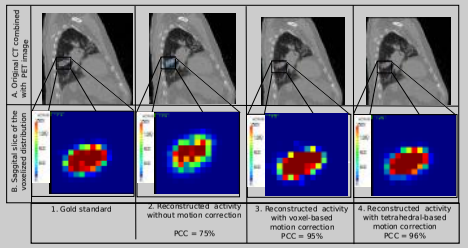
3D tetrahedral dose maps for different scenarios. left: Static simulatio. Center: Dose distribution taking into account organ motion using our method. Right: Dose distribution using the RBD method.



CDVH for different scenarios. The red curve corresponds to the static simulation, the green to our method and the blue one to the CDVH resulted from dose calculation using the RBD method.



Reconstructed activities on the patient case. The gold standard image is the reconstructed image when no motion was applied to the meshes. PCC: the Pearson Correlation Coecient.



CONCLUSION

We have deloped a software-plateform including both four dimensional dose calculation method based on deformable tetrahedral mesh models, and 4D PET image reconstruction.
In the same context of treatment verification, the tetrahedral deformable mesh models can be used in order to simulate Prompt Gamma events and simulate the efects of variable breathing on Prompt Gamma imaging.

Management of selfish and malicious behavior in distributed collaborative systems

Hayam Mousa, Sonia Ben Mokhtar, Omar Hasan, Osama Younes, Mohiy Hadhoud, Lionel Brunie

3rd year of PhD, Funded by the Egyptian Government, DRIM Team
hayam.mousa@liris.cnrs.fr - http://liris.cnrs.fr/membres?idn=hmousa

Abstract

Participatory sensing is an emerging distributed collaborative paradigm in which citizens voluntarily use their mobile phones to capture and share sensed data from their surrounding environment in order to monitor and analyse some phenomena (e.g., weather, road traffic, pollution, etc.). Participating users are not usually remunerated for their participation. Therefore, they don't have strong motivations to comply with the tasks' requirements. Thus, they can disrupt the system by contributing corrupted, fabricated, or erroneous data. Different reputation systems have been proposed in the state-of-the-art to monitor the participants' behavior and to estimate their honesty. A few of these methods concern about the estimation of the quality of contributions. In addition, existing reputation systems lack the resistance to malicious colluding participants. Much more work still needed to have an efficient and robust reputation system for such applications. In our work, we propose a more robust and efficient reputation system designed for participatory sensing applications. The system can efficiently estimate the quality of participants' contributions and assign them trust scores. These scores enable for more accurate data aggregation. Furthermore, it supports the participants' accountability characteristic through assigning a reputation score to each participant. The system also incorporates a mechanism to defend against massive collusion.

Introduction

Everyday, billions of people move around the world carrying a variety of handheld devices equipped with sensing, computing, and networking capabilities (e.g., smartphones, tablets, GPS watches, smart gears, music players and in-vehicle sensors). The functionalities, performance, and widespread use of such devices have helped toward the emergence of a new kind of applications called participatory sensing. These applications exploit both the mobility of the participants and the sensing capabilities of their devices to construct opportunistic mobile sensor networks.

Threat Model

Participants control the sensing process and take over the responsibility to capture and report their observations to a backend server. However, no restrictions are usually imposed about the participants' experience, concern, trustworthiness, and interest. Participants are usually volunteers. So, the quality of each contribution depends on the participant's behavior and concern. Hence, these applications are vulnerable to a set of attacks which can disrupt the system measurements. We have surveyed classified studied and compared these attacks and the previously proposed solutions to defend against such attacks. We also have defined the major limitations of those solutions [1].

DTSRS: Reputation System

We propose a reputation system to estimate the trustworthiness of participants' contributions in participatory sensing applications.

The system proposes a novel method to accurately evaluate the quality of contributions. It also incorporates some parameters to assign a trust score to each contribution. These parameters include the users' feedback, the proximity of a participant's location to the sensing area, and the reputation score assigned of its provider. Reputation score of the participant is then updated based on his instantaneous trust.

Conclusion Experimental results indicate that our system outperforms the state-of-the-art since it can accurately estimate the quality of contributions even if there exist a massive data disruption. In addition, the system can tolerate up to 70% of adversaries participating in the sensing campaign compared to 40% in reputation systems in the state-of-the-art. Furthermore, it can detect adversaries even if they strategically contribute some good data with high probability (e.g. 0.8).

Bibliography

[1] Hayam Mousa, Sonia Ben Mokhtar, Omar Hasan, Osama Younes, Mohiy Hadhoud, Lionel Brunie, Trust management and reputation systems in mobile participatory sensing applications: A survey, Computer Networks, Elsevier July (2015), doi:10.1016/j.comnet.2015.07.011

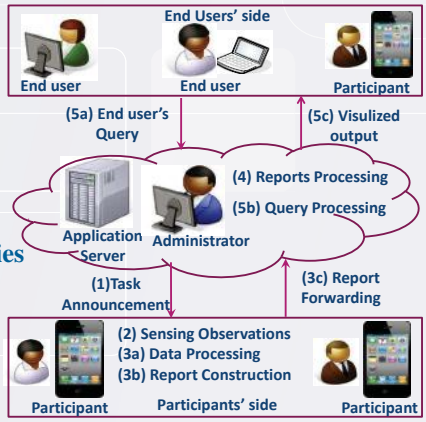
DTSRS: A **D**ynamic **T**rusted **S**et based **R**eputation **S**ystem Robust Against Malicious and Colluding Adversaries in Participatory Sensing Applications

Hayam Mousa, Sonia Ben Mokhtar, Omar Hasan, Osama Younes, Mohiy Hadhoud, Lionel Brunie

Laboratoire d'InfoRmatique en Image et Systèmes d'information
LIRIS UMR 5205 CNRS / INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / Ecole Centrale de Lyon

Participatory Sensing

- Applications
  - User Centric
  - Environment Centric
- Challenges
  - Attacks and Vulnerabilities
  - Trust and Reputation



Contribution

DTSRS: A novel reputation system resistant to the following attacks:

- Data Corruption
- Collusion of participants
- On-off attack

$$behavior = \begin{cases} Good & if\ R_{p_i} > \tau \\ Malicious & if\ R_{p_i} \leq \tau \end{cases}$$

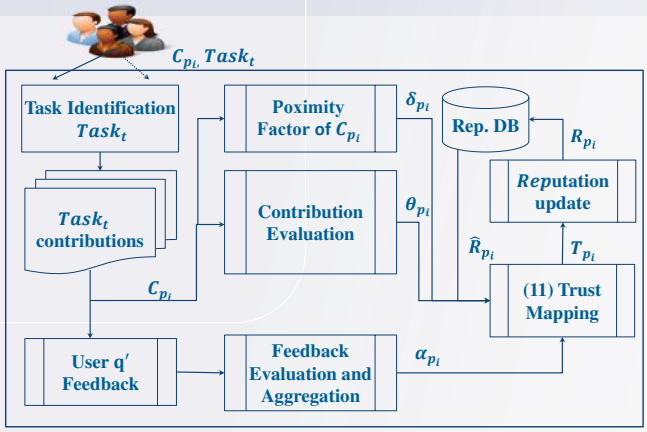
DTSRS reputation system

- Define the trusted set.
- Calculate the trust of a contribution

$$\mathbf{Trust}(C_{p_i}) = \mathbf{W1} \times \theta_{p_i} + \mathbf{W2} \times \alpha_{p_i} + \mathbf{W3} \times \delta_{p_i} + \mathbf{W1} \times \widehat{R}_{p_i}$$

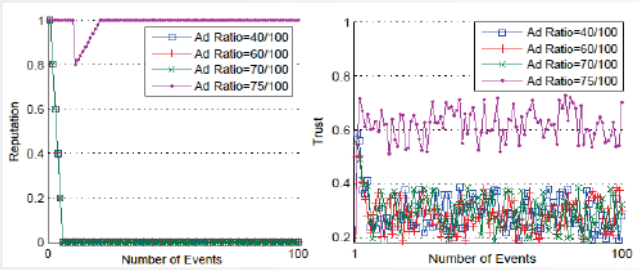
- Update the reputation of participants

$$R_{p_i} = \begin{cases} min\{\widehat{R}_{p_i} + \epsilon_r, 1\}\ if\ \theta_{p_i} > \tau \\ max\{\widehat{R}_{p_i} - \epsilon_p, 0\} if\ \theta_{p_i} < \tau \end{cases}$$

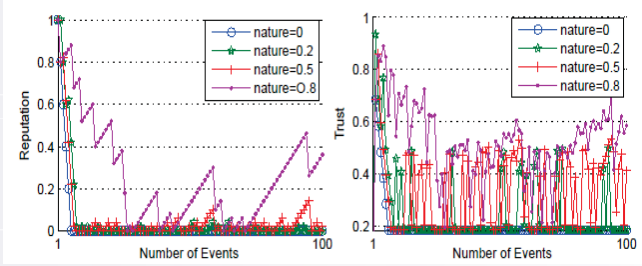


DTSRS strengths

- Efficient estimation of the quality of participants' contributions even if there exist a massive data disruption.
- Assigning accurate reputation score to each participant.
- Defence against massive collusion.
- Better performance compared with the state of the art reputation systems.
- Tolerate up to 70% of adversaries participating in the sensing campaign compared to 40% in systems from the state-of-the-art .
- Detection of adversaries even if they strategically contribute some good data with high probability (e.g. 0.8).



Impact of the adversary ratio on reputation and Trust of adversaries in the proposed DTSRS system.



Impact of the nature of the adversary on reputation and Trust of adversaries in the proposed DTSRS system.

References

[1] Hayam Mousa, Sonia Ben Mokhtar, Omar Hasan, Osama Younes, Mohiy Hadhoud, Lionel Brunie, Trust management and reputation systems in mobile participatory sensing applications: A survey, Elsevier, Computer Networks, July (2015), doi:10.1016/j.comnet.2015.07.011 .

[2] Hayam Mousa, Sonia Ben Mokhtar, Omar Hasan, Osama Younes, Mohiy Hadhoud, Lionel Brunie, DTSRS: A Dynamic Trusted Set based Reputation System Robust Against Malicious and Colluding Adversaries in Participatory Sensing Applications, Under revision

Étude de quelques paramètres de graphes : Décompositions et Dominations.

Fairouz BEGGAS, Hamamache KHEDDOUCI, Mohammed HADDAD

3ème année de thèse, Financement Gouvernement Algerien, Équipe GOAL
fairouz.beggas@liris.cnrs.fr – http://liris.cnrs.fr/membres?idn=fbeggas

Résumé de la thèse

Depuis leur apparition, les graphes sont considérés comme un outil de modélisation puissant dans une large variété de domaines scientifiques. Un graphe est une ensemble de sommets que des arêtes relient. Les problèmes modélisés par des graphes permet d'avoir une représentation assez facile pour les résoudre. Dans la première partie de cette thèse, on s'intéresse aux décompositions de graphes c'est à dire au fait de découper le graphe en sous-graphes disjoints.Ces sous-graphes entretiennent entre eux des relations particulières, dont l'exploitation permet de résoudre plus efficacement des problèmes. La deuxième partie de cette thèse touche un problème suscitant le plus d'attention actuellement qui est l'étude de la domination sur les graphes. Nous allons nous intéresser à une nouvelle variante de domination qui consiste à trouver un ensemble de sommets qui dominent un ensemble d'arrêtes qui ont une certaine particularité.

Problématique

Les deux problèmes majeures dans cette thèse est la décomposition d'arrêtes et la triangle domination d'arrêtes dans un graphe G. Le seul point en commun qui relie ces deux problèmes est le fait qu'il s'agit de paramètres à appliquer sur l'ensemble d'arrêtes d'un graphe G.

La décomposition

Si on a un problème difficile sur notre représentation du graphe, ce problème peut être résolu de manière plus simple avec une autre représentation ; l'idée n'est pas toujours de regarder le graphe différemment pour résoudre le problème plus vite, mais que l'on décompose le graphe en sous-graphes ce qui permet de résoudre le problème plus rapidement. Par exemple, si l'on divise un graphe en composantes connexes, de nombreux problèmes (routage, coloration, clique maximum...) peuvent être résolus (en parallèle) sur chaque composante, puis on fait l'union (ou le maximum, ou une opération plus complexe) des solutions pour obtenir la solution pour le graphe de départ. Il existe différents types de décompositions : simples ou multiples. Intuitivement, une décomposition en sous-graphes F de taille k permet de représenter le graphe d'origine G par un ensemble de copies du sous-graphe F, où chaque sommet du graphe G appartient à une seule copie du sous-graphe F, avec quelques contraintes. Notre étude touchera la multi-décomposition d'un multigraphe complet en cycles et étoiles.

Multigraphe complet On note λK_n un graphe complet doté de λ arêtes multiples et d'ordre n.

Le graphe Étoile Une étoile S_k est le graphe biparti complet $K_{1,k}$. On peut le définir comme un graphe connexe dont tous les sommets sauf un sont de degré 1.

Le graphe Cycle Le graphe cycle C_n est constitué d'un unique cycle de longueur n (pour $n \geq 3$). C'est un graphe connexe non-orienté d'ordre n à n arêtes. Il est 2-régulier.

La triangle domination

Le problème de domination consiste à trouver un ensemble de sommets (de taille minimum) dominant le reste des sommets d'un graphe. De nombreuses variantes d'intérêt à la fois théoriques et pratiques ont été proposées et ont été étudiés dans la littérature. Durant cette thèse, nous allons nous intéresser à une nouvelle variante de domination qui consiste à trouver un ensemble de sommets qui dominent l'ensemble des arrêtes d'un graphe tel que chaque arrête soit dominée par un sommet formant un triangle avec elle. L'essence de ce problème réside dans sa nature combinatoire ainsi que ses domaines d'application : réseaux de capteurs, etc. Notre graphe doit être triangulé.

Graphe triangulé Chaque arrête du graphe appartient à un triangle.

Références

[GD97] Dor, Dorit and Tarsi, Michael, Graph decomposition is NP-complete : A complete proof of Holyer's conjecture,*SIAM Journal on Computing, 1997.*

[ESM11] Dong, Dezun and Liao, Xiangke and Liu, Yunhao and Shen, Changxiang and Wang, Xinbing, Edge self-monitoring for wireless sensor networks,*Parallel and Distributed Systems, IEEE Transactions on, 2011.*

Étude de quelques paramètres de graphes : Décompositions et Dominations.

Fairouz BEGGAS, Hamamache KHEDDOUCI, Mohammed HADDAD

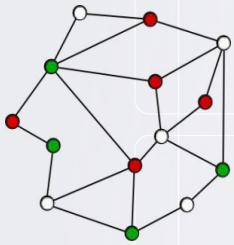
3 ème année de thèse, Financement Gouvernement Algérien, Équipe GOAL

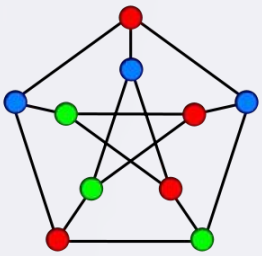
Laboratoire d'InfoRmatique en Image et Systèmes d'information

LIRIS UMR 5205 CNRS / INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / Ecole Centrale de Lyon

Résumé

- Depuis leur apparition, les graphes sont considérés comme un outil de modélisation puissant dans une large variété de domaines scientifiques. Un graphe est une ensemble de sommets que des arêtes relient.
- Les problèmes modélisés par des graphes permet d'avoir une représentation assez facile pour les résoudre. Dans la première partie de cette thèse, on s'intéresse aux décompositions de graphes c'est à dire au fait de découper le graphe en sous-graphes disjoints.
- Ces sous-graphes entretiennent entre eux des relations particulières, dont l'exploitation permet de résoudre plus efficacement des problèmes. La deuxième partie de cette thèse touche un problème suscitant le plus d'attention actuellement qui est l'étude de la domination sur les graphes.
- Nous allons nous intéresser à une nouvelle variante de domination qui consiste à trouver un ensemble de sommets qui dominent un ensemble d'arrêtes qui ont une certaine particularité.



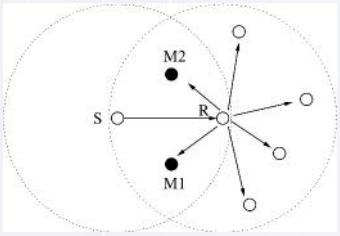


Problématique

- Les deux problèmes majeures dans cette thèse est la décomposition d'arrêtes et la triangle domination d'arrêtes dans un graphe G. Le seul point en commun qui relie ces deux problèmes est le fait qu'il s'agit de paramètres à appliquer sur l'ensemble d'arrêtes d'un graphe G.

La triangle domination

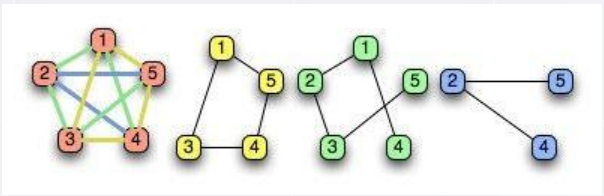
- Le problème de domination consiste à trouver un ensemble de sommets (de taille minimum) dominant le reste des sommets d'un graphe. De nombreuses variantes d'intérêt à la fois théoriques et pratiques ont été proposées et ont été étudiés dans la littérature.



- Durant cette thèse, nous allons nous intéresser à une nouvelle variante de domination qui consiste à trouver un ensemble de sommets qui dominent l'ensemble des arrêtes d'un graphe tel que chaque arrête soit dominée par un sommet formant un triangle avec elle. L'essence de ce problème réside dans sa nature combinatoire ainsi que ses domaines d'application: réseaux de capteurs, etc. Notre graphe doit être triangulé.
- Nous étudions différents points à savoir :
  - La complexité
  - Les bornes
  - Les classes de graphes
  - Généralisation sur les graphes pondérés
  - Etc,...

Décomposition

- Si on a un problème difficile sur notre représentation du graphe, ce problème peut être résolu de manière plus simple avec une autre représentation; l'idée n'est pas toujours de regarder le graphe différemment pour résoudre le problème plus vite, mais que l'on décompose le graphe en sous-graphes ce qui permet de résoudre le problème plus rapidement.
- Par exemple, si l'on divise un graphe en composantes connexes, de nombreux problèmes (routage, coloration, clique maximum...) peuvent être résolus (en parallèle) sur chaque composante, puis on fait l'union (ou le maximum, ou une opération plus complexe) des solutions pour obtenir la solution pour le graphe de départ.



Type de décomposition

- Il existe différents types de décompositions: simples ou multiples. Intuitivement, une décomposition en sous-graphes F de taille k permet de représenter le graphe d'origine G par un ensemble de copies du sous-graphe F , où chaque sommet du graphe G appartient à une seule copie du sous-graphe F , avec quelques contraintes. Notre étude touchera la multi-décomposition d'un multigraphe complet en cycles et étoiles.

Hand pose estimation by deep transductive learning

Natalia Neverova, Christian Wolf, Graham W. Taylor, Florian Nebout

4ème year of PhD, Funded by Financement *Investissements d'Avenir ("INTERABOT")*, ImagineTeam
natalia.neverova@liris.cnrs.fr – http://liris.cnrs.fr/natalia.neverova/

Abstract

The goal if this thesis is to develop cognitive and interaction capabilities for a mobile robot. To interact with its environment, a mobile robot needs to be capable to interpret data from its sensors, amongst which a camera is one of the most important ones. In this project, we focus on gesture and action recognition from videos taken from onboard cameras of small mobile robots. The first part of this project was dedicated to development of a multimodal deep learning framework performing gesture detection and classification based on video and depth streams, mocap data and audio recordings. In this work, we present the second part of the project, which considers exact hand pose estimation for fine grained gesture recognition and gesture parameter estimation.

Hand pose estimation and tracking from depth images, i.e. the estimation of joint positions of a human hand, is a first step for various applications: hand gesture recognition, human-computer interfaces (moving cursors and scrolling documents), human-object interaction in virtual reality settings and many more. While the estimation of full-body pose is now available at real-time in commercial products, at least in cooperative environments, the estimation of hand pose is more complex. In settings where the user is not directly placed in front of the computer, and therefore not close to the camera, the problem is inherently difficult. In this case, typically the hand occupies only a small portion of the image, and fingers and finger parts are only vaguely discernible.

We present a new method for the regression on depth based on semi-supervised learning using convolutional deep neural networks. An intermediate representation is first constructed based on a segmentation into parts. While recent works on body and hand pose estimation tend to perform direct regression from depth or color input to joint positions [1, 2], we argue that the intermediate representation is a powerful tool in the special context of semi-supervised learning.

In our setting, pose estimation is performed frame-by-frame without any dynamic information. A model is learned from two training sets: one containing labelled synthetic training images produced from 3D models by a rendering pipeline, and a second set containing unlabelled real images acquired with a consumer depth sensor. Our method does not rely on labels for the real data, and no explicit transfer function is defined or learned between synthetic and real data. Instead, a loss is defined on these data by extracting geometrical, structural and topological information related to a strong prior on the intermediate representation, i.e. on the segmentation of a hand into parts. The main arguments we develop are the following: – an intermediate representation defined as a segmentation into parts contains rich structural and topological information, since the label space itself is structured. Labels have adjacency, topological and geometric relationships, which can be leveraged and translated into loss for training; – a regression of joint positions is easier and more robust from a rich semantic representation like a segmentation into parts than from raw depth data, provided that this semantic segmentation is of high fidelity.

The prior knowledge we collect on unlabelled data is captured by several terms, some of which are defined on patches, and some of which are calculated on a full hand image: (i) contextual information is learned by an auto-context like model trained on the intermediate representation; (ii) a transductive learning method maps segmented patches from unlabelled real images to labelled patches from a very large set of patches rendered from synthetic data. The novelty here lies in the fact that we do not match input patches but patches taken from the intermediate representation, which include the to-be-inferred label and its local context; (iii) a global term evaluates topological properties in the input image similar to [3]. The combined information is fused into a semi-supervised training algorithm which minimizes empirical loss on labelled synthetic data and loss generated from structural terms calculated on real data.

Bibliography

[1] 1 Tompson, J., Stein, M., LeCun, Y., Perlin, K.: Real time continuous pose recovery of human hands using convolutional neural networks. In: SIGGRAPH (2014)

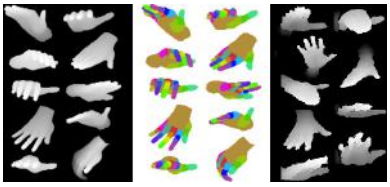
[2] Sun, M., Kohli, P., Shotton, J.: Conditional regression forests for human pose estimation. In: CVPR. (2012)

[3] Neverova, N., Wolf, C., Taylor ,G. ,Nebout, F.: Hand segmentation with structured convolutional learning. In: ACCV. (2014)

Hand segmentation with structured convolutional learning

Natalia Neverova*, Christian Wolf*, Graham Taylor**, Florian Nebout***
*Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
**University of Guelph, Canada ***Awabot
e-mail: natalia.neverova@liris.cnrs.fr

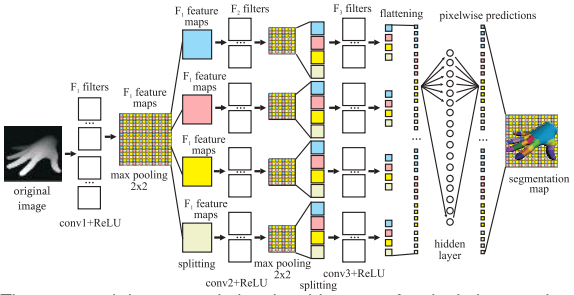
Motivation



Detailed hand pose estimation remains a challenge since fingers are often occluded and may only represent just a few pixels. Moreover, labelled data is difficult to obtain.
We propose a **deep learning based-approach for hand pose estimation**, targeting gesture recognition, that requires very **little labelled data**. It leverages both unlabelled data and synthetic data from renderings. The key to making it work is to integrate **structural information** not into the model architecture, which would slow down inference, but **into the training objective**. We show that adding unlabelled real-world samples significantly improves results compared to a purely supervised setting.

Deep architecture

- ◇ Two learning pathways: a **direct learner** f_d and a **context learner** f_c.
- ◇ The output of f_d is fed into f_c to **integrate the context** of the pixel.
- ◇ The **context learner** operates on **punctured neighborhood segmentation maps** where the middle pixel is missing.
- ◇ During **test time** only the **direct learner** is used.



The proposed deep convolutional architecture of a single learner that outputs full-resolution segmentation maps.

- ◇ **Training data**: **labelled synthetic** and **unabelled real** depth images.
- ◇ **Synthetic frames**: rendered using a **deformable 3D hand model** with large variety of view points and hand poses.
- ◇ The following **loss function** is used for learning:

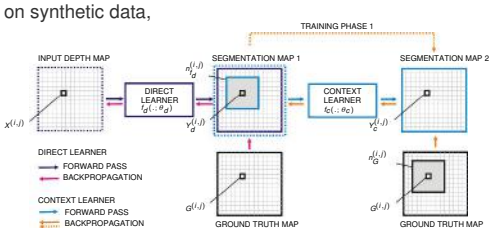
$$Q = Q_{sd} + Q_{sc} + Q_u,$$

Q_{sd} is responsible for **supervised** training of the **direct learner**,
Q_{sc} corresponds to the **context learner**,
Q_u – **unsupervised term** serving as a natural regulariser.

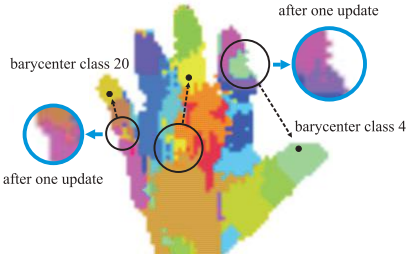
Semi-supervised structured learning

- ◇ The **unsupervised term** is defined as $Q_u = f(Q_{loc}, Q_{glb})$,

Q_{loc} – term capturing **local structure** by favoring local output patches which are consistent with a deep context model learned on synthetic data,



Q_{glb} – captures **global structure** favoring predictions with simple topological properties (one connected region per output label).



Both **local structure** and **global structure** are fused emphasising agreement between both unsupervised terms.

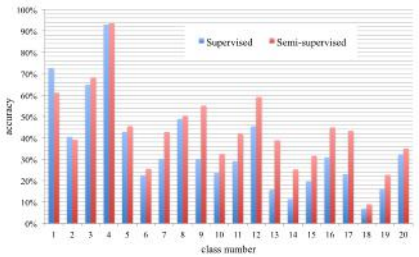
Experimental results

Performance of networks trained with different objective functions

Loss function	Training	Testing	Accuracy	Average per class
Q_{sd} (supervised baseline)	synth.	synth.	85.90%	78.50%
		real	47.15%	34.98%
$Q_{sd} + Q_{loc} + Q_{glb}$ (semi-supervised, ours)	all	synth.	85.49%	78.31%
		real	**50.50%**	**43.25%**

Performance improvement on a real image after updating parameters using different supervised and unsupervised terms, estimated as an average over 50 real images.

Terms	Q_{loc}	Q_{glb}^+	$Q_{glb}^+ + Q_{glb}^-$	$Q_{loc} + Q_{glb}^+ + Q_{glb}^-$	Q_{sd}
Requires labels	no	no	no	no	yes
Gain in % points	+0.60	+0.36	+0.41	**+0.82**	+16.05



Average accuracy per class obtained with the supervised method (in blue) and with semi-supervised structured learning (in red)

Challenging examples: output segmentation maps produced by the semi-supervised network for real-world images. In each pair of images the first one corresponds to the output of the baseline method, the second ones are results of our proposed algorithm on the same data.

Détection, localisation et typage de texte dans des images de documents hétérogènes par réseaux de neurones profonds

Bastien Moysset, Christian Wolf, Jérôme Louradour

2ème année de thèse, Financement Industriel, Équipe Imagine
bastien.moysset@liris.cnrs.fr – http://liris.cnrs.fr/membres?idn=bmoysset

Résumé de la thèse

Les systèmes de reconnaissance de l'écriture et d'extraction d'information dans les images de documents, développés depuis une vingtaine d'années, reposent sur des étapes préliminaires d'analyse de structure de documents. Le but étant d'extraire automatiquement le texte présent dans une image. Cette thèse mettra en place de nouvelles méthodes de localisation de lignes de texte suffisamment robustes pour pouvoir traiter des documents difficiles et hétérogènes. Pour cela, nous nous intéresserons aux Réseaux de Neurones Artificiels profonds. Ces types de modèles ont été tout récemment appliqués avec succès à des problèmes de détection et localisation d'objets dans des scènes naturelles. Le verrou scientifique principal concerne l'intégration de contexte dans les modèles actuels. De manière générale, il s'agit de modéliser les dépendances spatiales entre entités sémantiques dans une image (pixels, super-pixels, régions, objets, parties d'un objet etc.). Dans le contexte spécifique étudié dans ce projet, il s'agit de modéliser les dépendances entre les entités d'un document : caractères, mots, lignes, blocs de texte, blocs de graphisme.

Segmentation de Paragraphes

Dans un premier temps[1], nous avons utilisé des réseaux de neurones récurents pour segmenter les paragraphes en ligne en utilisant une architecture et une fonction de coût similaires à celles utilisées pour la reconnaissance de l'écriture ; c'est à dire en utilisant la Connectionist Temporal Classification (CTC) pour l'alignement des séquences. Cette technique ne nécessite pas l'annotation des positions des lignes, seule la connaissance du nombre de lignes dans le paragraphe est nécessaire. La limite de cette méthode est le caractère uniquement vertical de cette segmentation. De ce fait, elle ne peut pas être appliquée à des pages complètes à structures complexes, ni à des paragraphes penchés.

Segmentation de pages complètes

Nous nous sommes ensuites penchés sur la possibilité de détecter les lignes dans des pages complètes. Les difficultés principales auxquelles nous cherchons à répondre sont :

- La nécessité de pouvoir détecter un nombre variable d'objets.
- La nécessité de pouvoir détecter des objets qui se superposent.

Pour éviter des étapes de post-traitements basées sur des heuristiques, nous avons utilisé l'algorithme présenté par Ehran et al.[2] qui trouve directement les coordonnées des boîtes en utilisant un réseau de neurones comme régresseur, et qui remplit les deux conditions sus-cités.

Néanmoins, cet algorithme peine à généraliser correctement, malgré l'utilisation de data augmentation et de régularisation (dropout nottamment) lors de l'entrainement des réseaux. Nous pensons que cela est dû au nombre important de paramètres sur la dernière couche du réseau. Pour résoudre ce problème, nous avons apporté une nouvelle couche nommée Space Displacement Localization (SDL) pour permettre de partager les paramètres entre les différentes zones de l'image, permettant un gain significatif en performance sur la tâche de la compétition ANDAR[3].

Future work

La couche SDL ne permet en l'état que de détecter des points. Le travail à venir consistera à évaluer quelle est la meilleure méthode pour détecter des boîtes. Les options principales sont l'utilisation de deux réseaux pour détecter respectivement les points à droite et à gauche de la ligne puis associer ceux-ci entre eux ou la modification de la couche SDL pour qu'elle puisse prédire des boîtes. La suite du travail consistera à modéliser l'interraction que peuvent avoir entre elles les différentes lignes du document et se servir de ce contexte pour améliorer la détection.

Références

[1] Paragraph text segmentation into lines with Recurrent Neural Networks, B Moysset, C Kermorvant, C Wolf, J Louradour - ICDAR 2015

[2] Scalable object detection using deep neural networks, D Erhan, C Szegedy, A Toshev, D Anguelov - CVPR 2014

[3] Space Displacement Localization Neural Networks to locate origin points of handwritten text lines in historical documents, B Moysset, P Adam, C Wolf, J Louradour - HIP 2015

DETECTION, LOCALIZATION AND TYPING OF TEXT IN HETEROGENEOUS DOCUMENT IMAGES WITH DEEP NEURAL NETWORKS

Bastien Moysset

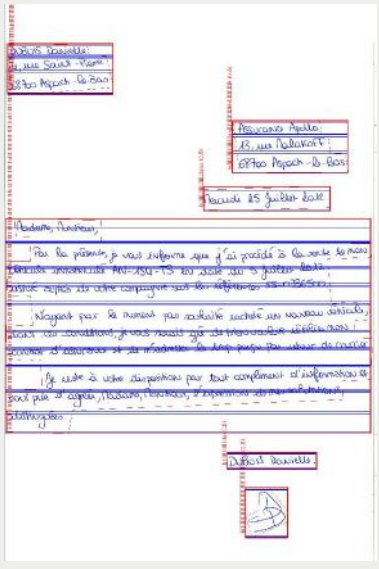
Motivations

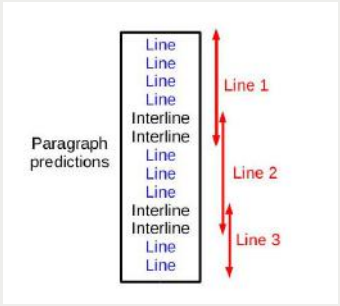
- Text recognition and information extraction need good detection and localisation of the text.
- Use of Machine Learning techniques helps to get higher level features to represent the text, thus to deal with highly heterogeneous databases.
- Techniques will have to be able to detect a variable number of objects and to deal with bounding boxes that overlap.

Heterogeneous database : the Maurdor example

- 8774 pages with 192536 zones, 3 languages, handwritten and printed.

Paragraph segmentation into lines with RNNs





- Use of a 2D-LSTM architecture similar to text recognition.
- Use of Connectionist Temporal Classification (CTC) to align sequences.
- Two labels: line and interline.
- Main advantage: No need for the location of lines in annotation.
- Main disadvantages: Works only at paragraph level and has problem with skewed texts.

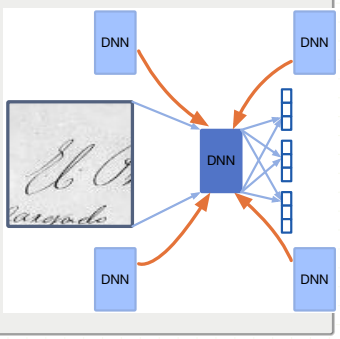
The Space Displacement Localization layer for position prediction.

- Inspired by Ehran et al.
- The neural network is used to predict point position hypothesis (regression).
- For each point, three values are predicted (X-value, Y-Value, Confidence).
- The cost function is:

$$E(X,\theta) = \alpha \sum_{ij} X_{ij} ||o_i(\theta) - g_j||^p + \sum_i X_{ij} \log\left(\frac{c_i(\theta)}{1 - c_i(\theta)}\right)$$

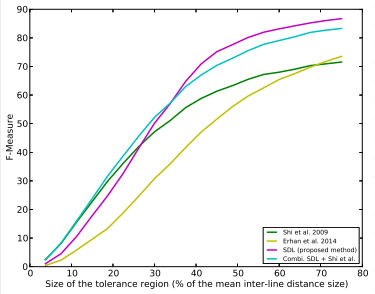
$$\text{s.t. } \forall j, \sum_i X_{ij} \leq 1 \quad \wedge \quad \forall i, \sum_j X_{ij} \leq 1$$

- SDL enables to share weights between the different localisations of the image.
- A 2D-LSTM recurrence is added to take context dependencies into account.



The Andar competition: prediction of the beginning of lines

- First application of the SDL, localisation of the line beginnings in historical documents.
- Results comparable to image-processing based techniques.



Perspectives and future work

- Use the SDL layer to detect boxes instead of points. Main options are :
  - Trust the LSTM to convey information about the width and the height of the lines.
  - Detect left and right of the text line separately and match them.
- Modelisation of the interraction between the different lines of the document (with graphical probabilistic models, or deformable part models).
- Study the invariance between different databases. Can we train some generic network or do we need to adapt the first layers to the new images?
- Tune the architecture of the RNN, especially the filter sizes and the number of layers, in order to allow more complex internal representations.

References

- D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014.
- B. Moysset, C. Kermorvant, C. Wolf, and J. Louradour, "Paragraph text segmentation into lines with recurrent neural networks," in *International Conference on Document Analysis and Recognition*, 2015.
- B. Moysset, P. Adam, C. Wolf, and J. Louradour, "Space displacement localization neural networks to locate origin points of handwritten text lines in historical documents," in *ICDAR 2015 Workshop on Historical Document Imaging and Processing*, 2015.

bm@a2ia.com

Réalité Augmentée sur mobile en contexte urbain

M.Ayadi, M.Scuturici, S.Miguet

2ème année de thèse, Financement Erasmus Mundus, Equipe IMAGINE
Mehdi.Ayadi@univ-lyon2.fr - https://liris.cnrs.fr/membres?idn=mayadi

Résumé de la thèse

Cette thèse s'insère dans le cadre de collaborations interdisciplinaires entre informaticiens, géographes et urbanistes. Ces collaborations ont permis l'élaboration du projet ANR « Skyline » et d'un projet du « LabEx IMU » sur la mesure géométrique du Skyline. Ces interactions ont permis de faire émerger le besoin d'outils et de techniques spécifiques, susceptibles d'avoir des retombées pour l'ensemble des communautés concernées. Nous proposons d'étudier les approches permettant à un utilisateur se déplaçant en milieu urbain, de visualiser, à l'aide d'un Smartphone ou d'une tablette, l'incidence d'un projet de construction sur le paysage urbain, une fois le bâtiment réalisé. Il ne s'agit pas de proposer un rendu photo-réaliste, mais plutôt une esquisse la plus exacte possible de la géométrie de la scène. La plupart des applications de Réalité Augmentée (RA) se basent aujourd'hui soit sur les instruments seulement, ou sur l'image pour incruster l'objet de synthèse. Nous proposons une approche dans laquelle des points d'amers sont extraits automatiquement des images et suivis tout au long de la séquence, permettant d'ancrer les objets de synthèse à ces amers visuels, et ainsi de donner à l'observateur une impression d'immersion bien meilleure. Ces amers visuels ne sont autres que les points définissant le *Skyline*.

Contexte

Avec le retour actuel des projets de constructions de tours dans les grandes villes (Londres, Paris et Lyon notamment), il devient un enjeu très important pour les architectes et urbanistes, mais également pour les usagers des quartiers, de pouvoir se rendre compte visuellement de l'impact d'un projet de construction sur la ligne d'horizon. Le projet Skyline est la base de nos travaux, et les résultats obtenus sont notre point de départ.

Objectifs

L'objectif de ce sujet est de proposer une approche, basée sur les technologies de Réalité Augmentée mobile, permettant d'incruster un objet de Synthèse en 3D, en l'occurrence un bâtiment, dans le scène réelle, et ce en se basant sur un marqueur très particulier : Le Skyline.

L'hypothèse principale que nous voulons valider :

- Est-ce que le Skyline peut jouer le rôle d'un marqueur dans une application de Réalité Augmentée en contexte urbain ?

Extraction du Skyline

Nous nous basons sur les résultats d'un contrat Post Doctoral effectué au LIRIS, nous permettent d'obtenir un Skyline s'adaptant à différents niveaux de profondeurs.

- Publication : En cours de finalisation d'un article : [A parametric approach for Skyline Extraction].

Skyline d'arrière plan

Skyline intermédiaire artificiel

➔ IHM + Extraction paramétrique du Skyline (Démonstration 1)

État de l'Art

Dans la littérature, deux principales approches sont proposées :

- App de RA mobile qui se contentent d'utiliser les instruments embarqués dans les Smartphones :
  - GPS + compas magnétique ➔ Génération d'une vue de synthèse
  - Gyroscope + Accéléromètre ➔ Evaluation grossière des paramètres de mouvement .[1]

  Inconvénient : Impression très peu réaliste de la scène (les objets de synthèse semblent « flotter » au gré du mouvement). (Démonstration 2)
- App de RA se basant sur l'image, à l'aide de Marqueurs spécifiques.
  Inconvénient Techniques contraignantes car besoin de marqueur spécifiques. (Démonstration 3).

Approche proposée

La pose est, en premier lieu, estimée en utilisant les instruments embarqués dans le Smartphone, qui sera ensuite corrigée grâce aux informations récupérées d'un dernier Capteur : La Caméra. En effet, certaines informations de l'image, et plus précisément les points du Skyline permettent de récupérer la pose de l'utilisateur. Ceci est expliqué ci-après :

1. Extraction d'un Skyline réel (Etape finalisée)
2. Instruments +modèle 3D (GrandLyon openData) ➔ Skyline théorique (en cours de finalisation)
3. Matching ➔ Correction de la pose (étape non encore entamée)

- Publication : En cours d'écriture d'un article [Survey of Augmented Reality in urban context].

Références

[1] T. Fukuda, T. Zhang, and N. Yabuki, "Improvement of registration accuracy of a handheld augmented reality system for urban landscape simulation," Front. Archit. Res., vol. 3, no. 4, pp. 386–397, 2014.

Réalité Augmentée sur mobile en contexte urbain

M. Ayadi

Sous l'encadrement de M. Scuturici et S. Miguet

2ème année de thèse, Financement Erasmus Mundus, Equipe IMAGINE
Mehdi.Ayadi@univ-lyon2.fr - https://liris.cnrs.fr/membres?idn=mayadi

Résumé de la thèse

Nous proposons d'étudier les approches permettant à un utilisateur se déplaçant en milieu urbain, de visualiser, à l'aide d'un Smartphone ou d'une tablette, l'incidence d'un projet de construction sur le paysage urbain, une fois le bâtiment réalisé. Il ne s'agit pas de proposer un rendu photo-réaliste, mais plutôt une esquisse la plus exacte possible de la géométrie de la scène. La plupart des applications de Réalité Augmentée (RA) se basent aujourd'hui soit sur les instruments seulement (GPS, accéléromètre, …), soit sur l'image (utilisation de marqueurs) pour incruster l'objet de synthèse. Nous proposons une approche dans laquelle des points d'amers sont extraits automatiquement des images et suivis tout au long de la séquence, permettant d'ancrer les objets de synthèse à ces amers visuels, et ainsi de donner à l'observateur une impression d'immersion bien meilleure. Nous voulons valider l'hypothèse selon laquelle des points caractéristiques de la ligne d'horizon (Skyline) peuvent jouer le rôle de marqueurs.

Contexte

Quel est l'impact d'un projet de construction sur la ligne d'horizon (Skyline) ?

Objectifs

L'objectif de nos travaux est de proposer une approche, basée sur les technologies de Réalité Augmentée mobile, permettant d'incruster un objet de Synthèse en 3D, en l'occurrence un bâtiment, dans la scène réelle, en se basant sur le Skyline.

L'hypothèse principale que nous voulons valider :

- Le Skyline peut-il jouer le rôle d'un marqueur dans une application de Réalité Augmentée en contexte urbain ?

Extraction du Skyline

Résultats d'un contrat Post-Doctoral au LIRIS :

- Extraction d'un Skyline à différents niveaux de profondeur.

Figure 1 Skyline d'arrière plan (Naturel)

Figure 2 Skyline intermédiaire (Artificiel)

→ IHM : Extraction paramétrique du Skyline.

Etat de l'Art (RA)

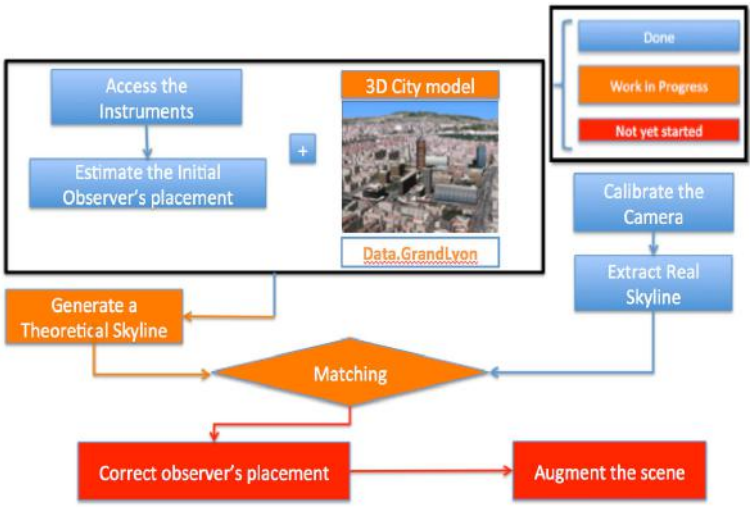
Deux principales approches :

- App de RA mobile qui se contentent d'utiliser les instruments embarqués dans les Smartphones :
  - ✓ GPS + compas magnétique
    ➔ Génération d'une vue de synthèse
  - ✓ Gyroscope + Accéléromètre
    ➔Evaluation grossière des paramètres de mouvement.

Inconvénient : Impression très peu réaliste de la scène (les objets de synthèse semblent « flotter » au wgré du mouvement) [1]

- App de RA se basant sur l'image (marqueurs) Inconvénient: Techniques contraignantes car besoin de marqueur spécifiques.

Approche proposée



Références

[1] *T. Fukuda, T. Zhang, and N. Yabuki, "Improvement of registration accuracy of a handheld augmented reality system for urban landscape simulation," Front. Archit. Res., vol. 3, no. 4, pp. 386–397, 2014.*

Multi-label learning with Ensemble paradigm

Ouadie GHARROUDI, Haytham ELGHAZEL, Alexandre AUSSEM

3ème year of PhD, Funded by FUI, DM2LTeam
ouadie.gharroudi@liris.cnrs.fr – http://liris.cnrs.fr/membres?idn=ogharrou

Abstract

Ensemble methods have been called the most influential development in Data Mining and Machine Learning in the last decade. Indeed, ensemble learning which combines multiple models to jointly accomplish one common task is shown to be very beneficial for enhancing the generalization ability of a single classifier. In the multi-label context, less attention has been given to this family of model. Although different ensemble models have been proposed for multi-label classification, no in-depth research has been addressed to analyse when ensemble models can be helpful in multi-label context and how they can address the feature selection issue. Our objectives in this thesis are to understand how the ensemble construction and aggregation steps affect the classification performance and to show how effective can be the ensemble models to handle the feature selection task in this specific classification.

Multi-label Learning

Multi-label classification is a challenging problem that emerges in several modern applications such as text categorization, gene function classification, and semantic annotation of images[4]. It studies the problem where each data sample is associated with a set of labels. From a computational perspective, the aim of multi-label classification is to obtain simultaneously a collection of binary classifications; where positive classes refer the relevant labels of the instances in the label space $\mathcal{L} = \{\lambda_1, \ldots, \lambda_Q\}$.

In multi-label context, ensemble methods are developed on top of the common problem transformation or algorithm adaptation methods [4]. The improvement of performances with this family of methods relies on the concept of diversity which states that a good ensemble model is an ensemble in which misclassified instances are different from one base-classifier to another. In multi-label classification, different strategies are used to build a group of diversified base-classifiers e.g., sub-resampling training data, feature subsets selection, random selection of labels, etc.

Ensemble multi-label models vote

An important step for the classification of unseen instances with the ensemble model is the aggregation of the base-classifiers predictions. In multi-label context several aggregation schemes are possibles (label 0/1 vote, probability distribution vote). Moreover, such aggregation schemes can only output prediction scores. Sometimes these label scores are insufficient and it is desirable to obtain a specific predicted label set. In this case, the predicted scores require to undergo a thresholding procedure $t(\cdot)$ to implement a decision function. For this purpose we conduct an extensive study of the influence of different voting and thresholding strategies on the final performance of ensemble multi-label models predictions [3] .

Calibrating ensemble model predictions

The base-classifier diversity is an important part of ensemble models that influences the model generalization ability and must be respected in both learning and prediction step (aggregation step). From this perspective, we propose three practical steps to learn and predict with ensemble k-labelsets models : One is to increase the diversity of the base classifiers in the ensemble, the second to smooth the label powerset probability estimates during the ensemble aggregation process, and the third to calibrate the label decision thresholds. The tree steps are gathered in a new ensemble models named Calibrated k-labelsets for Ensemble Multi-Label Classification (CkMLC) [1].

Feature selection with ensemble multi-label models

As an important part of the learning process, feature selection in multi-label attracted amount of research studies. However, few research were focused on the feature selection using ensemble models. In order to fill this gap we proposed two ensemble frameworks for estimating a feature importance with Random forest in multi-label context. We carried out an extensive comparison between the proposed frameworks and state-of-the-art multi-label feature selection algorithms in [2].

References

[1] Ouadie Gharroudi, Haytham Elghazel, and Alex Aussem. Calibrated k-labelsets for ensemble multi-label classification. In *ICONIP 2015 Proceedings (to appear)*.

[2] Ouadie Gharroudi, Haytham Elghazel, and Alex Aussem. A comparison of multi-label feature selection methods using the random forest paradigm. In *CAAI-2014. Proceedings*.

[3] Ouadie Gharroudi, Haytham Elghazel, and Alex Aussem. Ensemble multi-label classification: A comparative study on threshold selection and voting methods. In *ICTAI 2015 Proceedings (to appear)*.

[4] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 2012.

Calibrated k-labelsets for Ensemble Multi-Label Classification

Ouadie GHARROUDI , Haytham ELGHAZEL and Alex AUSSEM

Laboratoire d'InfoRmatique en Image et Systèmes d'information
LIRIS UMR 5205 CNRS / INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / Ecole Centrale de Lyon

Abstract:

RAndom k-labELsets (RAkEL) is an effective ensemble multilabel classification (MLC) model where each base-classifier is trained on a small random subset of k labels. However, the model construction does not fully benefit from the diversity of the ensemble and the label probability estimates obtained with RAkEL are usually badly calibrated due to the problems raised by the imbalanced label representation. In this paper, we propose three practical solutions to overcome these drawbacks. One is to increase the diversity of the base classifiers in the ensemble. The the second to smooth the label powerset probability estimates during the ensemble aggregation process, and the third to calibrate the label decision thresholds. Experimental results on various benchmark data sets indicate that the proposed approach outperforms significantly recent state-of-the-art MLC algorithms, including RAkEL and its variants.

Multi-label Learning:

	X_1	X_2	...	X_D	λ_1	λ_2	...	λ_Q
E_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,D}$	$\lambda_{1,1}$	$\lambda_{1,2}$	...	$\lambda_{1,Q}$
E_1	$x_{2,1}$	$x_{2,2}$	...	$x_{2,D}$	$\lambda_{1,1}$	$\lambda_{1,2}$	...	$\lambda_{1,Q}$
⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
E_N	$x_{N,1}$	$x_{N,2}$	...	$x_{N,D}$	$\lambda_{N,1}$	$\lambda_{N,2}$	...	$\lambda_{N,Q}$

Table 1. Multi-label data set

$$f : \mathcal{X} \longrightarrow \{0,1\}^Q$$

Calibrated k-labelsets for Ensemble Multi-Label Classification (CkMLC)

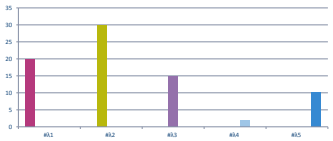
A- Label Powerset Construction

$$N = 10 \times ceil(\frac{\log(\alpha)}{\log(1-1/k)})$$

B- Base-classifier Aggregation

Selected Labelsets	Labels				
	λ_1	λ_2	λ_3	λ_4	λ_5
{λ2; λ3; λ5}		x	x		X
...	...	...	...	..	..
{λ1; λ3; λ4}	x		x	x	
# of selection	20	30	15	2	10

The random selection of the Labelsets leads to an imbalanced label representation in the ensembe.



$$\hat{P}_{\lambda_i} = \frac{h(x_t,\lambda_i)+1}{n+C}$$

C- The proposed Forward Threshold Calibration heuristique

Predicted labels probabilties		True label	
$P(\lambda_1=1/X)$	$P(\lambda_2=1/X)$	λ_1	λ_2
0.30	0.05	0	0
0.40	0.10	0	0
0.55	0.15	0	0
0.65	0.25	0	1
0.71	0.30	0	1
0.8	0.35	1	1
0.9	0.45	1	1

Algorithm 1 *Forward Multi-label Thresholds Calibration*

Require: *Oob* predictions probabilities $(\hat{Y})$, *Oob* real labels (Y), label set $\mathcal{L}$, multi-label performance measure to optimize $(MLmeasure)$. $=0$

$\mathcal{L}^* \leftarrow \emptyset$; $T \leftarrow \emptyset$
while $\mathcal{L} \neq \emptyset$ **do**
  $\lambda^*, t_{\lambda^*} \leftarrow \underset{\lambda \in \mathcal{L}, t \in [0,1]}{\text{argmax}} \; MLmeasure([\hat{Y}_{\mathcal{L}^*}/T \cup \{\hat{Y}_\lambda/t\}], [Y_{\mathcal{L}^*} \cup \{Y\}])$
  $\mathcal{L}^* \leftarrow \mathcal{L}^* \cup \lambda^*$
  $T \leftarrow T \cup t_{\lambda^*}$
  $\mathcal{L} \leftarrow \mathcal{L} \backslash \lambda^*$
end while
return T

Experimental evaluation:

Performance analysis of CkMLC against several sate-of-art multi-labe classifiation models

Table 3. Predictive performances in terms of $MicroF_1$.

↑	CkMLC^m	RAkEL	RAkEL++^{m}	CkMLC$^{0.5}$	TREMLC	FBR^m	ECC-RF
Arts	.650±.123	.620±.144•	.599±.149•	.539±.106•	.528±.107•	.508±.107•	.521±.091•
Birds	.560±.080	.498±.079•	.418±.110•	.476±.068•	.420±.056•	.470±.074•	.472±.057•
Business	.807±.057	.799±.044•	.796±.061•	.801±.043	.796±.045•	.757±.055•	.779±.040•
Education	.655±.119	.637±.144•	.585±.153•	.585±.127•	.568±.130•	.534±.106•	.545±.098•
Emotions	.798±.062	.728±.081•	.742±.080•	.778±.071•	.699±.078•	.727±.085•	.726±.063•
Enron	.716±.059	.698±.075•	.720±.083	.670±.060•	.654±.059•	.631±.071•	.678±.060•
Flags	.809±.043	.787±.039•	.804±.037	.806±.027	.764±.036•	.778±.044•	.773±.030•
Health	.769±.072	.755±.080•	.751±.087•	.736±.070•	.732±.068•	.674±.077•	.715±.057•
Image	.789±.070	.707±.095•	.707±.105•	.745±.091•	.681±.081•	.673±.090•	.685±.070•
Medical	.871±.031	.857±.034•	.862±.035•	.859±.032•	.860±.035•	.853±.032•	.853±.025•
Scene	.840±.059	.769±.077•	.773±.084•	.803±.072•	.746±.070•	.725±.077•	.752±.058•
Slashdot	.818±.021	.816±.030	.817±.029	.815±.019•	.816±.017	.814±.030	.807±.016•
(win/tie/loss)		(11/1/0)	(9/3/0)	(10/2/0)	(11/1/0)	(11/1/0)	(12/0/0)

•/◦ CkMLC^m is significantly better/worse, at level of significance of 5%.

Table 4. Predictive performances in terms of $MacroF_1$.

↑	CkMLC^M	RAkEL	RAkEL++^{M}	CkMLC$^{0.5}$	TREMLC	FBR^M	ECC-RF
Arts	.405±.131	.429±.136	.402±.140	.299±.075•	.282±.073•	.398±.114	.326±.079•
Birds	.384±.097	.324±.082•	.253±.098•	.276±.056•	.220±.035•	.336±.067•	.294±.049•
Business	.385±.125	.295±.098•	.329±.126•	.271±.079•	.245±.075•	.375±.110	.278±.073•
Education	.349±.117	.360±.123	.294±.109•	.256±.081•	.226±.068•	.316±.105•	.249±.071•
Emotions	.784±.078	.721±.083•	.736±.082•	.767±.078•	.689±.078•	.721±.085•	.715±.069•
Enron	.382±.135	.279±.076•	.386±.133	.204±.044•	.184±.032•	.320±.087•	.257±.066•
Flags	.738±.054	.724±.050	.757±.046	.742±.058	.664±.048•	.734±.052	.713±.052
Health	.440±.092	.439±.099	.445±.109	.352±.060•	.333±.048•	.416±.091•	.351±.059•
Image	.786±.074	.707±.093•	.707±.104•	.746±.090•	.682±.081•	.674±.089•	.687±.069•
Medical	.541±.091	.524±.083•	.525±.091•	.511±.086•	.481±.090•	.524±.083•	.430±.070•
Scene	.842±.058	.773±.076•	.778±.084•	.805±.074•	.748±.070•	.730±.075•	.756±.058•
Slashdot	.294±.093	.205±.063•	.263±.081•	.151±.029•	.127±.018•	.284±.088	.145±.026•
(win/tie/loss)		(8/4/0)	(9/3/0)	(11/1/0)	(12/0/0)	(8/4/0)	(11/1/0)

•/◦ CkMLC^M is significantly better/worse, at level of significance of 5%.

Conclusion :

In this paper, we discussed a novel strategy to build and aggregate k-labelsets in the context of ensemble multi-label classification. The proposed strategy extends and improves upon the original RAkEL algorithm in three ways: i) new randomization strategy using bagging in tandem with random klabelsets; ii) accounting for the imbalanced label representation when aggregating the base-classifiers predictions; and iii), a specific label threshold calibration procedure on out-of-bag instances. Experimental results on twelve benchmark data sets indicate that the proposed model outperforms the RAkEL algorithm and other recent state-of-the-art MLC algorithms. Future works will be conducted to analyze the thresholding strategy on different ensemble MLC approaches and to adapt, in a more principled way, the aggregation procedure to the specific loss function.

Statistical Learning under Selection Bias

Van Tinh TRAN, Alex AUSSEM

3rd year of PhD, Funded by INTEGRATE, DM2LTeam
van-tinh.tran@liris.cnrs.fr – http://liris.cnrs.fr/membres?idn=ttran

Abstract

Selection bias, which occurs when training and test joint distributions are different, i.e. $P_{tr}(x,y) \neq P_{te}(x,y)$, is pervasive in almost all empirical studies, including Machine Learning. However, when $P_{tr}(x,y)$ and $P_{te}(x,y)$ differ only in $P_{tr}(x)$ and $P_{te}(x)$ (known as *covariate shift*) or only in $P_{tr}(y)$ and $P_{te}(y)$ (known as *prior probability shift*), adaptation methods using importance weight were proven to be effective. In this thesis, we first present a general framework to correct the more general class of selection bias problem which includes covariate shift and prior probability shift as special cases. We then show that the method of importance weighting applied to covariate shift is sub-optimal and discuss a manner to optimally combine the important-weighted and the unweighted models in order to improve the predictive performance in the target domain. Both of these methods are supported by theoretical and experimental result.

Background

Selection bias, which occurs when training and test joint distributions are different, i.e. $P_{tr}(x,y) \neq P_{te}(x,y)$, is pervasive in almost all empirical studies, including Machine Learning, Statistics, Social Sciences, Bioinformatics, Epidemiology, Medicine, etc. It is therefore highly desirable to devise algorithms that remain effective under such distribution shifts. It is well known that one may account for the difference between $P_{tr}(x,y)$ and $P_{te}(x,y)$ by re-weighting the training points using the so-called importance weight, denoted as $\beta(x,y) = \frac{P_{te}(x,y)}{P_{tr}(x,y)}$. In general, the estimation problem with two different distributions $P_{tr}(x,y)$ and $P_{te}(x,y)$ is unsolvable, as the two terms could be arbitrarily far apart. However, when $P_{tr}(x,y)$ and $P_{te}(x,y)$ differ only in $P_{tr}(x)$ and $P_{te}(x)$ (known as *covariate shift*), the importance weight is reduced to $\beta(x) = P_{te}(x)/P_{tr}(x)$, which requires unlabeled examples $(P(x))$ to estimate. Similarly, when $P_{tr}(x,y)$ and $P_{te}(x,y)$ differ only in $P_{tr}(y)$ and $P_{te}(y)$ (known as *prior probability shift*), the importance weight is reduced to $\beta(y) = P_{te}(y)/P_{tr}(y)$, which requires a sample of label population $(P(y))$ to estimate.

Generalize covariate shift and prior probability shift

In [1], we show that, if we have a combination of biased data and unbiased data and qualitative probabilistic assumptions that are deemed plausible about our sampling mechanism, our problem becomes solvable. More specifically, we assume we have access to a S-control feature vector, X_s, and some additional sample of the form (x_s) that is drawn from the population as a whole, such that S is conditionally independent of (X,Y) given X_s. Despite being limited to specific or idealized situations, this framework includes covariate shift and prior probability shift as special cases. We also consider the case where X_s is not fully measured in the target population. This situation typically arises in various clinical studies, where some variables are too difficult or costly to measure in the target population.

Hybrid approach to covariate shift

Importance weighting in general is know to come at the expense of a reduction of the effective sample size. We show analytically in [2] that, while the unweighted model is globally more biased than the weighted one, it may locally be less biased on low importance instances. Our approach includes four steps:

- Define the metric to measure the global and local bias based on test distribution.
- Define alternative metric to measure the global and local bias based on training distribution, which is observed in training data.
- Establish the relationship between the global bias metric estimated on train distribution and local bias metrics estimated on test distribution.
- Show that on some local subsets of input feature space, the unweighted model is less biased than the weighted model.

References

[1] Van-Tinh Tran and Alex Aussem. Correcting a class of complete selection biaswith external data based on importance weight estimation. In *ICONIP*, 2015. to appear.

[2] Van-Tinh Tran and Alex Aussem. A practical approach to reduce the learning bias under covariate shift. In *Machine Learning and Knowledge Discovery in Databases*, pages 71–86. Springer, 2015.

Statistical Learning under Selection Bias

Van Tinh Tran and Alex Aussem

LIRIS Laboratory, University of Lyon 1

Summary

Selection bias is pervasive in almost all empirical studies, including Machine Learning. However, when training and test distribution differ only in input distribution (known as *covariate shift*) or only in output distribution (known as *prior probability shift*), adaptation methods using importance weight were proven to be effective. In this thesis, we first present a general framework to correct a more general class of selection bias problem which includes covariate shift and prior probability shift as special cases. We then show that the method of importance weighting applied to covariate shift is sub-optimal and discuss a manner to optimally combine the important-weighted and the unweighted models in order to improve the predictive performance in the target domain. Both methods are supported by theoretical and experimental result.

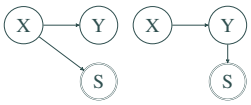
Background

Selection bias, which occurs when training and test joint distributions are different, i.e. $P_{tr}(x,y) \neq P_{te}(x,y)$, is pervasive in almost all empirical studies, including Machine Learning Statistics, Social Sciences, Economics, Bioinformatics, Biostatistics, Epidemiology, Medicine, etc. It is therefore highly desirable to devise algorithms that remain effective under such distribution shifts. It is well known that one may account for the difference between $P_{tr}(x)$ and $P_{te}(x)$ by re-weighting the training points using the so-called importance weight, denoted as $\beta(x,y) = \frac{P_{te}(x,y)}{P_{tr}(x,y)}$

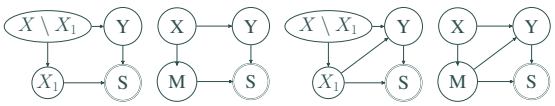
In general, the estimation problem with two different distributions $P_{tr}(x,y)$ and $P_{te}(x,y)$ is unsolvable, as the two terms could be arbitrarily far apart. However, when $P_{tr}(x,y)$ and $P_{te}(x,y)$ differ only in $P_{tr}(x)$ and $P_{te}(x)$ (known as *covariate shift*), the importance weight is reduced to $\beta(x) = P_{te}(x)/P_{tr}(x)$, which require unlabeled examples ($P(x)$) to estimate. Similarly, when $P_{tr}(x,y)$ and $P_{te}(x,y)$ differ only in $P_{tr}(y)$ and $P_{te}(y)$ (known as *prior probability shift*), the importance weight is reduced to $\beta(y) = P_{te}(y)/P_{tr}(y)$, which require a sample of label population ($P(y)$) to estimate.

Generalize covariate shift and prior probability shift

When $P_{tr}(x,y)$ and $P_{te}(x,y)$ differ only in $P_{tr}(x)$ and $P_{te}(x)$ (known as *covariate shift*) or only in $P_{tr}(y)$ and $P_{te}(y)$ (known as *prior probability shift*), it is known that importance weighting can correct the selection bias.



In [1], we generalized covariate shift and prior probability shift and showed that, if we have a combination of biased data and unbiased data and qualitative probabilistic assumptions that are deemed plausible about our sampling mechanism, our problem becomes solvable. More specifically, we assume we have access to a S-control feature vector, X_s, and some additional sample of the form (x_s) that is drawn from the population as a whole, such that S is conditionally independent of (X,Y) given X_s. In that case the bias is correct by weight each training example by $\beta(x_s) = \frac{p_{te}(x_s)}{p_{tr}(x_s)}$.



Despite being limited to specific or idealized situations, this framework includes covariate shift and prior probability shift as special cases.

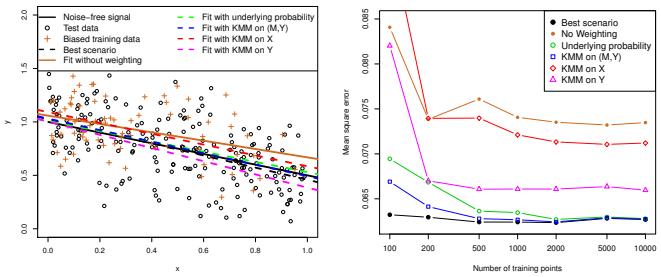


Figure 1: Experiment

Hybrid approach to covariate shift

Importance weighting in general is know to come at the expense of a reduction of the effective sample size. We show analytically in [2] that, while the unweighted model is globally more biased than the weighted one, it may locally be less biased on low importance instances. Our approach includes four steps:

- Define the metric to measure the global and local bias based on test distribution.
- Define alternative metric to measure the global and local bias based on training distribution, which is observed in training data.
- Establish the relationship between the global bias metric estimated on train distribution and local bias metrics estimated on test distribution.
- Show that on some local subsets of input feature space, the unweighted model is less bias then the weighted model. Discuss the method to identify those subset

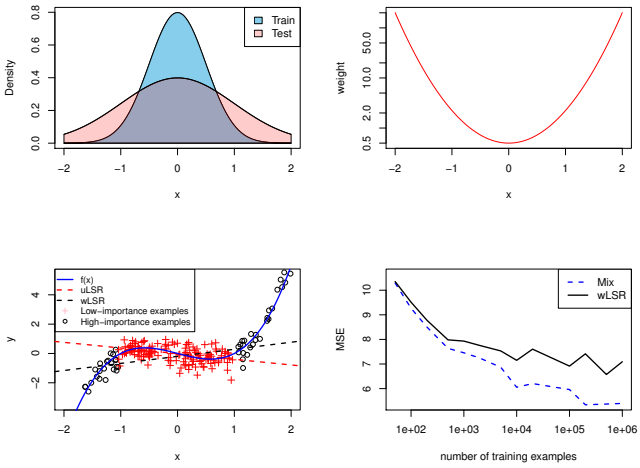


Figure 2: An illustrative example of fitting a function f(x) using a linear model with/without the weight importance scheme (wLSR/uLSR) and a combination of both (termed "Mix"). Top left: Test and train distribution of input x; top right: Importance weight $\beta(x)$; bottom left: true function, unweighted and weighted model on test data; bottom right: Mean Square Error vs training sample size.

Synthesized simple step sample selection distribution

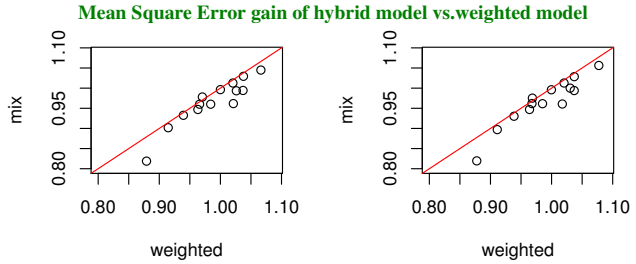


Figure 3: MSE gain of the mix model vs. MSE gain of the weighted model. Points below the diagonal line indicate that the mix model outperforms the weighted model. The weighted model was based on (a) the true selection probability and (b) based on the some noisy estimation of selection probability.

References

[1] Van-Tinh Tran and Alex Aussem. Correcting a class of complete selection biaswith external data based on importance weight estimation. In *ICONIP*, 2015. to appear.

[2] Van-Tinh Tran and Alex Aussem. A practical approach to reduce the learning bias under covariate shift. In *Machine Learning and Knowledge Discovery in Databases*, pages 71–86. Springer, 2015.

Bio inspired data mining algorithms taking advantage of evolution of evolution

Sergio Peignier, Christophe Rigotti, Guillaume Beslon

2ème year of PhD, Funded by EU-FET, BEAGLETeam
sergio.peignier@inria.fr – http://liris.cnrs.fr/membres?idn=peignier

Abstract

According to [1], bio-inspired optimization algorithms could be improved by incorporating knowledge from molecular and evolutionary biology. A promising source of advances in optimization is one of the important phenomena in evolutionary biology: the dynamic evolution of the genome structure. Several studies showed for instance that an evolvable genome structure allows evolution to modify the effects that evolution operators (e.g., mutations) have on individuals, a phenomenon known as $evolution\ of\ evolution$ [2]. My thesis takes place within the european project EvoEvo (FP7 funding, http://evoevo.eu/, ICT-610427) and aims to take advantage of $evolution\ of\ evolution$ mechanisms to achieve data mining tasks on dynamic data.

Chameleoclust: Subspace clustering using evolvable genome structure

A first major step in this phD project was to design and assess Chameleoclust, a first evolutionary algorithm taking advantage of evolution of evolution to takle the subspace clustering problem. Subspace clustering is recognized as more difficult than standard clustering since it requires to identify not only the clusters but also the various subspaces where the clusters hold. We proposed to tackle this problem with a bio-inspired algorithm that incorporates a genome having an evolvable structure to allow for evolution of evolution. The key intuition in the design of the Chameleoclust algorithm is to take advantage of such an evolvable structure to detect various numbers of clusters in subspaces of various dimensions and to self-tune the main evolutionary parameters (e.g., levels of variability).

The algorithm has been assessed using a reference framework for subspace clustering evaluation, and compared to state-of-the-art algorithms on both real and synthetic datasets [3]. The results obtained with the Chameleoclust algorithm show that evolution of evolution, through an evolvable genome structure, is useful to solve a difficult problem such as subspace clustering. The reader is referred to [4] for a detailed description of Chameleoclust.

Perspectives

Analysing dynamic data streams

The most immediate perspective is to test and evaluate Chameleoclust using dynamic data stream: we propose to take advantage of $evolution\ of\ evolution$ mechanisms to tackle the dynamic subspace clustering problem. In order to achieve this goal we have been testing our algorithm using sliding windows over static data sets to simulate dynamic data streams.

Designing data mining dedicated algorithms

Another perspective of this work is to produce data mining dedicated algorithms based on the knowledge and the principles obtained from the study of the $evolution\ of\ evolution$ mechanisms.

Assessing the algorithms

Further applications are also targeted in order to assess our algorithm in the context of dynamic data streams, e.g., analyze dynamic Wi-Fi contexts, analyze data from motion sensors attached to a dancer.

References

[1] W. Banzhaf, G. Beslon, S. Christensen, J. A. Foster, F. Képès, V. Lefort, J. F. Miller, M. Radman, and J. J. Ramsden, "Guidelines: From artificial evolution to computational evolution: a research agenda," *Nature Reviews Genetics*, vol. 7, no. 9, pp. 729–735, 2006.

[2] T. Hindré, C. Knibbe, G. Beslon, and D. Schneider, "New insights into bacterial adaptation through in vivo and in silico experimental evolution," *Nature Reviews Microbiology*, vol. 10, pp. 352–365, May 2012.

[3] E. Müller, S. Günnemann, I. Assent, and T. Seidl, "Evaluating clustering in subspace projections of high dimensional data," in *Proc. 35th Int. Conf. on Very Large Data Bases (VLDB 2009)*, (Lyon, France), pp. 1270–1281, 2009.

[4] S. Peignier, C. Rigotti, and G. Beslon, "Subspace clustering using evolvable genome structure," in *Proc. of the ACM Genetic and Evolutionary Computation Conference (GECCO 2015)*, pp. 1–8, 2015.

SUBSPACE CLUSTERING USING EVOLVABLE GENOME STRUCTURE+

Sergio Peignier*, Christophe Rigotti and Guillaume Beslon

GOAL

Definitions

. Dataset objects defined over a set of dimensions D = {d1, d2, . . . }.
. Cluster : Set of similar points.
. Clustering : Dataset →{Cluster1, Cluster2, ...}.
. Subspace Clustering : Dataset →{ (Cluster1, Subspace1), (Cluster2, Subspace2), ... }.
. Subspace Clustering : Different clusters may be defined in different subspaces.

Subspace clustering specificities
- Unknown number of clusters.
- Unknown subpace dimensionalities.

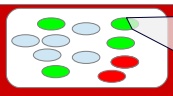
⟺

- Variable genome length
- Non-functional elements

Evolvable Genome Structure

ALGORITHM

Generation T

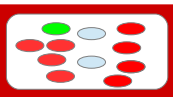


Selection

Exponential Ranking (reproduction probability) with elitism.

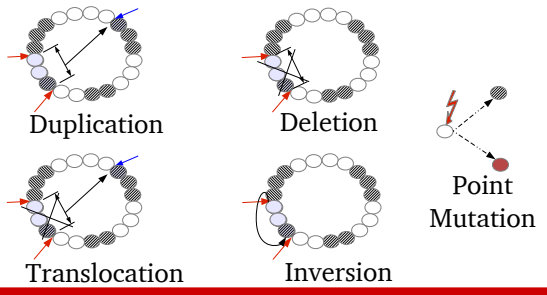
$$p_\alpha = (s-1)\frac{s^{N-r_\alpha}}{s^N-1}$$

Progeny



Mutations

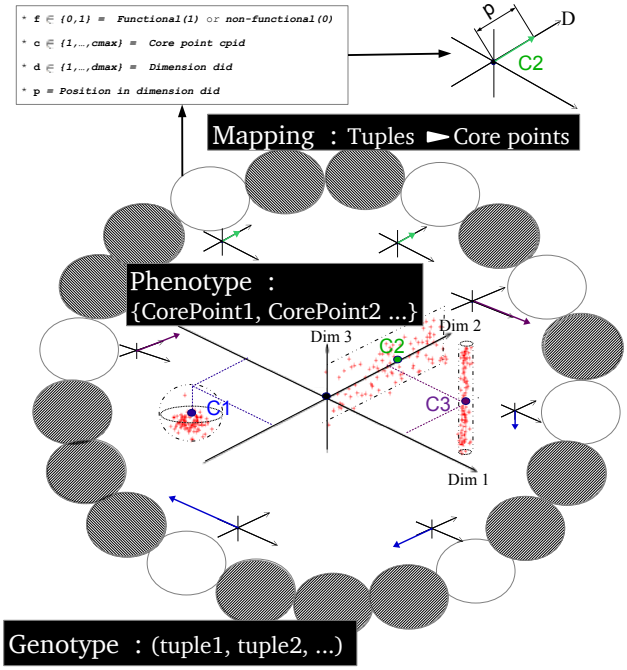
- Mutation rate : Probability of mutation per tuple.



Generation T+1

Global evolution model inspired by In silico evolution formalisms
[Crombach and Hogeweg, 2007] [Knibbe et al., 2007]

Chameleoclust Individual



Defining the fitness: Individual clustering → Quality

- Dataset S defined over a set of dimensions D = {d1, d2 . . . }
- Core point $p_c \in \Phi$ **(Phenotype)**
- S_{p_c} : Set of dataset objects associated to p_c

Manhattan segmental distance ([Aggarwal et al.,1999])
Between $\mathbf{y_1 = (y_{1,1}, y_{1,2}, \ldots)}$ and $\mathbf{y_2 = (y_{2,1}, y_{2,2}, \ldots)}$

$$d_{\mathcal{D}}(y_1, y_2) = \sum_{i \in \mathcal{D}} \frac{|y_{1,i} - y_{2,i}|}{|\mathcal{D}|}$$

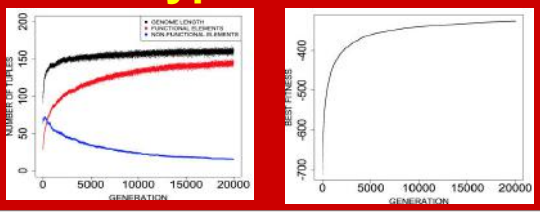
Assignment mismatch
Each object $x \in S$ is assigned to the core point $P_c \in \Phi$ for which $\mathcal{E}(x, p_c)$ is minimal.

$$\mathcal{E}(x, p_c) = \frac{|\mathcal{D}_{p_c}| \cdot d_{\mathcal{D}_{p_c}}(x, p_c) + |\mathcal{D} \setminus \mathcal{D}_{p_c}| \cdot d_{\mathcal{D} \setminus \mathcal{D}_{p_c}}(x, \mathcal{O})}{|\mathcal{D}|}$$

Fitness
Individual clustering quality : dispersion around core points

$$\mathcal{F} = -\sum_{p_c \in \Phi} \sum_{x \in S_{p_c}} \mathcal{E}(x, p_c)$$

Typical Run



EXPERIMENTS RESULTS

Evaluation experiments

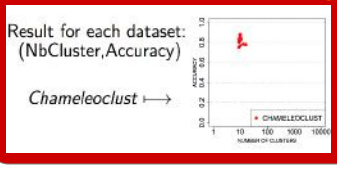
- Compare to state-of-the-art algorithms
- Reference evaluation framework for subspace clustering [Müller et al., 2009] :
  + 7 real world benchmark datasets
  + 16 synthetic benchmark datasets

Results : Real world data

Dataset	NumClusters	AvgDim
breast	15.2	5.32
diabetes	60.8	2.06
glass	34.8	2.67
liver	62.8	1.93
pendigits	17.5	5.61
shape	14.9	7.27
vowel	48.7	2.59

Regulation of the number of clusters found and their dimensionality with a single parameter setting

Results : Synthetic data



Competitive accuracy and realistic number of clusters

Results : Real world data (shape)

		CLIQUE	DOC	MINECLUS	SCHISM	SUBCLU	FIRES	INSCY	PROCLUS	P3C	STATPC	Chameleoclust
Accuracy	max	0.76	0.79	0.79	0.74	0.7	0.51	0.76	0.72	0.61	0.74	0.82
	min	0.76	0.54	0.80	0.49	0.64	0.44	0.48	0.71	0.61	0.74	0.73
RNIA	max	0.07	0.90	1.00	0.26	0.05	0.25	0.37	0.61	0.17	0.55	0.6
	min	0.07	0.82	1.00	0.01	0.04	0.20	0.24	0.37	0.17	0.55	0.48
NumClusters	max	486	53	64	8835	3468	10	185	34	9	9	16
	min	486	29	32	90	3337	5	48	34	9	9	13
AvgDim	max	3.3	13.8	17.0	6.0	4.5	7.6	9.8	13.0	4.1	17	10
	min	3.3	12.8	17.0	3.9	4.1	5.3	9.5	7.0	4.1	17	7.3
RunTime	max	235	2E+06	46703	712964	4063	63	22578	593	140	250	314
	min	235	86500	3266	9031	1891	47	11531	469	140	171	287

- Competitive performances with respect to other algorithms.
- Good compromise between the different evaluation measures.

+ Peignier, S., Rigotti, C., Beslon, G.: Subspace clustering using evolvable genome structure. In: Proc. of the ACM Genetic and Evolutionary Computation Conference (GECCO 2015). pp. 1–8 (2015)

* Sergio.Peignier@inria-lyon.fr

This research has been supported by EU-FET grant EvoEvo (ICT- 610427)

AGGREGATING AND MANAGING BIG REALTIME DATA (AMBED) IN THE CLOUD:

APPLICATION TO INTELLIGENT TRANSPORT FOR SMART CITIES

Gavin Kemp, Catarina Ferreira da Silva, Genoveva Vargas-Solar, Parisa Ghodous, Christine Collet

2nd year of PhD, Funded by ACR7, SOC Team
gavin.kemp@liris.cnrs.fr - http://liris.cnrs.fr/membres?idn=gkemp

Abstract

The increasing power of computer hardware and the sophistication of computer software have brought many new possibilities to information world. On one side the possibility to analyze massive data sets has brought new insight, knowledge and information. On the other, it has enabled to massively distribute computing and has opened to a new programming paradigm called Service Oriented Computing particularly well adapted to cloud computing. The possibility to analyze big data brings new insights into obscure and useful correlations in data providing undiscovered knowledge. Applying big data analytics to the transport data has brought better understanding to the transports network revealing unexpected choking points in cities. This technology is still largely inaccessible to small companies due to their limited computational resources and complex for large ones due to the time needed to develop a big data analytical system. Applying these new technologies to the transport industry can bring new understanding of town transport infrastructures. The objective of our work is to manage and aggregate cloud services for managing big data and assist decision making for transport systems. Thus this poster presents our approach for developing data storage, data cleaning and data integration services to make an efficient decision support system. Our services will implement algorithms and strategies that consume storage and computing resources of the cloud. For this reason, appropriate consumption models will guide their use. Proposing big data management strategies for data produced by transport infrastructures, whilst maintaining cost effective systems deployed on the cloud, is a promising approach.

Architecture

Infrastructure: In cloud computing everything is viewed as a service (XaaS). We are constructing the big data infrastructure around the five step for big data analytics proposed by Jagadish and co. [1]:

Data acquisition service: hardware and infrastructure services that transfer to NoSQL data stores adapted to the format of the data, the data acquired by the vehicles, users, and sensors deployed in cities (e.g. roads, streets, public spaces). We are using node.js service to collect data from data.grandlyon.fr [2], tweeter and bing.

Information extraction and cleaning service: at this level information is extracted from the unstructured data (e.g. video stream). It is also responsible for scanning the data store, seeking for outliers and duplicates of data.

Integration and aggregation services: the real time service combines data from multiples sources (e.g. video, sensor, weather data) into a single URL. The historic data provides a unified unql view to the data analytical services to perform unql queries.

Big data analysis service and decision support service: these services provide a synthetic vision of the data. The first service provides algorithms to analyze the data. The decision support services provide the interface to understand the information.

Scenario

This architecture is being tested on a scenario where a taxi company needs to embed decision support in electric vehicles, to help their global optimal management. The company uses electric vehicles that implement a decision cycle to reach their destination while ensuring optimal recharging, through mobile recharging units. The decision making cycle aims at ensuring vehicles availability both temporally and spatially; and service continuity by avoiding congestion areas, accidents and other exceptional events. In this perspective, the company uses data coming from multiple open databases as well as service available on the cloud to analyze the open data to support their decision support services.

Bibliography

[1] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, *Big Data and Its Technical Challenges*, vol. 57, no. 7. 2014.

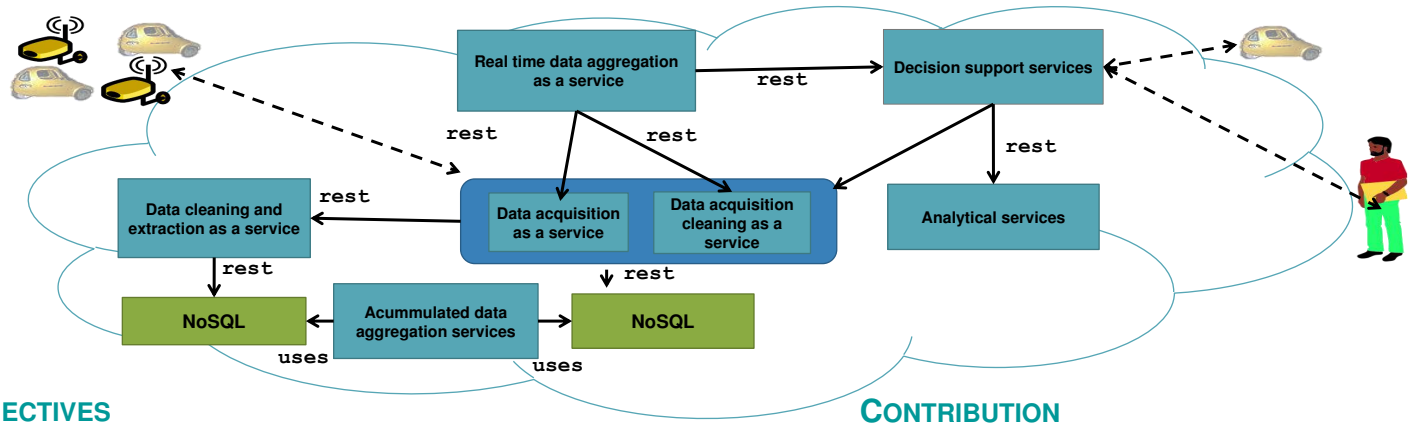
[2] GrandLyon, “Smart Data,” 2015. [Online]. Available: http://data.grandlyon.com/.

AGGREGATING AND MANAGING BIG REALTIME DATA (AMBED) IN THE CLOUD

Application to intelligent transport for smart cities

Gavin Robert Kemp

gavin.kemp@liris.cnrs.fr



OBJECTIVES

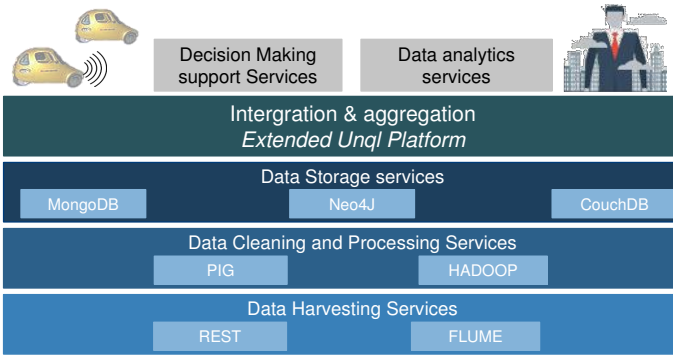
- Develop services using big data for decision making
- Use cloud and streaming as tools
- Insure that big data, cloud and streaming work well together

CONTRIBUTION

Big data cloud service oriented architecture

- Data acquisition
- Information extraction and cleaning
- Real-time data integration and aggregation
- Big data analysis and decision support

CLOUD BIG DATA SERVICES FOR TRANSPORT



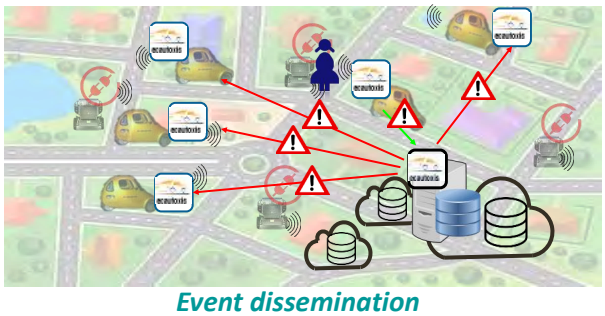
AUTONOMOUS VEHICLES CASE STUDY

Help pilot vehicles

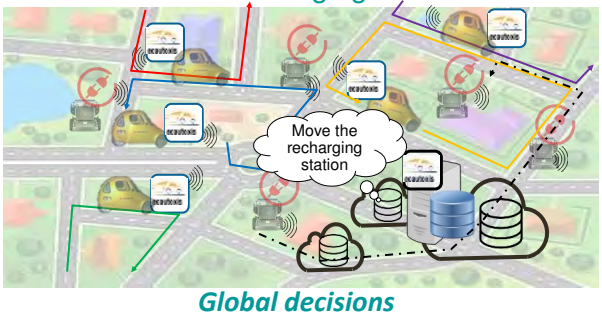
- Avoiding accidents through event dissemination
- Avoiding congestion

Vehicles availability

- Spatio-temporal location
- Optimal charging
- Mobile recharging units



Event dissemination



Global decisions

PhD supervisors

Catarina Ferreira Da Silva, catarina.ferreira@univ-lyon1.fr, LIRIS

Parisa Ghodous, parisa.ghodous@univ-lyon1.fr, LIRIS

Genoveva Vargas-Solar, genoveva.vargas@imag.fr, LIG-LAFMIA

Christine Collet, christine.collet@grenoble-inp.fr, LIG

Toward an Integrated Evolutionary Model to study Evolution of Evolution

Charles Rocabert, Carole Knibbe, Guillaume Beslon

3rd year of PhD, funded by the European Commission under FP7
charles.rocabert@liris.cnrs.fr - http://liris.cnrs.fr/membres?idn=crocaber

Abstract

Variation and Selection are the two core processes of Darwinian Evolution. Yet, both are directly regulated by many processes that are themselves products of evolution. Microorganisms efficiently exploit this ability to dynamically adapt to new conditions: evolution seems to have optimised their own ability to evolve, as a primary mean to react to environmental changes. We call this process Evolution of Evolution (EvoEvo). Here, we propose to use an integrated evolutionary model including a complex and evolvable genotype-to-phenotype mapping to study EvoEvo. As exemplified in the poster, the integrated evolutionary model will allow us to decipher the EvoEvo strategies and to offer new hypothesis and predictions on the evolution of microorganisms.

Introduction

Life on Earth evolved for billion years in ever changing environments, undergoing smooth or brutal, cyclic or unseen variations. To survive in such conditions, being adapted to the current environment is not enough. Extant organisms had to deal with the evolutionary competition but they also had to deal with the variations of their environment to stay adapted despite the rapid and sometimes profound crises they had to cope with. How did they do so is an open question. Did the extant organisms survive by chance or did they survive because, being regularly confronted to such crises, they evolved reaction/adaptation mechanisms?

Background

Experimental evolution, where fast replicating organisms (e.g. bacteria or viruses) are evolved in controlled environments for thousands of generations, allows to recover precisely the evolution history of lab strains by reviving frozen samples and performing data analysis. Many evolution results have shown that microorganisms are able to evolve at an amazing speed: in virtually all experimental frameworks that use bacteria or viruses, important phenotypic innovations have emerged in only a few tens of generations and in many cases, evolution tend to be partly reproducible. Microorganisms efficiently use mutation and selection to dynamically adapt to new conditions. Thus, evolution seems to have optimized their own ability to evolve, as a primary mean to react to environmental changes. We call this process “Evolution of Evolution” (EvoEvo, [Hindré2012]).

Experimental evolution, despite its explanatory and statistical power, remains a long and costly process. An alternative is to simulate evolution in a computer. However, Evolution of Evolution implies the interaction of a wide range of biological structures (e.g. genome, genetic regulation network, metabolic network, ...), so we need to develop complex models. Following this idea, *in silico* experimental evolution is a growing field in evolutionary biology,and a lot of theoretical questions have been deciphered with this approach.

A new model to study EvoEvo

In the context of the EvoEvo project, we have designed an integrated model to study Evolution of Evolution. The aim of this individual based, multilevel model is to study the evolutionary process and to unravel the EvoEvo strategies that result from a pure Darwininan evolution. Here we propose to include five levels in our model: the genome, the genetic network, the metabolic network, the fitness and the environment.

In the poster, we will present our formalism, inspired from [Beslon2010] and [Crombach2008], and describe the integrated evolutionary model. Then, we will present a quick example of the evolutionary dynamic observed in the model.

Bibliography

[Beslon2010] Beslon, G., Parsons, D. P., Pena, J. M., Rigotti, C., Sanchez-Dehesa, Y.: *From digital genetics to knowledge discovery: Perspectives in genetic network understanding*. Intelligent Data Analysis, 14(2), 173-191 (2010).

[Crombach2008] Crombach, A., Hogeweg, P.: *Evolution of evolvability in gene regulatory networks*. PLoS computational biology, 4(7), e1000112 (2008).

[Hindré2012] Hindré, T., Knibbe, C., Beslon, G., Schneider, D.: *New insights into bacterial adaptation through in vivo and in silico experimental evolution*. Nature Reviews Microbiology, 10(5), 352-365 (2012).

Towards an integrated model to study Evolution of Evolution

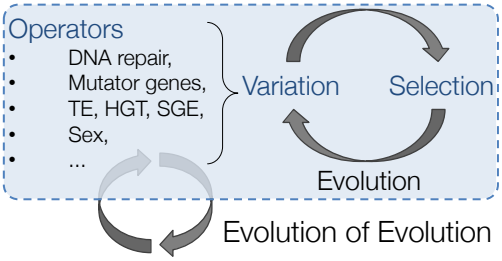
Charles Rocabert*, Carole Knibbe, Guillaume Beslon

Université de Lyon, INRIA, CNRS, LIRIS, UMR 5205, France

*charles.rocabert@inria.fr

Key concepts: What is Evolution of Evolution ?

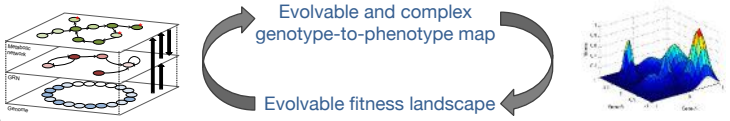
Variation and **Selection** are the two core processes of Darwinian Evolution. Yet, both are directly regulated by many processes that are themselves products of evolution (e.g. DNA repair, mutator genes, transposable elements, horizontal transfer, stochasticity of gene expression, sex, network modularity, niche construction...). **This results in the ability of evolution to self-modify its operators, hence its dynamics**. We call this process "**Evolution of Evolution**" or EvoEvo. Different EvoEvo strategies have been proposed in the literature, including **regulation of variability**, **robustness/evolvability strategies**, **bet-hedging**... However, most of these strategies are poorly characterized and the conditions under which they evolve as well as their consequences are generally unknown.



Key concepts: Why an Integrated model of evolution ?

(A) – EvoEvo is an **integrative concept** exactly as fitness is. Robustness/evolvability/variability of the phenotype is the result of the interaction of robustness/evolvability/variability at all the organization levels of the organism, including its interactions with its environment.

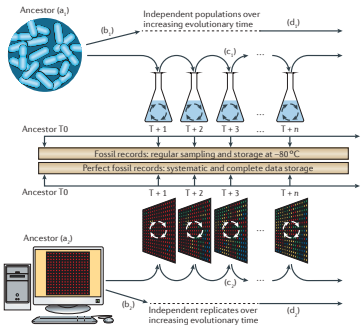
That is why **a computational model of EvoEvo must be integrated**, including the main organization levels of the genotype-to-phenotype map.



(B) – Experiments of ***in vivo* experimental evolution** are mimicked in a computational framework to produce hypothesis and predictions on the Evolution of Evolution theory.

« *Simulation models can be used to mimic complex systems, but unlike nature, can be manipulated in ways that would be impossible, too costly or unethical to do in natural systems.* » (Peck, 2004).

Peck, S. L. (2004). Simulation as experiment: a philosophical reassessment for biological modeling. Trends in Ecology & Evolution, 19(10), 530-534.



Overview of our Integrated evolutionary model

Artificial chemistry (Achem):
- Molecules are integers
- Metabolic reactions are operations on integers
- Genome units and proteins are n-tuples

Let's $c_i(t)$ be the concentration of the TF i at time t and $coE_i(t)$ the concentration of its co-enzyme. Depending on the state of the free activity A_i^F and the bound activity A_i^B, the concentration of the TF i that is active $a_i(t)$ at time t is:

(6) $a_i(t) = \begin{cases} min(c_i(t), coE_i(t)), & IF(A_i^F \text{ is false AND } A_i^B \text{ is true}) \\ max(c_i(t) - coE_i(t), 0), & IF(A_i^F \text{ is true AND } A_i^B \text{ is false}) \\ c_i(t), & IF(A_i^F \text{ is true AND } A_i^B \text{ is true}) \\ 0, & IF(A_i^F \text{ is false AND } A_i^B \text{ is false}) \end{cases}$

(1) $A_s(t) = \sum_j c_j(t) A_{js}$

with $c_j(t)$ the concentration of the transcription factor j at time t and A_{js} the affinity of this transcription factor for the binding site s.

Then, the transcription rate e_i over time of the promoter i is given by the Hill-like function:

(3) $e_i(t) = \beta_i \cdot \left(\frac{\theta^n}{o_i(t)^n + \theta^n}\right) \cdot \left(1 + \left(\frac{1}{\beta_i} - 1\right)\left(\frac{E_i(t)^n}{E_i(t)^n + \theta^n}\right)\right)$

with β_i the basal expression level of the promoter i, n and θ being constant coefficients that determine the shape of the Hill function.

From (1), we deduce the activity of the enhancer site $E_i(t)$ and of the operator site $O_i(t)$ flanking the promoter i:

(2) $\begin{cases} E_i(t) = \sum_{s \in enhancer_i} A_s(t) \\ O_i(t) = \sum_{s \in operator_i} A_s(t) \end{cases}$

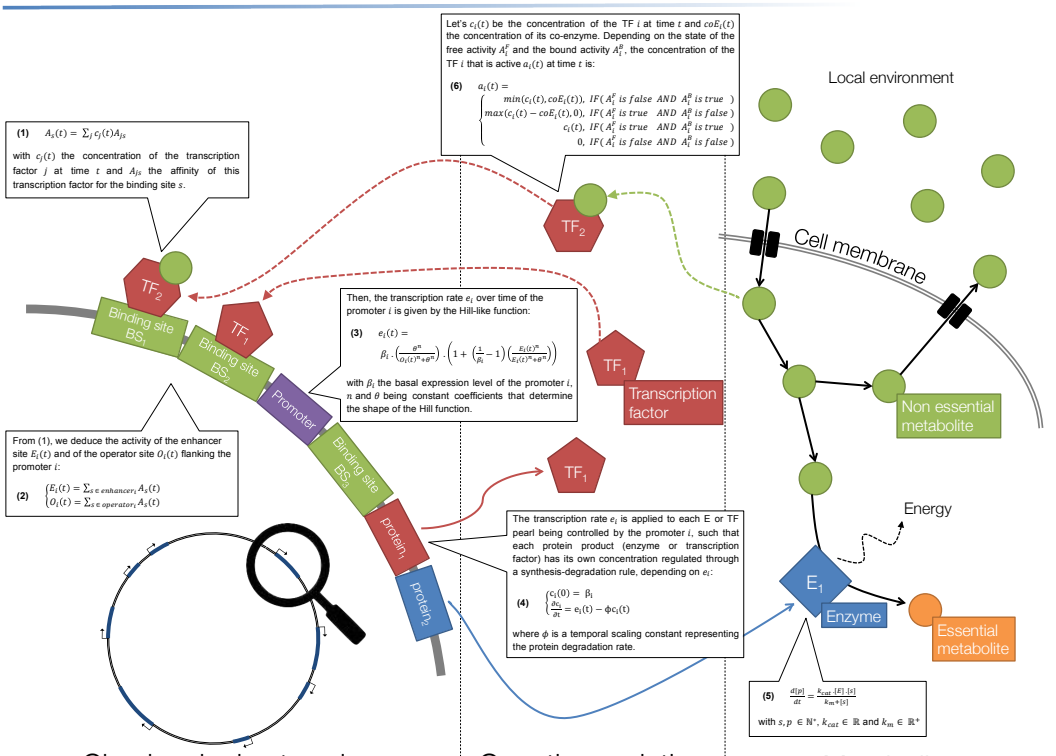
The transcription rate e_i is applied to each E or TF pearl being controlled by the promoter i, such that each protein product (enzyme or transcription factor) has its own concentration regulated through a synthesis-degradation rule, depending on e_i:

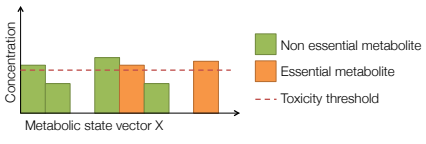
(4) $\begin{cases} c_i(0) = \beta_i \\ \frac{dc_i}{dt} = e_i(t) - \phi c_i(t) \end{cases}$

where ϕ is a temporal scaling constant representing the protein degradation rate.

(5) $\frac{d[p]}{dt} = \frac{k_{cat} \cdot [E] \cdot [s]}{k_m + [s]}$

with $s, p \in \mathbb{N}^*$, $k_{cat} \in \mathbb{R}$ and $k_m \in \mathbb{R}^+$





Fitness function $w = f(X_{ne}, X_e, T)$

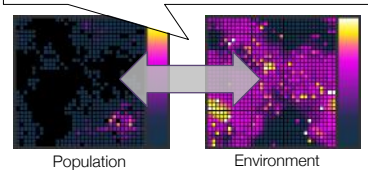
The environment level is characterized by the following parameters:
- Environment size $W.H$ ($W.H$ thus being the maximal carrying capacity of the environment),
- Diffusion coefficient D,
- Degradation rate D_g,
- Metabolites inflow I_i (for each metabolite i),
- Metabolites outflow O_i (for each metabolite i),
- The migration coefficient mig,
- A minimum threshold.

Note that the dynamics of inflow/outflow can be constant (i.e. applied at each time step) or time dependent.

At a given time t each lattice cell of coordinate (x, y) is characterized by:
- The individual that occupies the cell (possibly null)
- The list of free metabolites that are available and their concentrations $C_i(t)$

Given the parameters of the environment, the dynamics of a free metabolite i in a cell is then given by ($C_{i,j}$ being the concentration of metabolite i in the lattice cell j):

(7) $C_i(t+1) = C_i(t) - DgC_i(t) + \sum_{j \text{ in neighbours}} D\,C_{i,j}(t) - 8DC_i(t) + I_i(t) - O_i(t)$



Circular single strand genome

Genetic regulation network

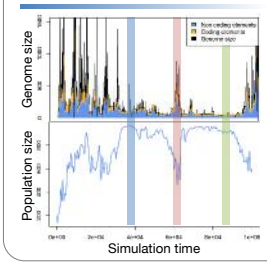
Metabolic network

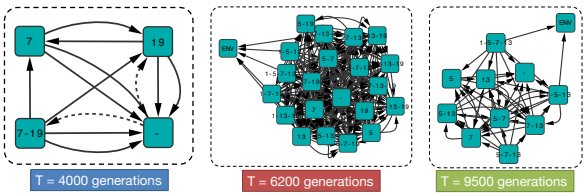
Population and environment

Evolution (mutations, selection)

Results

Population evolves through cycles of stabilizing selection periods and environmental crisis. Trophic networks implying producers and parasites evolve, producers enriching their environment and favouring parasites fixation. Parasites deplete their environment, leading to crisis. At each crisis, we observe genome size explosions, trophic network complexification and phenotypic innovation.





Trophic group

Consuming link

« Helping » link

This work is supported by the European Commission 7th Framework Programme (EvoEvo project FP7-ICT-610427)

www.evoevo.eu

Traitement joint de nuage de points et d'images pour l'analyse et la visualisation des formes 3D

GUISLAIN Maximilien, DIGNE Julie, CHAINE Raphäelle, MONNIER Gilles

2ème année de thèse, Financement CIFRE, Équipe GEOMOD
maximilien.guislain@liris.cnrs.fr – https://liris.cnrs.fr/membres?idn=mguislai

Résumé de la thèse

Cette thèse CIFRE, en partenariat avec la société Technodigit, s'inscrit dans le cadre d'une utilisation des informations complémentaires fournies par des instruments de numérisation. La problématique de cette thèse est d'enrichir les informations géométriques acquises par un scanner laser, typiquement des nuages de points d'une scène urbaine, avec d'autres informations issues par exemple d'une série de photos prises dans cette même scène. Ainsi la géométrie pourra être consolidée et de nouveaux outils seront proposés pour traiter les données fournies en masse par les systèmes LIDAR (LIght Detection And Ranging) modernes.

Problématique

Les données issues de sytèmes LIDAR modernes tels que ceux provenant de scanners mobiles embarqués (tel que le *Leica Pegasus : two*) fournissent à la fois un nuage de points dense et des photos de l'environnement. La plupart de ces scanners sont couplés à un capteur CCD qui permet d'attribuer une couleur aux points mesurés par le laser. Cependant, les couleurs ainsi attribuées ne sont pas d'une définition suffisante pour permettre de capter certains détails importants, comme par exemple des variations dans une texture [2]. La ré-attribution de ces couleurs à posteriori nécessite une parfaite concordance entre les données LIDAR et les images, concordance qui n'est pas toujours suffisamment précise en utilisant les informations de géolocalisation embarquées du LIDAR.

Approches existantes et pistes étudiées

Le positionnement d'images sur une géométrie complémentaire peut être effectué de plusieurs manières, par exemple en utilisant des descripteurs images tels que SIFT comme proposé par Gonzalez et al.[1] ou Moussa et al.[2]. Cependant, ces techniques ne peuvent être appliquées que dans le cas où le nuage de points est composé d'informations riches (réflectance laser, couleurs) et ne sont d'aucune utilité si l'on ne possède que des informations géométriques (positions et normales estimées des points). Une seconde technique couramment utilisée pour les images est basée sur la maximisation de l'information mutuelle. Elle peut être adaptée pour être utilisée sur la géométrie et permet d'utiliser des informations de plusieurs modalités [4] [3].

Le recalage par information mutuelle, bien qu'il présente de bons résultats en utilisant des données fortement corrélées entre elles (couleurs/couleurs, couleurs/réflectance) s'est révélé décevant dans les scènes urbaines complexes ne possédant que des informations géométriques. Un recalage par descripteurs ne fonctionne pas non plus dans ces conditions, l'absence de texture rendant impossible de trouver des points d'intérêt commun. Á l'inverse la comparaison des Histogram of Oriented Gradients (HOG) entre les images des nuages de points et les photos donne des résultats plus robustes que l'information mutuelle dans un panorama de ville sans information supplémentaire.

Résultats et travaux futurs

Dans le cas d'un recalage à 6 degrés de libertés (translation et rotation), sur des données réelles, le recalage est effectué avec une précision d'environ 3 à 10 pixels, pour des images d'une résolution de 2046 par 2046. Il est à noter toutefois que ces données étant issues d'une expérimentation réelle, une imprécision persiste, ceci étant dû aux imprécisions de calibrations des camera couleur.

La prochaine étape envisagée en plus de l'amélioration des temps de calcul du recalage pourrait être la colorisation des nuages de points à partir d'images mutliples, sans ajout de bruit ou de fausses couleurs. Un autre aspect qui découlera directement du recalage pourrait être la détection et la complétion automatique d'ombres projetées dans un nuage de points.

Références

[1] Diego González-Aguilera, Pablo Rodríguez-Gonzálvez, and Javier Gómez-Lahoz. An automatic procedure for co-registration of terrestrial laser scanners and digital cameras. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(3) :308–316, 2009.

[2] Wassim Moussa. *Integration of digital photogrammetry and terrestrial laser scanning for cultural heritage data recording*. PhD thesis, University of Stuttgart, 2014.

[3] Geoffrey Pascoe, Will Maddern, and Paul Newman. Robust direct visual localisation using normalised information distance. 2015.

[4] Zachary Taylor and Juan Nieto. Automatic calibration of lidar and camera images using normalized mutual information. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 2013.

Traitement joint de nuage de points et d'images pour l'analyse et la visualisation des formes 3D

Maximilien GUISLAIN[1,2], Julie DIGNE[1], Raphäelle CHAINE[1], Gilles MONNIER[2]
[1] Laboratoire d'Informatique en Image et Systèmes d'information - Université Lyon 1 - équipe GEOMOD
[2] TECHNODIGIT - Hexagon Metrology

1. Abstract

Cette thèse CIFRE, en partenariat avec la société Technodigit, s'inscrit dans le cadre d'une utilisation des informations complémentaires fournies par des instruments de numérisation. La problématique de cette thèse est d'enrichir les informations géométriques acquises par un scanner laser, typiquement des nuages de points d'une scène urbaine, avec d'autres informations issues par exemple d'une série de photos prises dans cette même scène. Ainsi la géométrie pourra être consolidée et de nouveaux outils seront proposés pour traiter les données fournies en masse par les systèmes LIDAR (LIght Detection And Ranging) modernes

2. Objectifs

Recalage d'image sur une géométrie correspondante

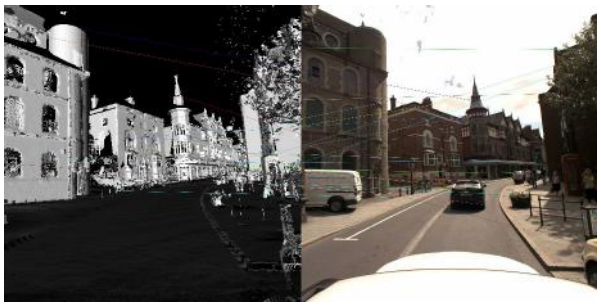
- Avoir un recalage sur de grands nuages ed points
- Doit fonctionner sur des parties peu denses
- Doit fonctionner avec peu d'informations

3. Contributions

- Processus de génération d'images à partir d'un nuages de points LIDAR.
- Méthode de recalage iterative et rapide.
- Métrique de distance entre images robuste et fonctionnelle.

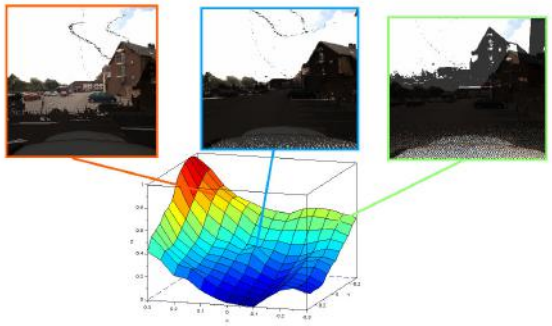
3. Etat de l'Art

Basé sur descripteurs



Descripteur SIFT [1] - non fonctionnel sur les normales

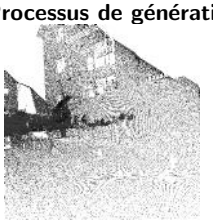
Basé sur l'information mutuelle



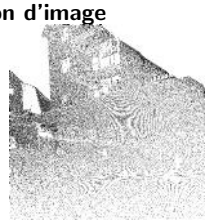
Maximisation d'information mutuelle [3] [2] - instable sur les normales

4. Résultats

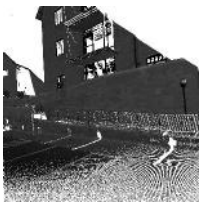
Processus de génération d'image



Projection

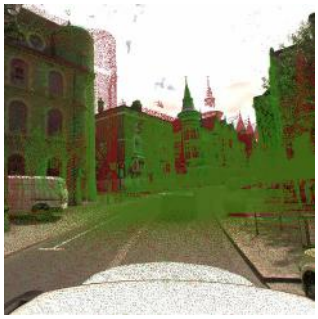


Résolution de l'Occultation



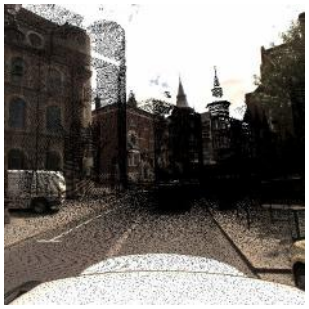
Interpolation

Résultats obtenus par la méthode proposée

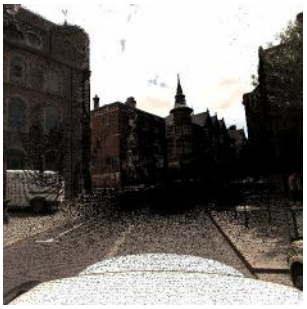


recalage initial (en rouge) et résultat de la méthode (en vert)

Comparaison des résultats



Approche classique



Approche proposée

Approche classique (details)

Approche proposée(details)

5. Conclusion

Proposition d'une méthode de recalage d'image sur géométrie

- Aucun *a priori* sur les données d'entrée
- Robustesse de la méthode face au bruit et au manque d'informations

[1] D. González-Aguilera, P. Rodríguez-González, and J. Gómez-Lahoz. An automatic procedure for co-registration of terrestrial laser scanners and digital cameras. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(3) :308–316, 2009.

[2] G. Pascoe, W. Maddern, and P. Newman. Robust direct visual localisation using normalised information distance. 2015.

[3] Z. Taylor and J. Nieto. Automatic calibration of lidar and camera images using normalized mutual information. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 2013.

Ces travaux ont été financés par Technodigit - Part of Hexagon Group.

Fouille de traces unitaires de produits manufacturés

Olivier Cavadenti, Mehdi Kaytoue, Jean-François Boulicaut

3ème année de thèse, Financement CIFRE, Équipe DM2L
olivier.cavadenti@liris.cnrs.fr – https://liris.cnrs.fr/membres?idn=ocavaden

Résumé de la thèse

De nos jours, de très grandes quantités de données issues de la traçabilité de produits manufacturés sont disponibles. Ces traces unitaires permettent aux industriels de connaître leur système logistique et de repérer d'éventuels problèmes. Cependant, il reste difficile de découvrir des contrefaçons ou des marchés gris à partir de ces traces. Si les experts ont une connaissance a priori de leur domaine, ils ne connaissent pas en détail les évènements responsables de ces anomalies. Nous cherchons à produire des méthodes de fouille de données utilisant la connaissance experte ainsi que les comportements contenus dans les traces unitaires pour y découvrir des comportements anormaux (contrefaçons, marchés gris, détournements de produits) et décrire ces ensembles d'objets grâce à des évènements co-occurrents (*pattern mining*). Les enjeux scientifiques sont importants étant donnés les volumes importants de données et l'environnement bruité.

Modélisation de la filière logistique

Nous avons établi le concept de modèle de filière. Celui ci permet de décrire la filière industrielle sous la forme d'un graphe augmenté. L'ensemble des opérations possibles permettent une analyse poussée du réseau de flux d'objets (k-connexité, flux minimum, visualisation sous différentes granularités spatiales et temporelles,...). En plus d'apporter une meilleure connaissance du réseau logistique, il permet lors du processus de fouille d'anomalies de caractériser l'attendu.

Détection d'anomalies dans les traces unitaires

Traces, trajectoires et motifs Le processus de fouille des trajectoires de produits se fait à partir des traces brutes de mobilité. Nous transformons ces traces en contextes de fouille, en différents langages (itemsets, séquences, ...) que nous appelons des trajectoires. Ainsi, nous pouvons exécuter différents solveurs pour extraire des motifs connus de la littérature (motifs fréquents, motifs rares,...) et permettre une analyse des collections de traces unitaires.

Utilisation de la connaissance experte Le modèle de filière permet de fixer une connaissance a priori de la filière sous forme de graphe augmenté. Il sert à déterminer ce qui est normal et permet de classifier les trajectoires (par exemple si un produit est allé d'un site à un autre alors que ce n'était pas un cheminement possible pour le modèle de filière).

Description des anomalies L'étape de classification des trajectoires permet de construire trois bases de trajectoire : normales, anormales et suspectes. Il convient ensuite de décrire chaque base de traces par des motifs présents uniquement dans une base et non dans les autres.

Modèles prédictifs pour les traces

Exploitation du comportement des objets mobiles Les traces unitaires codent le comportement des objets, ce qui permet de découvrir des comportements communs à un ensemble d'objets. Nous avons exploité cette propriété en cherchant à découvrir à partir des traces de comportement si ces objets étaient guidés par un même processus caché. Pour cela, nous avons utilisé un modèle prédictif qui classifie les traces et extrait à partir de la matrice de confusion résultante les paires d'objets dont la prédiction est confuse (elle est répartie entre les deux objets) ce qui indique qu'un processus de même type guide ces objets (par exemple, une personne détourne régulièrement des objets à un endroit différemment de ce qui est attendu).

Application à des données réelles de comportements Nous avons appliqué cette approche à des données de jeux vidéos issues du jeu de stratégie Starcraft 2. En utilisant les actions du jeu effectuées par des joueurs à travers des avatars virtuels , l'approche permet d'extraire efficacement les paires d'avatars d'un unique joueur. La méthode a des débouchés chez les éditeurs de jeux vidéos pour la découverte de tricheurs mais aussi dans le domaine du sport électronique, et a donné lieu à deux publications dans des conférences internationales [MLSA15] [DSAA15].

Références

[MLSA15] O. Cavadenti, V. Codocedo, J. F. Boulicaut, M. Kaytoue - *Identifying Avatar Aliases in Starcraft 2* - Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2015 workshop, 11 September 2015, Porto, Portugal.

[DSAA15] O. Cavadenti, V. Codocedo, J. F. Boulicaut, M. Kaytoue - *When Cyberathletes Conceal Their Game : Clustering Confusion Matrices to Identify Avatar Aliases* - International Conference on Data Science and Advanced Analytics 2015, Paris, France (Acceptance Rate : 14.7%).

When Cyberathletes Conceal Their Game

Clustering Confusion Matrices to Identify Avatar Aliases

Olivier Cavadenti, Victor Codocedo, Jean-François Boulicaut, Mehdi Kaytoue

Université de Lyon, CNRS, INSA-Lyon, LIRIS, France

2015 IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015), October 19-21 2015, Paris, France

Motivations and Problem

Video games and esports

- A lucrative and flourishing industry
- Electronic-sports: professional cyberathletes with teams, commentators, sponsors, growing cash prizes, compete in international tournaments, ...
- World-wide fan base following video game live streaming such as *Twitch.tv*
- Millions of games played on a daily basis

New challenges in video game analytics

- For game editors: security issues, bugs & cheaters detection, balance issues, building fun yet challenging automated agents ("AIs"), ...
- For sport structures: player profiling, match preparation, outcome prediction, ...
- Identify who is the real player behind an avatar

The avatar aliases identification problem

- The trace generation model in StarCraft II: only the *generates* mapping is known

User —owns→ Account —contains→ Avatar —generates→ Trace

- Players have official (public) avatars but also un-official avatars to conceal their games while training on the Web

Our goal is to identify groups of avatars that belong to the same player

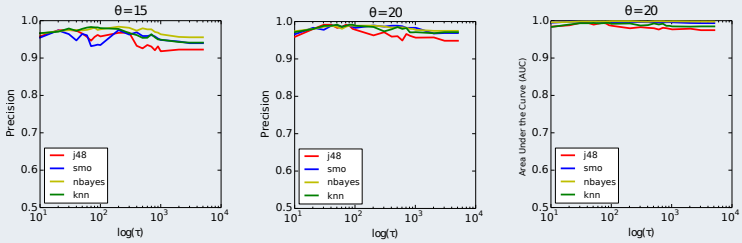
Identifying Players from Behavioural Data (or *gamelogs*)

Predicting an Avatar from his Keyboard Usage on StarCraft II (Blizzard Ent. ©, 2010)

Principle: *Given a database of behavioral traces T taking avatars as class labels in L, we construct a predictive model and its a confusion matrix*

- A game trace is the series of actions performed by the player, e.g. TAiLS generates the sequence $\langle Base, hotkey1a, s, hotkey1s, ...\rangle$ means that *TAiLS selects first with the mouse a building called Base, assigns it to hotkey **1**, selects some units with the mouse (s), etc.*
- Hotkeys offer a personalized configuration to drastically increase players velocity
- We use hotkey usage distributions as main features to train our classifiers
  - We consider only the τ first seconds of a game
  - We consider a trace only if its avatar is represented by at least θ examples

Hotkeys usage allows accurate avatar prediction after a few seconds of game (no aliases considered in the dataset WCS Season 2 EU 2014)



Unscrambling Confusion Matrices to Identify Avatar Aliases

Principle: *Given a normalized confusion matrix $\tilde{C}^\rho$, avatars of the same player locally and strongly concentrate the confusion*

Finding pairs of avatars such that

- $\tilde{C}^\rho_{ij} \simeq \tilde{C}^\rho_{ji} \simeq \tilde{C}^\rho_{ii} \simeq \tilde{C}^\rho_{jj}$
- $\tilde{C}^\rho_{ij} + \tilde{C}^\rho_{ji} + \tilde{C}^\rho_{ii} + \tilde{C}^\rho_{jj} \simeq 2$

	...	l_i	l_j	...
...	...			
l_i	...	C_{ij}	C_{ij}	...
l_j	...	C_{ji}	C_{jj}	...
...	...			

(l_1, l_2) are potentially avatars of the same player

	l_1	l_2	l_3	l_4	l_5
l_1	**0.6**	**0.4**	0	0	0
l_2	**0.4**	**0.55**	0.05	0	0
l_3	0	0	**0.8**	0.15	0.05
l_4	0	0.05	0	**0.7**	**0.25**
l_5	0	0	0	**0.5**	**0.5**

Method

- Each row and column of $\tilde{C}^\rho$ correspond to an avatar $l \in L$
- A value $\tilde{C}^\rho_{ij}$ gives the proportion of traces of the avatar l_i classified by ρ as the avatar l_j
- Each row is considered as a fuzzy set
- Fuzzy sets and their intersections form a $\sqcap$-semi-lattice
- Closed sets are extracted using Formal Concept Analysis
- Closed sets are scored (sum of the remaining membership degrees)
- Pairs are finally generated and ranked with a cosine similarity (to favor membership degrees concentrated on the diagonal), with two avatars $a_i, a_j \in A$:

$$cluster\_score(a_i, a_j) = cosine(\langle \tilde{C}^\rho_{ii}, \tilde{C}^\rho_{ij} \rangle, \langle \tilde{C}^\rho_{ji}, \tilde{C}^\rho_{jj} \rangle)$$

Example

$$\delta(l_1) = \{l_1^{0.6}, l_2^{0.4}, l_3^0, l_4^0, l_5^0\}$$
$$\delta(l_2) = \{l_1^{0.4}, l_2^{0.55}, l_3^{0.05}, l_4^0, l_5^0\}$$
$$\mathrm{d} = \delta(l_1) \sqcap \delta(l_2) = \{l_1^{0.4}, l_2^{0.4}, l_3^0, l_4^0, l_5^0\}$$
$$support(\mathrm{d}) = \{l_1, l_2\}$$
$$score(\mathrm{d}) = \sum_{j=1}^{|L|} \mathrm{d}^j = 0.8$$
$$cluster\_score(l_1, l_2) = 0.999$$

Experimental Results

Datasets

- $10,108$ games from the community website http://spawningtool.com with $3,805$ avatars

Validation with a ground truth

- Several supervised classification methods used
- As there exists no ground truth, we automatically inserted (SUG) and identified (URL, NAMES) a controlled number of avatar aliases
- High recall measures for KNN and J48
- Good MAP and AUC values for all classifiers
- High precision when considering the top of the ranking (as the number of candidates is very large)

Results

- Several pairs are detected as false positives with a high score: these are the pairs we are looking for!
- Example: avatars *EGaLive* and *aLiveRC* belong to the same player (expert knowledge)

Parameters: $\gamma = 0.2$, $\theta = 20$, $\lambda = 0.9$, $\tau = 90$

Surrogates

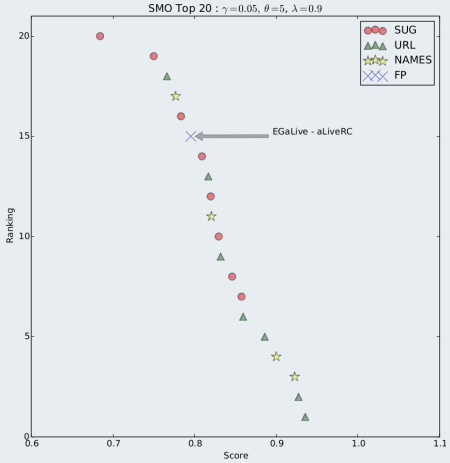
Classifier	F1	MAP	Recall	AUC	Precision	P@10
J48	0.468	0.824	0.805	0.904	0.33	1.0
nBayes	0.226	0.740	0.390	0.915	0.16	0.8
SMO	0.312	0.971	0.536	0.993	0.22	1.0
KNN	0.567	0.822	0.976	0.882	0.4	0.9

Surrogates & URLS

Classifier	F1	MAP	Recall	AUC	Precision	P@10
J48	0.588	0.907	0.606	0.866	0.57	1.0
nBayes	0.443	0.857	0.457	0.864	0.43	1.0
SMO	0.257	0.912	0.266	0.945	0.25	1.0
KNN	0.670	0.937	0.691	0.874	0.65	1.0

Surrogates & URLS & Names

Classifier	F1	MAP	Recall	AUC	Precision	P@10
J48	0.689	0.983	0.606	0.935	0.8	1.0
nBayes	0.560	0.943	0.492	0.906	0.65	1.0
SMO	0.258	0.949	0.227	0.960	0.3	1.0
KNN	0.758	0.967	0.667	0.792	0.88	1.0



E. Q. Yan, J. Huang, G. K. Cheung. *Masters of Control: Behavioral Patterns of Simultaneous Unit Group Manipulation in StarCraft 2.* Proc. of the 33rd ACM Conf. on Human Factors in Computing Systems (CHI), pp. 3711–3720, 2015.

M. Kaytoue, A. Silva, L. Cerf, W. Meira Jr., and C. Raïssi. *Watch me playing, I am a professional: a first study on video game live streaming.* Companion Proc. of the 21st Int. Conf. on World Wide Web, pp. 1181–1188, 2012.

B. Ganter, S. O. Kuznetsov. *Pattern Structures and Their Projections.* Proc. of the 9th Int. Conf. on Conceptual Structures (ICCS), Springer, 129–142, 2001.

T. L. Taylor. *Raising the Stakes: E-Sports and the Professionalization of Computer Gaming*, MIT Press, 2012.

This research has been partially funded by the French National Project FUI AAP 14 Tracaverre 2012-2016.

Formalisation et mise en œuvre de méthodes heuristiques de fouille de données massives et hétérogènes

Guillaume Bosc, Mehdi Kaytoue, Jean-François Boulicaut

2ème année de thèse, Financement Contrat Doctoral, Équipe DM2L
guillaume.bosc@liris.cnrs.fr – http://liris.cnrs.fr/membres?idn=gbosc

Résumé de la thèse

La fouille de données massives et hétérogènes constitue un nouveau défi que ce soit du point de vue algorithmique, pour le traitement des données massives, ou de l'aspect de la méthode, pour gérer la variété des types des données. Les algorithmes heuristiques sont une solution au contexte *big data*, tandis que la prise en compte de données hétérogènes peut être abordée par l'élaboration d'une mesure de qualité spécifique ainsi que d'une exploration originale et efficace de l'espace de recherche.

Introduction

De plus en plus de données mettent en œuvre des objets décrits dans plusieurs bases de données, issues de différents domaines, e.g., des molécules odorantes peuvent à la fois être décrites par des propriétés physicochimiques, mais également par les odeurs auxquelles elles sont associées.L'objectif de mes travaux de thèse est alors de développer des approches *génériques* de fouille de données permettant, à partir de tels jeu de données, d'extraire des sous-ensembles d'objets (e.g., des molécules) pour lesquels plusieurs univers de description correspondent. De plus, le rapport expressivité/calculabilité est primordial dans de telles approches : plus l'expressivité du langage utilisé dans les descriptions est riche, plus les résultats obtenus peuvent être bons du point de vue de la mesure de qualité mais plus la complexité algorithmique est importante.

Etat de l'art

Plusieurs techniques de fouille de données permettent, en partie, d'aborder la problématique posée par ce sujet de thèse. Les méthodes de *découverte de sous-groupes*, mais aussi de *redescription mining* et les techniques de *parallel universes* traitent de l'extraction de motifs à partir de plusieurs univers de description en utilisant des langages plus ou moins expressifs. Cependant, ces approches ne sont pas toutes adaptées aux données massives et hétérogènes, c'est à dire décrites par différents types de descripteurs (numériques, booléennes, graphes ou séquences).

Pour faire face aux données massives, plusieurs algorithmes heuristiques ont été développés pour n'explorer que partiellement l'espace de recherche en se basant sur des hypothèses mathématiquement fondées afin de garantir la découverte d'un maximum global (ou local dans le pire des cas). Ces algorithmes heuristiques sont basés sur une mesure évaluant la qualité d'un résultat obtenu. Le choix de cette mesure de qualité est alors déterminant car elle influe directement sur le type de résultats obtenus.

Premières approches

Les premières approches réalisées pour aborder ce problème consistent à la mise en place d'une exploration heuristique classique : la recherche en faisceau (*beam-search*). Cette technique heuristique permet de simplement explorer les candidats les plus prometteurs de l'arbre de recherche et ainsi abandonner l'exploration des candidats jugés les moins bons [1]. Nous avons également défini une nouvelle mesure de qualité permettant de traiter le cas où un des univers de description est composé d'attributs booléens dont certains sont sur-représentés et d'autres sont sous-représentés parmi l'ensemble des objets (résultat à soumettre). Ce cas particulier correspond, dans la littérature, aux jeux de données multi-labels dont les distributions sont déséquilibrées. D'autre part, nous avons travaillé sur la fouille de motifs de type *séquence* appliquée à la découverte de stratégies discriminantes dans le domaine du sport électronique [2].

Conclusion

Les travaux futurs à réaliser concernent différents aspects du problème. L'heuristique *beam-search* ne permet pas une exploration efficace et pertinente de l'espace de recherche et est inadaptée dans le cas de données massives. De plus, nous souhaitons mettre en place une exploration permettant d'utiliser le plus efficacement les différents types de données à disposition (représentations 2D et 3D des molécules, projections, etc...) ainsi que l'ontologie du domaine.

Références

[1] G. Bosc, M. Kaytoue, M. Plantevit, F. De Marchi, M. Bensafi, and J.F. Boulicaut. Vers la découverte de modèles exceptionnels locaux : des règles descriptives liant les molécules à leurs odeurs. In *EGC*, 2015.

[2] G. Bosc, M. Kaytoue, C. Raïssi, J.F. Boulicaut, and P. Tan. Mining Balanced Patterns in Real-Time Strategy Games. In *ECAI*, pages 975–976, 2014.

OlfaMining

Comment caractériser l'identité d'une odeur ?

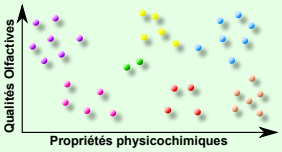
Moustafa Bensafi[1], Guillaume Bosc[2], Fabien De Marchi[2], Mehdi Kaytoue[2], Roland Kotto Kombi[2], Marc Plantevit[2]

[1]Centre de Recherche en Neurosciences de Lyon, France [2] Université de Lyon, LIRIS CNRS, France

Motivations

L'Olfaction : un processus complexe . . .

- Capacité à percevoir des odeurs.
- Existence de liens entre les propriétés physicochimiques et les qualités olfactives des molécules [1,2].

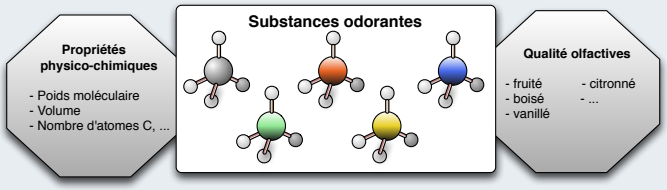


. . . **dont la compréhension a des enjeux en**

- Recherche fondamentale en neurosciences.
- Industrie (agroalimentaire, parfumerie, . . .).
- Santé (anosmie, . . .).

⇝ Comment **caractériser** et **décrire** le lien existant entre les **propriétés physicochimiques** d'une molécule et ses **qualités olfactives** ?

Matériel et Méthodes



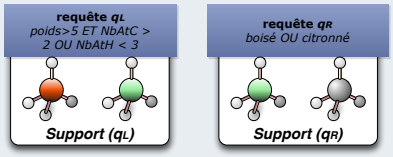
Dravnieks
- 138 molécules
- 4885 propriétés
- 146 qualités

Boelens
- 263 molécules
- 4885 propriétés
- 30 qualités

Arctender
- 1689 molécules
- 1704 propriétés
- 74 qualités

Fouille de redescriptions (*redescription mining* [3])

Principe *Chercher des descriptions ou requêtes dans chacune des vues (propriétés et qualités) qui couvrent presque les mêmes substances odorantes.*



- **Redescription** (q_L, q_R): deux requêtes définies sur des langages à expressivité variée ($\vee$, $\wedge$, $\neg$, ...)
- **Précision** : coefficient de Jaccard à maximiser

$$\mathcal{J}(q_L, q_R) = \frac{|Support(q_L) \cap Support(q_R)|}{|Support(q_L) \cup Support(q_R)|}$$

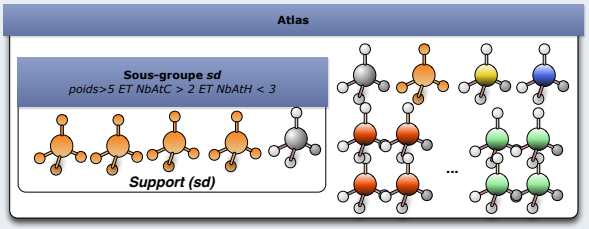
- **Test statistique** : avec $p_L = \frac{|supp(q_L)|}{|\mathcal{O}|}$ et $p_R = \frac{|supp(q_R)|}{|\mathcal{O}|}$

$$pval(q_L, q_R) = \sum_{k=|supp(q_L)\cap supp(q_R)|}^{|\mathcal{O}|} \binom{|\mathcal{O}|}{k} (p_L p_R)^k (1 - p_L p_R)^{|\mathcal{O}|-k}$$

- **Algorithme** : approche heuristique (*beam-search*)

Découverte de sous-groupes (*Subgroup discovery* [4])

Principe *Trouver et décrire des sous-groupes de molécules odorantes statistiquement caractéristiques d'une (ou plusieurs) qualité(s) d'odeur.*

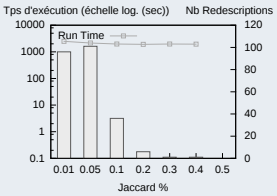


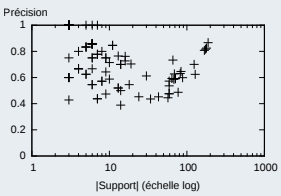
- **Sous-groupe :** décrit par une conjonction de paires attribut-valeur, supportée par un ensemble de molécules DESCRIPTION-PHYSICO-CHIMIQUE ⟶ QUALITÉ D'ODEUR
- **Précision :** quantifie la divergence entre la distribution des valeurs de la projection du sous-groupe et du jeu entier sur l'espace de modèles (divergence de Kullback-Leibler)
- **Algorithme :** approche heuristique (*beam-search*) due à la taille exponentielle de l'espace de recherche (comme pour la fouille de redescriptions)

Résultats

r_1 = (**VANILLE**, [19.403 ≤ MV ≤ 19.5106] OU [1.267 ≤ VE2_X ≤ 1.292] ET [11.574 ≤ MP ≤ 14.625] ET [1.511 ≤ IC3 ≤ 3.461] OU [3.342 ≤ VR3_X ≤ 3.342] ET [10.0 ≤ D/DTR11] ET [2.949 ≤ SPPOSLOG_H2 ≤ 4.385])

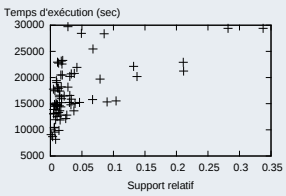
⇝ Support : 18, Similarité : 0.7.

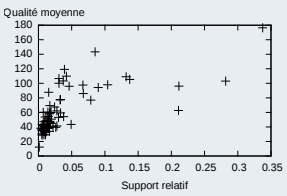




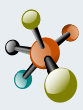
95% DES MOLÉCULES SATISFAISANT LA CONDITION SE07 > 5.86500 ET SV > 6.84200 ET XLOGP < 1.90000 ET X0SOL < 2.29750 ET VE1_L < 1.35400 ET VE2_X < 1.33100 ET IC3 < 3.78100 ET SV99 < 0.32250 SONT ASSOCIÉES À LA **VANILLE**

⇝ Support : 20.





Conclusion & Perspectives



- Prise en compte des représentations 2D et 3D des molécules.
- Réduction des temps de calculs pour des langages expressifs (heuristiques, parallélisation, . . .).

[1] K. Kaeppler and F. Mueller. Odor classification: a review of factors in influencing perception-based odor arrangements. Chemical senses, 38(3):189-209, 2013.
[2] C. Sezille and M. Bensafi. De la molécule au percept. Biofutur, (346):24-26, 2013.
[3] E. Galbrun, P. Miettinen: From black and white to full color: extending redescription mining outside the Boolean world. Statistical Analysis and Data Mining 5(4): 284-303 (2012).
[4] P. K. Novak, N. Lavrac, G. I. Webb: Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. Journal of Machine Learning Research 10: 377-403 (2009)

TRAVAIL RÉALISÉ DANS LE CADRE DU PROJET TRANSVERSE LIRIS OLFAMINING ET DU PROJET DEFISENS MI CNRS.

UN GRAND MERCI à toute l'équipe organisatrice de la Journée des Thèses 2015 bis

Comité de Direction
Marie-Neige Chapel et Rubiela Carrillo Rozo

Comité d'Organisation
Abdoulaye Diakité, Joseph Garnier et Hélène Perrier

Comité de Communication / Web
Carine Touré et Romain Deville

Comité de Relecture
Marie-Neige Chapel, Rubiela Carrillo Rozo, Cheick Hito Kacfah Emani, Sébastien Dufromentel, Jérémy Levallois, Awa Diattara, Julien Salotti, Matthieu Giroux, Ouadie Gharroudi et Maxime Gasse

Un grand merci également à Marie-Neige Chapel pour la création de l'affiche officielle