CADDY : Entreposage, fouille et visualisation de séries temporelles de trafic routier en milieu urbain

Projet ACIMD 2003 CADDY - Ecole Centrale Paris, laboratoire MAS et Université Paris Dauphine, laboratoire LAMSADE. Autres partenaires : ENST et INRETS

Olivier Carles, Florian De Vuyst, Georges Hébrail, Bernard Hugueney, Marc Joliveau, Geneviève Jomier, Maude Manouvrier, Yosr Naija, Raphael Rispoli, Gérard Scémama et Laurent Steffan

Journées PaRISTIC – Panorama des Recherches Incitatives en STIC – LaBRI Bordeaux, 21-23 novembre 2005

Nous contacter :

Florian De Vuyst florian.de-vuyst@ecp.fr,

Geneviève Jomier genevieve.jomier@dauphine.fr

Georges Hébrail hebrail@enst.fr

1 - Thèmes

- Trafic routier en réseau urbain
- Entreposage de données
- Fouille de données
- Séries temporelles multivariées
- Gestion des flux de données
- Apprentissage et modèles dynamiques
- Visualisation avancée
- Supervision et aide à la décision

2 - Objectif

- Définir des schémas de données pour des données temporelles définies sur un réseau spatial (routier)
- Définir des requêtes performantes sur ces données
- Proposer des algorithmes de fouille de données
- Contrôler les temps d'analyse des flux de données
- Trouver de nouveaux concepts de visualisation pour l'exploration et l'aide à la décision
- Valider notre méthode sur des données réelles
- Fournir un prototype logiciel

7 - Perspectives

Finalisation du projet

- Restitution de l'information adaptée à l'usage
- Analyse en ligne avec adaptation des temps d'analyse
- Nettoyage de données spatiotemporelles
- Gestion des requêtes en continu
- Visualisation avancée via VTK
- Réalisation d'un prototype logiciel

Après le projet

- Projet de collaboration avec les pays du Sahel sur des donnés d'hydrologie
- Projet de continuation du projet avec l'INRETS (trafic routier)
- Projet de collaboration avec le CEA
- Projet de collaboration avec EDF

Autres champs d'intérêt

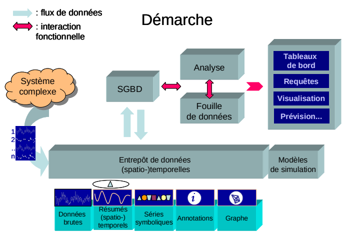
- Processus industriels et chaîne logistique
- Communauté HPC pour la recherche de caractéristiques dans des gros volumes de résultats de simulation numérique
- Sécurité informatique et détection d'intrusion

3 - Existant

- **Analyse spatio-temporelle**
  1) F. Duchêne, C. Garbay & V. Rialle, *Similarity measure for heterogeneous multivariate time-series*, http://www-sop.inria.fr/orion/personnel/Florence.Duchene/publications/FDuchene_Eusipco_2004.pdf
  2) M. Sebag, N. Tarrisson, O. Teytaud, J. Lefevre & S. Baillet, *A Multi-Objective Multi-Modal Optimization Approach for Mining Stable Spatio-Temporal Patterns.*, http://eprints.pascal-network.org/archive/00000994/01/1077.pdf
  3) Chiu, B. Keogh, E., & Lonardi, S. "Probabilistic Discovery of Time Series Motifs". In the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 24 - 27, 2003. Washington, DC, USA. pp 493-498.
  4) Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, Sharad Mehrotra:"Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases"ACM SIGMOD Record, 30(2), 151-162.
  5) D. Ganesan *et al., Coping with irregular spatio-temporal sampling in sensor networks*, http://berkeley.intel-research.net/sylvia/irregularity.pdf
  6) N. Mamoulis *et al., Mining, Indexing, and Querying Historical Spatiotemporal Data*, http://www.cs.bu.edu/~gkollios/Papers/kdd04-periodic.pdf
  7) F. Mörchen & A. Ultsch, *Mining Hierarchical Temporal Patterns in Multivariate Time Series*, http://www.mathematik.uni-marburg.de/~databionics/en/downloads/papers/moerchen04mining.pdf
  8) A.A. Freitas, *A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery*, http://citeseer.ist.psu.edu/freitas01survey.html
  9) W.S. Torgerson, *MultiDimensional Scaling: I.* Theory and method, Psychometrica, vol. 17, pp. 401--419, 1952
  10) S.H. Jeong *et al., A Generic Algorithmic Framework for Aggregation of Spatio-Temporal Data*, http://citeseer.ist.psu.edu/692918.html
- **Stockage de données spatio-temporelles**
  1) J.F. Roddick *et al., Spatial, Temporal and Spatio-Temporal Databases - Hot Issues and Directions for PhD Research*, http://www.cs.ust.hk/~dimitris/PAPERS/SIGREC04.pdf
  2) M. Hadjieleftheriou *et al.,Complex Spatio-Temporal Pattern Queries*, http://cs-people.bu.edu/marioh/papers/vldb05-stp.pdf
  3) S. Gupta *et al., Roads, Codes, and Spatiotemporal Queries*, http://www.sigmod.org/sigmod/pods/proc04/pdf/P-12.pdf
- **Flux de données**
  1) B.Babcock, S.Babu, M.Datar, R.Motwani, J.Widom, *Models and issues in data stream systems*, ACM PODS 2002, http://dbpubs.stanford.edu/pub/2002-19
  2) M.Garofalakis, J.Gehrke, R.Gastogi, *Querying and mining data streams: you only get one look*, VLDB'02 Tutorial, http://www.bell-labs.com/user/rastogi/streams-tutorial.ppt
  3) L.Golab, M.Tamer Özsu, *Issues in data stream management*, SIGMOD Records, Vol.32, n°2, Juin 2003.
  4) C. Giannella *et al., Mining Frequent Patterns in Data Streams at Multiple Time Granularities*, http://maids.ncsa.uiuc.edu/documents/readings/fpstm03.pdf
  5) S. Papadimitriou *et al., Streaming Pattern Discovery in Multiple Time-Series*, http://www.db.cs.cmu.edu/Pubs/Lib/spirit_vldb05/spirit_vldb05.pdf
  6) J. Han, *Mining unusual patterns in data streams*, http://www.cs.umbc.edu/NGDM02/slides/ngdm02.ppt
- **Visualisation**
  1) X. Li, *New Methods of Visualization of Multivariable Spatio-temporal Data, Thesis (2005)*, http://www.itc.nl/library/Papers_2005/msc/gfm/xia.pdf
  2) T. Duce *et al., Visualization of Historical Wildfire Data: Application of a DX-Oracle Interface*, http://opendx.npaci.edu/publications/environmental/mtwildfire/index.html
  3) S. Shekhar *et al., High-performance visualization of Spatial traffic data*, http://www.cs.umn.edu/research/shashi-group/alan_planb.pdf
  4) A. Debrégeas , G. Hébrail, *Interactive interpretation of Kohonen maps applied to curves,International* Conference on Knowledge Discovery and Data Mining (KDD'98), New-York, Août 1998.

4 - Démarche

- Réseau routier d'une ville moyenne = 420 capteurs
- 400 000 données élémentaires par jour

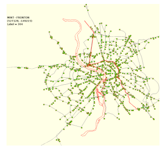


Spécificités

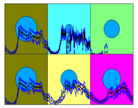
- Caractère (spatio)-temporel ou temporel multivarié des données
- Nettoyage de données spatio-temporelles
- Algorithmes intégrés de résumés (spatio)-temporels
- Multi-représentation
- Nouveaux concepts de visualisation de données
- Réseau ou graphe d'interaction sous-jacent

Site web du projet CADDY : http://norma.mas.ecp.fr/wikimas/Caddy

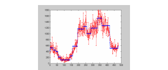
6 – Exemples de résultats



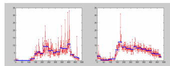
Réseau routier urbain considéré : Arcs, nœuds, axes, capteurs, fleuve, canaux



La figure ci-dessus montre le résultat d'une classification automatique d'extraits de séries temporelles, dont la durée est d'une journée, par carte de Kohonen. Chaque case du tableau correspond à une classe trouvée automatiquement : la surface du disque au centre de la case indique l'effectif de la classe, la courbe noir représente le prototype de celle-ci alors que les courbes rouges indiquent la dispersion de celle-ci en représentant les individus correspondant à une fois et demi la variance de la classe. Les classes trouvées par classification automatique ont ensuite été regroupées manuellement par un expert du domaine, suivant des classes indiquées par la couleur des cases. Cette intervention n'est pas fastidieuse car seul un nombre très restreint de classes présentées sous la forme d'un tableau est à considérer par l'expert qui n'a donc pas à évaluer chacune des très nombreuses courbes.

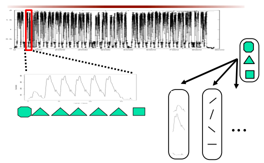


(a) Flux de véhicules par heure pour un capteur

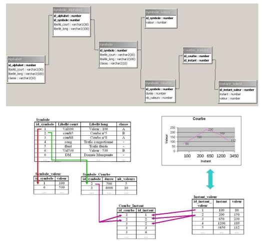


(b) Taux d'occupation pour deux capteurs différents (%)

Résumés PAA (Piecewise Aggregate Approximation) par algorithme original de segmentation automatique



La figure ci-dessus illustre des résultats de représentations symboliques de longues séries temporelles. Sur la droite, on trouve différents exemples d'alphabets de symboles qui correspondent soit à des formes typiques de durée fixe (ici journalières), soit à des classes de paramètres de modèles (ici des classes de pentes de modèles linéaires d'ordre 1). Les classes de comportements journaliers sont trouvées automatiquement et peuvent ensuite être étiquetées (ici des activités "normale", "faible" et "nulle"). De même, les classes de variations (pentes) permettant de décrire au mieux les séries temporelles sont trouvées automatiquement (ici "augmentation modérée", "forte augmentation", "baisse modérée", "stabilité"). Sous la série temporelle, on a représenté l'extrait correspondant à la semaine entourée en rouge. L'extrait de la série temporelle est encore constitué de centaines de points, alors qu'il est correctement représenté par sept épisodes journaliers associés à des symboles issus d'un alphabet de trois symboles.



5 - Contributions

AGREGATION DE SERIES TEMPORELLES
Résumés symboliques pour la fouille de données

Représentations symboliques pour la fouille de données appliquée à des bases de données de séries temporelles

La fouille de données est l'extraction d'informations à partir de très gros volumes de données.

Les représentations symboliques associant des épisodes définis sur le domaine temporel à des éléments d'un alphabet symbolique permettent de répondre aux deux aspects de la problématique.

1. **Gros volumes de données**
   - **Très importante réduction de dimension obtenue en regroupant des points en épisodes**. En pratique, les phénomènes mesurés peuvent l'être à des échelles de l'ordre de quelques minutes. Cependant les comportements intéressants ont souvent des durées bien supérieures. Le regroupement des instants de mesure en épisodes ayant une cohérence sémantique permet alors de réduire la quantité d'informations à traiter de plusieurs ordres de grandeur.
   - **Possibilité d'utiliser des algorithmes combinatoires.** L'utilisation de symboles appartenant à un alphabet symbolique de cardinalité très réduite (de l'ordre de la dizaine) permet d'effectuer facilement de l'optimisation combinatoire.
2. **Manipulation d'unités d'information à un niveau d'abstraction supérieur**
   - **Construction automatique de classes d'épisodes**. De l'information est extraite lors de la construction de l'alphabet symbolique par la découverte de comportements caractéristiques. Les connaissances a priori du domaine peuvent intervenir de différentes façons, par exemple pour spécifier les caractéristiques suivant lesquelles effectuer la classification (niveau, pente, comportement sur une période donnée,etc.). Les différents algorithmes de construction de représentations symboliques permettent alors de découvrir des classes de comportement permettant de décrire les séries temporelles auxquelles l'utilisateur peut associer une sémantique propre au domaine.
   - **Interaction avec les experts du domaine.** L'utilisation d'une représentation symbolique permet aux utilisateurs d'interagir avec les modèles pour enrichir ceux-ci par leur connaissance du domaine. En effet, la sémantique associée aux symboles rend possible l'expression de connaissances extrinsèques afin d'ajuster les modèles prédictifs basés sur des représentations symboliques.

MODELES DYNAMIQUES
Résumés par double ACP espace-temps et prévision court-terme

Double Analyse en Composantes Principales espace + temps et matrices résumés

Désignation : M : échantillonnage en temps ; N : degré d'approximation en temps ; J : dimension spatiale ; K : degré d'approximation en espace.

Une série spatiotemporelle discrète S multivariée sur [0,T] est une matrice dans $M_{JM}(\mathbb{R})$. On suppose N<<M, K<<J et K de l'ordre de N.

Etant donné un ensemble de jours correspondant au même jour de la semaine, on effectue N clichés aléatoires correspondant à différentes séries temporelles univariées $S(x_j)$, où les capteurs x_j sont tirés aléatoirement. Ces N prises de vue permettent de fabriquer une famille orthonormée $(\phi^n)_{n=1,N}$ de vecteurs d'analyse dans $\mathbb{R}^M$. Etant donné un ensemble de jours correspondant au même jour de la semaine, on effectue K clichés aléatoires correspondant à différents vecteurs d'état $S(t^m)$. Ces K clichés permettent de fabriquer une famille orthonormée de K vecteurs propres $(\psi_k)_{k=1,K}$ de $\mathbb{R}^J$. On note $P=col(\phi^n)_n \in \mathcal{M}_{MN}$ et $Q=col(\psi_k)_k \in \mathcal{M}_{JK}$. La double ACP espace-temps consiste à appliquer une ACP spatiale et une ACP temporelle, c'est-à-dire

$$\Pi S = QQ^T SPP^T.$$

La matrice A de petite taille (K×N) coefficients $A = Q^T S P$ peut être vue comme le résumé spatiotemporel de la série d'origine S. On l'appelle la matrice-résumé de S.

Classification des matrices résumés par cartes auto-organisatrices (cartes SOM)

On calcule d'abord une matrice-résumé pour chaque série temporelle multivariée journalière (tous les capteurs sont pris en compte). On effectue ensuite une classification des matrices-résumés pour chaque jour de la semaine (lundi, mardi, ...). Une carte SOM constituée d'une cinquantaine de représentants permet d'identifier la majorité des situations habituelles ainsi que les situations atypiques pour une journée donnée de la semaine.

Algorithme de prévision court-terme

On suppose que l'instant présent est t^m. On se donne une fenêtre temporelle $[t^{(m-M_w)}, t^m]$ pour l'identification du processus, M_w étant un entier définissant la fenêtre temporelle, $m-M_w>0$, $M_w << M$. A partir de la fenêtre temporelle, on définit une matrice diagonale de fenêtre temporelle χ de coefficients 0 ou 1. On recherche alors la ou les matrices-résumés, représentantes dans la carte SOM, qui réalisent le minimum

$$\min_{A \in \mathcal{M}_{KN},\ A \in SOM} \frac{1}{2} \| \left(S - QAP^T \right) \chi \|^2 \quad (1)$$

La série temporelle multivariée prédite est alors

$$S^{pred}(t^p) = S(t^m) - (Q\bar{A}P^T)(t^m) + (Q\bar{A}P^T)(t^p)$$

où A dans la carte SOM réalise le minimum..

CLASSIFICATION
Classe d'épisodes et classes sémantiques

Classification d'épisodes extraits de séries temporelle

Les séries temporelles analysées sont souvent issues de la mesure de phénomènes en rapport avec des activités humaines. Celles-ci présentent alors des périodicités liées aux cycles d'activité que sont les jours et les semaines. Il apparaît alors pertinent de découper les séries temporelles en épisodes dont la durée peut être déteminée à l'aide de connaissances a priori ou bien par analyse spectrale. Ces épisodes de même taille pouvant être considérés comme des éléments d'un même espace vectoriel, il est facile d'en effectuer des classifications automatiques suivant une distance euclidienne.

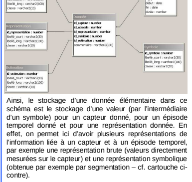
Lorsqu'il s'agit de créer des classes d'épisodes associées à une sémantique du domaine, l'utilisateur final doit valider et annoter la classification. Bien que deux extraits de séries temporelles proches au sens de la distance euclidienne soient généralement associés à une même classe d'équivalence suivant la sémantique du domaine, la réciproque n'est pas forcément vraie (deux extraits de séries temporelles peuvent appartenir conceptuellement à une même classe d'équivalence sans être proches suivant cette distance point à point).

Dans ce dernier cas, il importe de permettre à l'utilisateur de constituer les classes "sémantiques" en regroupant des classes trouvées automatiquement suivant un critère de minimisation de distance euclidienne intraclasse. C'est dans cette optique que nous avons utilisé des cartes auto-organisatrices de Kohonen suivant une topologie en deux dimensions permettant une visualisation des classes sous la forme d'un tableau.

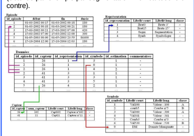
ENTREPOT DE DONNEES CADDY
Stockage Multi-Représentations de Données Temporelles Multivariées

Une approche générique sous forme de schéma en étoile

Le schéma global de l'entrepôt de données permet de stocker des données temporelles issues d'un ensemble de capteurs répartis géographiquement sur le territoire. L'entrepôt est stocké dans une base de données relationnelle (utilisation du logiciel ORACLE). La structure relationnelle retenue s'apparente à celle d'un schéma en étoile, solution classique pour le stockage d'un cube de données au sein d'une base de données relationnelle. Une table centrale (table 'Donnée' ci-dessous) stocke des éléments d'un produit cartésien suivant quatre dimensions principales : Capteur, Episode, Représentation, et Symbole (l'utilité de la dimension Estimation n'est pas détaillée ici).



Ainsi, le stockage d'une donnée élémentaire dans ce schéma est le stockage d'une valeur (par l'intermédiaire d'un symbole) pour un capteur donné, pour un épisode temporel donné et pour une représentation donnée. En effet, on permet ici d'avoir plusieurs représentations de l'information liée à un capteur et à un épisode temporel, par exemple une représentation brute (valeurs directement mesurées sur le capteur) et une représentation symbolique (obtenue par exemple par segmentation – cf. cartouche ci-contre).



Une gestion fine des épisodes temporels

Afin de faciliter les requêtes adressées à l'entrepôt de données CADDY, la structure de l'entrepôt intègre une gestion fine et pré-calculée des épisodes temporels. Un épisode temporel est défini à la base par une date de début et une date de fin. La relation d'inclusion des épisodes les uns dans les autres est précalculée. Les épisodes standard (jours, semaines, mois, années) font l'objet d'un traitement spécifique.

Interrogation efficace de l'entrepôt de données

La structure de l'entrepôt de données CADDY a été conçue pour répondre efficacement aux requêtes des utilisateurs. L'efficacité se décline à trois niveaux :

- Réponse à des requêtes proches des besoins utilisateurs, même lorsque ceux-ci se situent à un niveau d'abstraction élevé : ceci est réalisé par l'approche multi-représentations (en particulier avec les représentations symboliques).
- Réponse aux requêtes avec des bons temps de réponse : ceci est réalisé par le pré-calcul de nombreuses informations, certes au détriment du volume de stockage et de la rapidité du chargement des données.
- Facilité d'écriture des requêtes : ceci est réalisé à la fois par l'approche multi-représentations et par la gestion fine des épisodes temporels.

Gestion de représentations symboliques de données temporelles

L'entrepôt CADDY permet de stocker des données temporelles simultanément sous une représentation brute et sous une représentation symbolique. Ainsi, pour un même capteur, il est possible d'associer des symboles à des épisodes dont la longueur est différente des épisodes initialement mesurés. Aux symboles sont associés : (1) un libellé à forte connotation sémantique (par exemple 'Trafic congestionné'), (2) éventuellement une courbe type correspondant au symbole.

Les représentations symboliques sont calculées automatiquement à la demande de l'utilisateur ou bien à l'entrée des données brutes dans le système, par intégration des algorithmes de segmentation de courbes et de recherche de formes type journalières.

Exemples de requêtes typiques à l'entrepôt CADDY

La base de données utilisée en test correspond à des capteurs de trafic routier urbain (données fournies par l'INRETS). L'entrepôt de données permet de répondre efficacement aux requêtes typiques suivantes :

- *Quel était le taux d'occupation sur la ville X dans la rue y le 21 Janvier 2004 à 12h40 ?*
- *Déterminer les jours pendant lesquels le trafic a été congestionné sur le tronçon routier numéro 18 de la ville X*
- *Déterminer les épisodes ayant le même comportement quotidien au niveau du débit de voitures pour le capteur 17.*