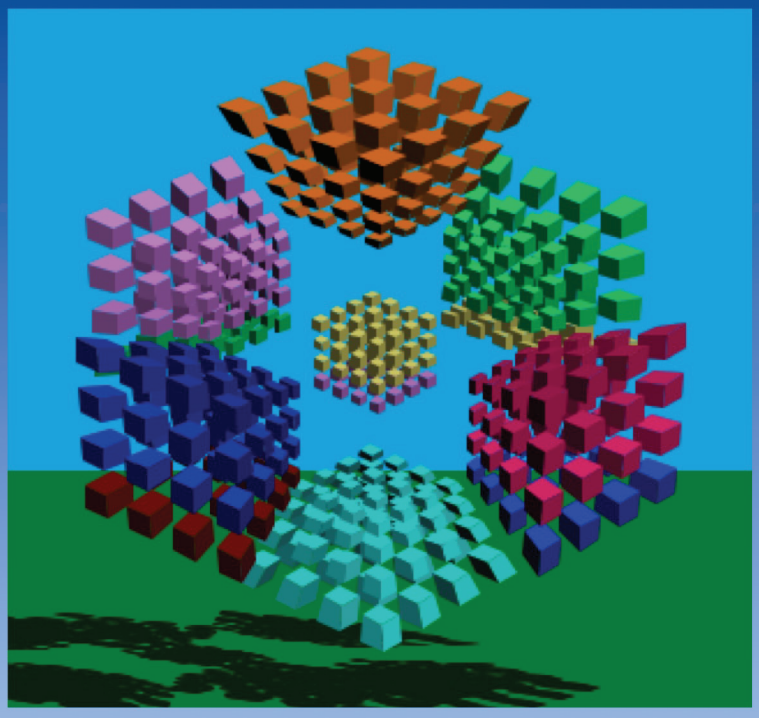
EGC'2013

13e Conférence Francophone sur l'Extraction et la Gestion de Connaissances

Université Paul Sabatier – IRIT – Toulouse

29 janvier 2013 – Journée Ateliers/Tutoriels



FOuille de données Spatio-Temporelles et Applications (FOSTA)

Université de Toulouse

Université Paul Sabatier TOULOUSE III

IRIT CNRS - INPT - UPS - UT1 - UTM

EGC

Atelier FOuille de données Spatio-Temporelles et Applications (FOSTA)

Organisateurs : Nazha Selmaoui-Folcher[1], Christophe Rigotti[2], Frédéric Flouvat[1]

[1] PPME-Labex Corail, Université de la Nouvelle-Calédonie

[2] LIRIS CNRS, EPC INRIA Beagle, INSA de Lyon

PRÉFACE

L'atelier FOSTA (FOuille de données Spatio-Temporelles et Applications) se situe dans la lignée des évènements organisés depuis plusieurs années au niveau international sur la fouille de données spatio-temporelles. Il a pour objectif d'offrir un espace de discussion et d'échanges pour les chercheurs de la communauté francophone travaillant sur des données spatiales et spatio-temporelles, et permettre le partage des derniers développements dans ce domaine.

Depuis quelques années, les avancées technologiques au niveau des capteurs, des techniques d'acquisitions (images satellitaires, capteurs sismiques, capteurs de température, ou encore de salinité, etc.) et la prolifération des appareils mobiles géolocalisés ont généré une grande quantité de données spatiales et spatio-temporelles. La collecte de ces masses de données géographiques et leur mise à disposition dans des systèmes d'information soulèvent de nouveaux défis. En effet, nous sommes face à de grandes bases données hétérogènes et multi-échelles, où la dimension spatiale est particulièrement importante. Les méthodes classiques (télédétection ou analyse spatiale) atteignent leurs limites pour traiter ces données et pour par exemple en extraire des tendances cachées.

Les informaticiens, géo-informaticiens et thématiciens collaborent de plus en plus étroitement pour fournir des solutions innovantes et efficaces. Ces travaux se développent notamment à partir de méthodes de fouille de données spatio-temporelles, qui s'avèrent indispensables ici, en permettant de tirer parti des informations de géolocalisation et des mesures acquises sous forme de séries temporelles.

Cette édition de l'atelier FOSTA encouragera, nous l'espérons, les réflexions sur les travaux dans le domaine de la fouille de données spatio-temporelles au sens large et de ses applications.

Nous tenons à remercier les auteurs pour leurs contributions, les membres du comité l'organisation d'EGC pour leur précieux support et les membres du comité de lecture pour leur participation. Nous sommes aussi très reconnaissant à Maguelonne Teisseire et Sandra Bringay, d'avoir accepté de donner une conférence invitée lors de cet atelier. Nous remercions aussi l'Agence Nationale de la Recherche (ANR) pour son soutien au travers du projet FOSTER (ANR-2010-COSI-012).

Nazha Selmaoui-Folcher
PPME-Labex Corail
Univ. de la Nouvelle-Calédonie

Christophe Rigotti
LIRIS CNRS
EPC INRIA Beagle
INSA de Lyon

Frédéric Flouvat
PPME-Labex Corail
Univ. de la Nouvelle-Calédonie

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Sandro Bimonte
Sandra Bringay
Bruno Crémilleux
Frédéric Flouvat
Thomas Guyet
Florence Le Ber
Nicolas Meger
René Quiniou
Marc Plantevit
Pascal Poncelet
Chrisophe Rigotti
Nazha Selmaoui-Folcher
Maguelonne Teisseire
Emmanuel Trouvé
Karine Zeitouni

TABLE DES MATIÈRES

Communications longues

Conférence invitée

Détection automatique de « faits marquants » dans la presse écrite

Ludovic Moncla*, Mauro Gaio*
Alain du Boisdulier**

*Laboratoire LIUPPA, BP-1155, 64013 PAU Université Cedex
prenom.nom@univ-pau.fr,
**DIS, Cré@ticité C - Technopole Izarbel, 64210 Bidart
adb@docimsol.com

Résumé. A l'heure des réseaux sociaux et des flux continus d'informations, la recherche d'information ciblée dans le domaine de l'actualité peut s'avérer complexe pour un utilisateur.

Nous proposons dans cet article une chaîne de traitements opérant sur un flux et permettant de détecter automatiquement les « faits marquants » de l'actualité. Notre chaîne s'appuie sur une méthode de classification non-supervisée des articles en provenance de la presse, utilisant des techniques de regroupement d'après des similarités spatiales, temporelles et thématiques.

Nous utilisons la notion de « faits marquants » afin de définir un événement ou un ensemble d'événements comme ayant une certaine redondance dans un ensemble de sources données et pour une fenêtre temporelle donnée.

Nous nous appuyons sur les méthodes du « Topic Detection and Tracking », ou encore du « New Event Detection » afin d'obtenir notre contribution et nous utilisons pour nos expérimentations une implémentation du « Latent Dirichlet Allocation ». La combinaison d'une méthode de clustering thématique à un classement spatial dans un intervalle temporel, nous permet de faire émerger les événements de l'actualité selon ces trois composantes représentatives.

1 Introduction

D'après une étude [1] menée aux Etats-Unis, Internet apparaît comme le deuxième média de consultation de news. Une autre étude, publiée par Yahoo en 2010 [2] montre que plus de 9 internautes sur 10 réalisent chaque mois des requêtes sur un moteur de recherche, dont près de 80% pour s'informer de l'actualité (politique, économie, société, etc).

L'information d'actualité issue d'agence de presse est non-structurée et évolutive, une information faisant l'actualité aujourd'hui ne le sera peut être plus dans deux mois. La validité d'une information d'actualité est donc par définition associée à une fenêtre temporelle définie

1. Pew Research en décembre 2010
2. Yahoo ! Search Academy 2010 $http://www.flickr.com/photos/yahoo_presse/4477558968/$

par un début t_0 et une fin t_n et à un contexte, celui-ci pouvant être thématique, spatial ou les deux.

Nous utilisons la notion de « faits marquants » afin de définir une information comme ayant un intérêt particulier dans une fenêtre temporelle donnée [t_0, t_n]. Dans le cadre d'une information d'actualité, cette notion représente l'intérêt collectif porté sur un événement ou sur un ensemble d'événements liés. La détection de « faits marquants » peut être ramenée à une démarche de détection et de classification d'événements. Les faits marquants peuvent être reliés entre eux par des relations thématiques, temporelles ou spatiales et peuvent être catégorisés par thématiques générales (politique, économie, culture, etc) ou par événements (guerre en Irak, tsunami au Japon, etc) pouvant eux-même être localisés géographiquement et temporellement.

Pour une fenêtre temporelle et une échelle spatiale données, un événement devient un « fait marquant », uniquement s'il est traité par un ensemble important d'articles d'actualité provenant de sources différentes. La fenêtre temporelle détermine la durée durant laquelle un événement est considéré comme un « fait marquant ».

Nous utilisons une définition d'un événement proche de celle donnée par Saval A. (2009), qui définit un événement comme la combinaison d'une propriété sémantique, d'un intervalle temporel et d'une entité spatiale. Nous généralisons la propriété sémantique par ce que nous appelons la composante thématique, et celle d'entité spatiale par un ou des espaces englobants à différentes échelles. Cela nous permet d'étendre la notion d'événement. Nous ne souhaitons pas extraire chaque sous-événement comme le fait Serrano et al. (2012) mais au contraire nous souhaitons conserver l'événement dans sa globalité, qu'il soit localisé à différents instants ou en différents lieux. La notion de validité d'une information d'actualité met en avant comme clef initiale la composante temporelle dans le traitement de données , cependant ce n'est pas la seule composante à prendre en considération. D'après Wegener (2000) tout événement se produit en un lieu et à un moment donné, précisant que la composante spatiale apparaît alors comme essentielle dans la description d'un événement. La localisation temporelle d'un événement dans l'actualité peut se révéler complexe. Doit-on considérer un événement comme faisant partie de l'actualité si la date de parution de l'article est en relation avec la fenêtre temporelle, ou bien si les données et relations temporelles présentes au sein du contenu de l'article relient l'événement à cette fenêtre ? D'autre part des études montrent l'importance de la composante spatiale dans la recherche d'information. D'après Hill (2006), 70% des documents textuels contiennent des références à des lieux géographiques.

Nous proposons de prendre également en compte la composante thématique pour venir renforcer, le cas échéant, les 2 autres composantes et permettre la compréhension et l'analyse d'une information protéiforme que constitue l'actualité.

La section 2 fait un état de l'art des méthodes existantes pour la détection, le suivi de sujets et le regroupement en classes de documents textuels. La section 3 présente notre contribution afin de répondre à la problématique de détection entièrement automatisée de « faits marquants ». Enfin la section 4 propose l'expérimentation de notre chaîne de traitement sur un corpus d'actualités.

2 État de l'art

Notre objectif est de permettre dans un flux continu de regrouper automatiquement les articles de presse en fonction de thèmes similaires, selon un périmètre temporel et spatial afin de faire émerger les « faits marquants » de l'actualité. De la même façon que Bossard et Poibeau (2008), nous avons décidé de regrouper les articles par classes d'événements.

Nous décomposons le processus de détection de « faits marquants » en différentes étapes, faisant intervenir différentes techniques. Une première étape consiste à classer les articles pour chaque unité temporelle selon des critères spatiaux. Puis une deuxième étape, la *classification thématique*, permet de regrouper les articles par similarités en *classes d'événements*. Le corpus que nous exploitons étant un corpus d'articles d'actualité, enrichi automatiquement au fil des heures, une troisième étape consiste à détecter l'arrivée d'un nouvel article et à indexer les faits ou événements qui le composent afin d'enrichir les classes d'événements existantes ou d'en créer de nouvelles.

2.1 Détection et suivi d'événements

La détection et le suivi d'événements font partie d'un sous-domaine de la recherche d'information. Il s'agit de repérer dans un flux continu de données un événement particulier (exemple : catastrophes naturelles). Les articles de presse sont particulièrement adaptés à ce type de détection et de recherche d'information. Un grand nombre d'études comme (AlSumait L., 2008), (Rigouste et al., 2006) ou (Blei et al., 2003) basent leurs expérimentations et leurs évaluations sur un corpus issu d'articles de presse Reuters.

2.1.1 Topic Detection and Tracking : TDT

Le TDT est un programme de recherche initié en 1997 par le DARPA (Allan et al., 1997). Il se décompose en 5 axes de recherche : la segmentation du flux de données (changement de topics), le suivi de topics, la détection de topics, la détection de nouveaux articles et enfin la détection de liens (topics liant deux articles).

La notion de topic se réfère à celle d'événement comme pour Wayne (1998). Ces axes de réflexion correspondent à notre problématique de recherche. Nous devons être capable, dans un flux continu d'informations, de détecter un événement important et représentatif de l'actualité puis de regrouper les articles traitant des mêmes événements.

2.1.2 New Event Detection : NED

La détection en-ligne d'événements est la classification d'un flux de données triées temporellement (Papka et Allan, 1998). Cette méthode fait appel au modèle vectoriel introduit par Salton et al. (1975), qui représente chaque document par un vecteur. Cela permet de calculer un coefficient de similarité entre les différents documents. Cette technique est largement utilisée en recherche d'information pour déterminer les documents les plus pertinents en fonction de la requête formulée par l'utilisateur.

La *Détection de Nouveaux Événements* étudiée par Allan et al. (1998) vise à détecter l'apparition d'un nouvel événement dans un flux de données. La NED fait parti de l'initiative TDT. Elle opère dans un processus en ligne (Papka et Allan, 1998) et traite chaque article un par un

lors de leur arrivée. L'algorithme peut être modélisé de la manière suivante (Allan et al., 1998) : construire une signature (résumé) puis déterminer un seuil afin de comparer les articles, de les associer ou d'identifier un nouvel événement.

Un nouveau document est ajouté à la classe la plus proche si la mesure de similarité entre le document et la classe existante est supérieure à un seuil donné, sinon une nouvelle classe est créée (Binsztok et al., 2004).

2.2 Regroupement et classification

Dans le contexte d'un processus entièrement automatique le regroupement d'articles se fait selon une méthode de classification non-supervisée (en anglais *clustering*), il s'agit d'un repérage de classes thématiquements homogènes dans un corpus textuel (Rigouste et al., 2006). Il existe différentes approches de classification non-supervisées. Dans Turenne (2001) elles sont regroupées en huit familles de deux types : les approches par partitionnement et les approches hiérarchiques.

Les techniques de classification par partitionnement segmentent l'espace en régions disjointes ou non selon les techniques employées. Les approches par hiérarchisation représentent le corpus de manière arborescente.

Des travaux comme Steinbach et al. (2000) ou Xiao (2010), se proposent de comparer et d'évaluer les différentes techniques de clustering de documents textuels, nous ferons ici simplement une présentation générale et succincte des techniques les plus utilisées, afin de déterminer laquelle correspond le mieux à nos besoins.

L'algorithme des *k-moyennes* (Forgy, 1965), est une des techniques de classification non supervisée la plus utilisée. Elle est rapide, demande peu d'espace mémoire et s'adapte à la classification en ligne de flux de données (Beringer et Hüllermeyer, 2003). L'algorithme partitionne les données en K classes (ou clusters) disjointes. Par itération successives, il place un prototype (ou centroïde) au centre d'un espace de données représentant la classe considérée. Le nombre de classes doit être fixé au préalable par l'utilisateur. L'algorithme se termine lorsque les classes sont stables (i.e lorsque qu'aucune donnée ne change de classe lors d'une itération). Cette technique « ancienne » est un classique, elle est utilisé dans différents domaines (notament en traitement d'image), et elle a connu de nombreuses améliorations et implémentations (*kernel k-means*, *global k-means*, *Gaussian-means*, etc)

La *Décomposition en Valeurs Singulières* (en anglais *Singular Value Decomposition* : SVD) (Golub et Loan, 1996) permet un partitionnement en classes qui se chevauchent, grâce à la décomposition d'une matrice termes/documents en un produit de 3 matrices utilisées notamment dans le cadre de travaux sur l'association mots-clés/documents (*Latent Semantic Indexing* : LSI). Un document est représenté par un vecteur de termes. Cette méthode est donnée comme moins efficace sur des gros corpus non homogènes (P. et al., 2001), de plus elle est coûteuse en temps et en espace mémoire. Elle est souvent utilisée pour réduire la taille de matrices conséquentes grâce à l'extraction de caractéristiques importantes.

La classification hiérarchique représente une collection de classes sous forme d'arborescence. Il existe deux approches de classification hiérarchique, l'approche ascendante et l'approche descendante. L'approche ascendante procède par fusions successives de classes déjà existantes, alors que l'approche descendante procède par segmentation. Le regroupement s'opère en fonction d'un calcul de distance entre les classes. La classification hiérarchique est une méthode obtenant de bons résultats (Steinbach et al., 2000), mais qui reste limitée à cause de sa complexité quadratique ($O(n^2)$), comparée à une méthode ayant une complexité linéaire ($O(n)$) comme les *K-moyennes*.

Le *Latent Dirichlet Allocation* (LDA) est un modèle probabiliste proposé par Blei et al. (2003), dans lequel la notion de thématique est centrale. Chaque document est vu comme un sac de mots, l'algorithme considère un document comme une distribution de probabilité sur des topics (sujets) au lieu de considérer la distribution sur les mots. De cette manière un mot peut se retrouver associé à plusieurs topics. Par exemple le mot « site » n'aura pas la même signification dans un article parlant de sites internet ou dans un article parlant de sites archéologiques.

Il existe une version en-ligne du LDA (*OLDA : Online Latent Dirichlet Allocation*) (AlSumait L., 2008) qui permet de rajouter les documents de façon incrémentale aux classes, afin d'avoir des classes à jour à l'arrivée de chaque nouveau document.

LDA ne permet pas seulement de regrouper les documents en classes mais permet également d'extraire des topics de ces documents. Il s'agit d'une méthode efficace pour des documents multi-topics (Xiao, 2010).

Il existe une multitude de solutions de classification et de détection d'événements.Toutes ne s'adaptent pas à la classification en ligne non supervisée. Dans le cadre du NED, la classification doit se faire « en ligne » c'est-à-dire traiter au fur et à mesure les documents qui arrivent dans un flux de données, à l'inverse de la classification « hors-ligne » qui traite un corpus déjà complet. La classification « en ligne » est plus complexe, elle nécessite la mise en place d'un processus de récupération des nouveaux articles, ainsi que d'un traitement enrichissant le corpus existant. L'étape de classification doit être adaptée, notre choix s'est porté sur la méthode OLDA décrite par AlSumait L. (2008), car elle permet non seulement de classifier les documents mais aussi de détecter l'émergence de nouveaux événements dans un flux.

3 Contribution

Nous proposons une chaîne de traitement (fig. 1) permettant de regrouper en classes, dans un flux d'articles de presse, les articles traitant d'un même sujet dans un espace géographique et pour une fenêtre temporelle donnée. Nous proposons ensuite d'identifier les « faits marquants » afin de faciliter le repérage et l'analyse d'un flux d'informations.

Nous avons fait le choix de commencer par le regroupement temporel car c'est ce qui nous permet de définir l'actualité. Nous opérons dans un système en-ligne, la fenêtre temporelle nous permet d'éliminer très simplemement les articles et ainsi réduire le travail des étapes de classification suivantes. Les classifications spatiale et thématique peuvent ensuite être interverties. Nous avons fait le choix de faire la classfiication thématique en fin de traitement, car il

s'agit du traitement le plus lourd. De cette manière, l'algorithme de classification thématique opère sur des corpus réduits.

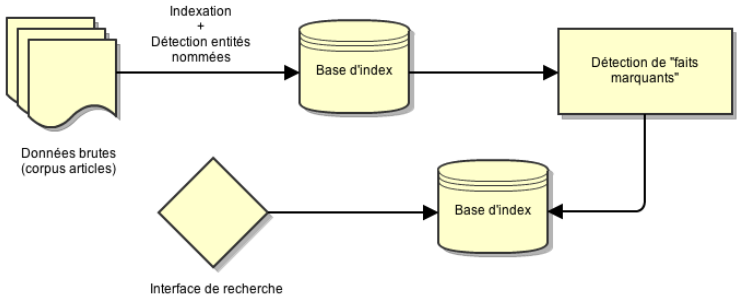


FIG. 1 – *Processus du traitement de l'information*

3.1 Processus de détection automatique de «faits marquants»

Nous regroupons les articles en classes d'événements (section 3.1.1) contenant les articles traitant d'un même sujet, pour une période et des échelles spatiales données. Afin de terminer le processus de détection de « faits marquants » (section 3.1.2), nous repérerons les classes contenant le plus d'articles.

3.1.1 Regroupement et classification

Regroupement temporel L'information d'actualité d'agence de presse étant non-structurée et évolutive, la localisation temporelle d'un événement peut se révéler complexe. Nous ne cherchons pas, dans nos travaux, à connaître la localisation temporelle précise d'un événement mais uniquement à savoir si l'article le relatant se trouve dans une fenêtre temporelle délimitant «l'actualité». Lorsqu'un nombre important d'articles traitent un même événement (passé, présent ou futur) alors on parle d'un «fait marquant» de l'actualité. Nous considérerons dans cette proposition qu'un événement fait parti de l'actualité, dès lors que la date de parution de l'article est en relation avec la fenêtre temporelle. Si un ensemble d'articles du jour traitent par exemple de la Seconde Guerre mondiale, nous ne nous intéressont pas à sa localisation temporelle mais uniquement au fait qu'elle soit traitée par un nombre important d'articles d'actualité. Ce choix permet d'avoir un processus rapide car il permet d'éliminer la phase d'interprétation de toutes les marques temporelles contenues dans les articles.

La composante temporelle est limitée par une fenêtre temporelle glissante délimitant un intervalle de temps. Notre contribution s'appuie sur un corpus enrichi quotidiennement, l'utilisation d'une fenêtre glissante permet d'adapter notre méthode à l'exploitation en ligne d'un flux continu et important d'informations.

Au terme de ce premier regroupement nous obtenons un ensemble de clusters catégorisant les articles dans chacune des fenêtres temporelles constituant l'intervalle.

Regroupement spatial A partir de ce premier niveau de regroupement (temporel) nous allons maintenant rafiner et segmenter les « clusters » précédemment obtenus par un regroupement spatial. Nous utilisons les entités nommées de lieux, extraites du contenu de chaque article, pour les catégoriser dans des échelles spatiales choisies au préalable.

Une fois les classements temporel et spatial réalisés, nous obtenons un index d'articles construit selon ces deux aspects. Il nous reste à traiter la composante thématique afin d'obtenir nos classes d'événements.

Regroupement thématique Pour le regroupement thématique nous avons choisi de travailler uniquement sur les titres des articles. Il s'agit de textes de taille réduite, décrivant le contenu de l'article avec un vocabulaire restreint. Dans la grande majorité des cas, ils sont représentatifs de la nature de l'événement traité par l'article.

La classification d'un nouvel article ne s'effectue pas sur le corpus entier existant mais uniquement en comparaison avec les classes d'événements présentes dans la fenêtre temporelle qui délimite la validité de l'actualité. De cette façon, les classes regroupant des articles dont les dates de parution dépassent la fenêtre temporelle, ne seront plus mises à jour lors de l'arrivée de nouveaux articles. Dans le cas où ceux-ci traitent d'un événement similaire à une classe existante hors de la fenêtre temporelle alors une nouvelle classe sera formée.

L'objectif est de détecter de manière automatique l'arrivée d'un nouvel événement (Allan et al., 1998) dans le flux, puis une fois que les étapes de regroupement temporel et spatial sont réalisées, il s'agit de déterminer si c'est un événement que nous avons déjà traité ou s'il s'agit d'un nouvel événement. Dans le cas d'un nouvel événement, la méthode OLDA crée une nouvelle classe avec l'article le contenant. Alors que dans le cas d'un événement déjà traité, OLDA ajoute l'article le contenant dans la classe correspondante. De cette façon les classes d'articles traitant d'un même événement sont maintenues à jour à l'arrivée de chaque nouvel article.

Le modèle de génération de la méthode est défini par AlSumait L. (2008) de la manière suivante :

1. Pour chaque classe k = 1, ... , K
   2. $\beta_k^t = B_k^{t-1}\omega^\delta$ avec B une matrice et ω^δ un vecteur de distance
   3. $\Phi_k^t \sim Dir(.|\beta_k^t)$
4. Pour chaque document d,
   5. $\theta_d^t \sim Dir(.|\alpha^t)$
   6. Pour chaque termew_i dans le document d,
      7. Choisir un topic z_i selon une distribution multinomiale θ_d^t ; $(p(z_i|\alpha^t))$
      8. Choisir un mot w_i selon une distribution multinomiale Φ_{z_i} ; $p(w_i|z_i, \beta_{z_i}^t)$

Nous obtenons ainsi une indexation des articles selon des critères de similarités temporelle, spatiale et thématique.

3.1.2 Repérage des « faits marquants »

Afin de terminer le processus de détection de « faits marquants » nous devons sélectionner les classes contenant le plus d'articles. Pour ce faire nous utiliserons un critère de redondance que l'on peut aussi définir, à la manière de Binsztok et al. (2004), comme étant l'activité d'un événement. Plus une classe d'événements possède de nouveaux articles, plus elle sera représentative de l'actualité. Binsztok et al. (2004) proposent de mesurer l'activité d'un événement, cependant nous nous basons sur un seuil empirique, déduit de nos expérimentations (section 4), afin de déterminer à partir de combien de nouveaux articles une classe fait partie des « faits marquants ».

Grâce à notre méthode de classification nous pouvons détecter les « faits marquants » dans un flux d'actualité selon un périmètre temporel ou spatial. Par exemple, nous pouvons proposer à l'utilisateur les « faits marquants » ne concernant que l'Europe, la France ou encore les « faits marquants » du mois dernier, etc.

4 Expérimentation

4.1 Processus de détection automatique de « faits marquants »

Nous menons notre expérimentation sur un flux d'articles d'actualité alimenté en continu. Nous avons développé dans le cadre de nos expérimentations la chaîne de traitements permettant de récupérer les articles de presse quotidiennement ainsi qu'une interface de recherche permettant de visualiser le résultat de notre classification. Notre corpus est enrichi en moyenne de 20 nouveaux articles par jour provenant de différentes sources (ex : France 24, Euronews, etc.).

Comme nous l'avons vu précédemment, nous travaillons à partir de la date de parution pour la localisation temporelle. Pour la localisation spatiale nous utilisons les entités spatiales contenues dans l'article complet. Par contre nous avons déterminé, d'après nos expérimentations, que l'algorithme de classification thématique est plus performant sur des textes courts et peu bruités. Nous avons donc décidé de n'utiliser pour la classification thématique que les titres des articles. En effet il s'agit d'un résumé de l'article où chaque mot a son importance. Nous pourrions avoir de meilleures performances en trouvant un compromis entre l'utilisation de titres seuls ou des articles dans leur ensemble. Ces trois informations représentent les trois composantes que nous avons définies comme étant représentatives d'une donnée d'actualité.

La taille de la fenêtre temporelle impacte directement les résultats obtenus. Le choix de cette taille est arbitraire, et nous pourrons par la suite laisser ce choix à l'utilisateur en fonction des besoins. Nous pouvons par exemple, si le corpus est assez important, proposer les « faits marquants » d'une journée particulière, d'un mois ou d'une année complète.
Nous définissons, pour nos expérimentations, une fenêtre temporelle que nous limitons à 10 jours et que nous faisons glisser chaque jour afin d'enrichir les classes d'événements au fur et à mesure.

Pour le regroupement spatial, nous appliquons un traitement par patrons lexico-syntaxiques afin d'extraire les entités nommées spatiales. Nous interrogeons un index géographique [3] afin

3. Index géographique ou gazetter : www.geonames.org

de connaître le pays ou le continent pour une entité spatiale considérée, cela afin de regrouper automatiquement les articles selon trois échelles différentes (mondiale, continentale et nationale). Des problèmes se posent lorsque des entités spatiales sont détectées à tord (bruit) et ce qui peut amener à classer l'article à la mauvaise échelle.

A partir de cette classification temporelle et spatiale nous appliquons l'algorithme permettant de classer les articles par événements. Le tableau 1 propose un échantillon obtenu grâce à l'utilisation de l'algorithme LDA [4] implémenté par Xuan-Hieu Phan et Cam-Tu Nguyen d'après Heinrich (2005). Nous avons utilisé pour cet exemple un corpus de 115 articles de presse francophone répartis du 10 au 20 mai 2011 et nous avons paramétré l'algorithme pour obtenir 25 topics composés de 7 mots.

Il s'agit de l'association de mots présents dans les différents titres. On note bien l'association de mots pour former un sujet (topic), on reconnaît par exemple l'*affaire DSK* (Topic 1), la crise économique en Grèce (Topic 2) ou encore les révolutions récentes dans le monde arabe (Topic 3). On note aussi la présence d'un même mot dans plusieurs topics (ex : fmi), l'algorithme LDA est capable d'identifier un mot dans des contextes différents.

Topic 1	Topic 2	Topic 3
dsk	crise	révolution
fmi	grèce	arabes
agression	dette	manifestation
new-york	politique	dictature
affaire	**fmi**	libye
directeur	europe	egypte
justice	allemagne	khadafi

TAB. 1 – *Topics obtenus avec l'algorithme LDA*

L'algorithme fournit aussi l'association documents / topics, avec pour chaque document un score de similarité correspondant aux différents topics. Le regroupement d'articles se fait grâce à cette association. Nous devons définir expérimentalement un seuil (dépendant de la taille du corpus) en fonction duquel les articles seront associés à un ou plusieurs topics. Une fois l'association documents / topics réalisée nous obtenons notre classification thématique. Cette étape est faite sur l'ensemble des classe obtenues préalablement avec les classifications temporelle et spatiale.

Une fois notre processus de classification terminé, nous obtenons une nouvelle base d'index ajoutant des liens de similarité entre les articles, cela nous permet non seulement de détecter les « faits marquants » de l'actualité, mais aussi de proposer à l'utilisateur des articles similaires lors d'une recherche ou de la consultation d'un article en particulier.

4. JGIbbLDA Implémentation Java du LDA : http://jgibblda.sourceforge.net

5 Conclusion et Perspectives

Au travers de cet article, nous proposons une chaîne de traitement simple et rapide capable de détecter et de faire émerger les « faits marquants » dans un flux continu d'actualité. Nous avons défini les composantes thématique, temporelle et spatiale comme représentatives d'une donnée d'actualité ou d'un événement, elles sont donc au centre de notre problématique. Nous apportons une solution au problème de la détection automatisée en appliquant des algorithmes et des techniques d'extraction de topics et de classification de texte lors de l'arrivée d'un nouvel article et selon des périmètres temporels et spatiaux.

Nous avons fait le choix, pour la gestion de la composante temporelle, d'utiliser une fenêtre temporelle glissante et de nous servir uniquement des dates de parution des articles comme repère temporel. Cette solution simple nous permet de traiter un nombre important de sources d'information en temps réel, et de pouvoir associer chaque article à une fenêtre temporelle sans avoir besoin de lourds calculs. En effet une autre solution serait d'extraire la composante temporelle d'un article directement depuis son contenu. Cette solution s'avère plus complexe à mettre en oeuvre, elle pose notamment des problèmes de représentation de la temporalité, elle soulève la question des repères utilisées et complexifie la gestion de la composante temporelle pour une classe d'événements. La durée d'un événement peut s'étaler dans le temps (mois, année, etc), on utilisera alors dans ce cas la notion de sous-événement définissant un événement comme étant un ensemble de sous-événements. Par exemple dans la fenêtre temporelle courante, un article peut évoquer un événement actuel et le rattacher à un événement antérieur. Le même problème se pose pour la localisation spatial, c'est pourquoi nous utilisons la notion d'espace englobants à différentes échelles.

Notre processus de détection automatique de « faits marquants » obtient des résultats encourageants. Cependant, nous avons identifié des erreurs lors de l'étape de classification thématique. Des articles associé à tord (bruit) ou des articles non-associés alors qu'il auraient du l'être (silence). Nous devons maintenant analyser plus en détails ces premiers résultats.

Une perspective d'évolution est la mise en place d'une méthode d'évaluation de notre chaîne de traitements. Notamment pour évaluer le taux d'erreurs dû à l'étape de classification. Cette méthode d'évaluation doit prendre en compte des expérimentations basées sur un corpus d'articles Reuters de référence afin de pouvoir comparer nos résultats avec d'autres travaux existants. Une deuxième perspective est la catégorisation des articles, en effet à l'heure actuelle notre chaîne de traitement ne permet pas d'associer les classes d'événements à des catégories générales. Dans le cadre d'un traitement automatique et d'une classification non supervisée, une des grandes difficultés est d'attribuer un nom cohérent aux différentes classes d'événements. Cette action s'apparente à un résumé automatique et représentatif des classes d'articles considérées. Dans le cas d'une information d'actualité il nous semble plus important d'associer les articles à des thématiques générales (i.e : catégorisation) que d'attribuer un nom à chaque classe d'événement.

Remerciements

Nous tenons à remercier l'entreprise DIS [5] pour le financement et l'aide apportée dans la réalisation de ces travaux.

Références

Allan, J., J. Carbonell, G. Doddington, J. Yamron, et Y. Yang (1997). Topic detection and tracking pilot study final report. Technical report.

Allan, J., R. Papka, et V. Lavrenko (1998). On-line new event detection and tracking. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 37–45.

AlSumait L., Barbara D., D. C. (2008). On-line lda : Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining,*, pp. 3–12.

Beringer, J. et E. Hüllermeyer (2003). Online clustering of data streams. Technical report, University of Marburg.

Binsztok, H., T. Artières, et P. Gallinari (2004). Un modèle probabiliste de détection en ligne de nouvel évènement. Technical report, lip6.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Bossard, A. et T. Poibeau (2008). Regroupement automatique de documents en classes événementielles. In *TALN 2008 Avignon*.

Forgy, E. (1965). Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics 21*, 768–780.

Golub, G. H. et C. F. V. Loan (1996). *Matrix computation*. JHU Press.

Heinrich, G. (2005). Parameter estimation for text analysis. Technical report, University of Leipzig, Germany.

Hill, L. L. (2006). *Georeferencing : The Geographic Associations of Information.* MIT Press, Cambridge, MA,.

P., H., S. H., et D. C (2001). On the use of singular value decomposition for text retrieval. *Computational Information Retrieval (SIAM)*, 45–156.

Papka, R. et J. Allan (1998). On-line new event detection using single pass clustering. Technical report, UM-CS-1998-021.

Rigouste, L., O. Cappé, et F. Yvon (2006). Quelques observations sur le modèle lda. In *Actes des Journées Internationnales d'Analyse statistique des données textuelles*, pp. 819–830.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Communications of the ACM 18, 11*, 613–620.

Saval A., M. Bouzid, e. S. B. (2009). A semantic extension for event modelisation. *21st IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2009)*.

5. Document Image Solutions - Cré@ticité C - Technopole Izarbel, 64210 Bidart - www.docimsol.eu

Serrano, L., M. Bouzid, T. Charnois, et B. Grilheres (2012). Vers un système de capitalisation des connaissances : extraction d'événements par combinaison de plusieurs approches. *SOS-DLWD'2012 at EGC'2012*.

Steinbach, M., G. Karypis, et V. Kumar (2000). A comparison of document clustering techniques. In *TextMining Workshop, KDD 2000*.

Turenne, N. (2001). Etat de l'art de la classification automatique pour l'acquisition de connaissances à partir de textes. Technical report, Inra.

Wayne, C. L. (1998). Topic detection and tracking (tdt). Technical report, National Security Agency.

Wegener, M. (2000). Spatial models and gis : new potential and new models. *International Journal of Geographical Information Science 15*, 487–488.

Xiao, Y. (2010). A survey of document clustering techniques and comparision of lda and movmf. Technical report.

Summary

Nowadays, with social networks and huge data streams, the information retrieval in the field of news can be difficult for a user.

We propose in this paper a processing chain operating on a stream for automatically detecting the "highlights" of the news. Our chain is based on an unsupervised classification of articles from the press, using clustering techniques based on spatial, temporal and thematic similarities.

We use the notion of "highlights" to define an event or set of events as having a certain redundancy in a set of data sources and for a given time window.

We rely on the methods of the "Topic Detection and Tracking", or the "New Event Detection" to achieve our contribution and our experiments we use an implementation of the "Latent Dirichlet Allocation". The combination of thematic and spatial clustering in a time interval, permits to bring out the current events by representing these three components.

Découverte de patrons de mobilité dans un réseau

Ahmed Kharrat [*,**] Karine Zeitouni[*]
Sami Faiz[***]

[*]Laboratoire PRISM
45, av. des Etats-Unis
78035, Versailles, France
Karine.Zeitouni @ prism.uvsq.fr
[**]SupCom
Cité Technologique des Communications
2083 El Ghazala Ariana, Tunisie
Ahmed.kharrat@supcom.rnu.tn
[***]Laboratoire LTSIRS
ENIT - Campus Universitaire – El Manar – Tunis – Tunisie
sami.faiz@insat.rnu.tn

Résumé. La fouille de données relatives aux objets mobiles est devenue un sujet important de recherche ces dernières années. Dans cet article, nous nous intéressons à la fouille de trajectoires d'objets mobiles dans un réseau telles que celles des véhicules dans le réseau routier. Nous proposons une méthode d'extraction de patrons de mobilité (PM) qui se caractérise par la prise en compte du réseau sous-jacent et de la variabilité temporelle de ces patrons. De plus, nous proposons d'enrichir les PM découverts par des mesures ou des résumés statistiques issues des trajectoires. Les patrons constituent ainsi des résumés spatio-temporels et descriptifs de groupes de trajectoires.

1 Introduction

Dans toute base de données, le gros volume de données et les valeurs trop détaillées rendent leur analyse fastidieuse. Les techniques OLAP et de fouille de données visent justement à synthétiser les données par des agrégations ou par l'extraction de motifs plus synthétiques et plus représentatifs de la base, et donc faciles à appréhender pour l'analyste. S'agissant des bases d'objets mobiles, les données sont particulièrement volumineuses et sont intrinsèquement trop détaillées (à l'origine des séquences de coordonnées de positions datées). En effet, dans la plupart des applications, des données trop détaillées pénalisent leur exploitation en raison des coûts élevés de stockage, d'accès et de transferts via les réseaux de communication. Ces données présentent souvent des similarités cachées (particulièrement pour les trajectoires contraintes par les réseaux routiers). Ces similarités se basent sur le contenu (des trajectoires) et sont complexes et difficiles à exhiber. Les patrons (ou motifs) de mobilité s'avèrent être le moyen le plus approprié pour représenter ces similarités. L'extraction de ces patrons est une

des techniques de fouille de données les plus réputées par leur performance d'analyse des données.
Cet article vise la recherche de patrons de mobilité décrivant un comportement fréquent sur un espace de référence. Le terme de patron est largement utilisé par diverses disciplines. Le terme de mobilité quant à lui, concerne le mouvement dans l'espace géographique d'individus, de biens ou d'informations. Un patron de mobilité est une forme récurrente apparaissant dans la succession des différentes valeurs prises par un groupe d'individus pour une dimension donnée (motivation, localisation, activité, etc.) (Chardonnel et al., 2004).

En termes de contributions, nous proposons dans cet article :

- Une nouvelle forme de patrons de mobilités PM adaptée aux données spécifiques d'objets mobiles tout en prenant en considération les caractéristiques de la mobilité dans un réseau, à savoir la densité du réseau routier et sa variabilité au cours du temps, ainsi que les mesures associées à cette mobilité.
- Une méthode de découpage automatique du temps.
- Un algorithme efficace de découverte de patrons de mobilité selon la définition proposée.

Le reste de cet article est organisé comme suit. Nous présentons tout d'abord dans la section 2 un bref état de l'art. Nous proposons ensuite quelques concepts préliminaires dans la section 3. Nous détaillons dans la section 4 l'algorithme de découverte de chemins denses NETSCAN. Dans la section 5, nous exposons notre approche de découverte de patrons de mobilité. La section 6 est dédiée aux résultats d'expérimentations. Enfin, une conclusion et des perspectives sont proposées à la section 7.

2 État de l'art

Outre le domaine des trajectoires d'objets mobiles, plusieurs domaines de recherche se sont focalisés sur la détection et la fouille de patrons pour faciliter l'analyse et l'exploration de grands volumes de données (Agrawal et Srikant, 1995). Dans le domaine des bases de données spatiotemporelles, Meratnia et de By (2002) ont présenté une méthode d'agrégation de trajectoires d'objets mobiles en s'appuyant sur une présentation raster de l'espace de référence. Le but de l'agrégation est d'identifier des trajectoires similaires et de les représenter par une seule trajectoire. Une approche similaire a été proposée dans (Giannotti et al., 2007) où la notion de patron de trajectoire a été introduite. Elle consiste en une séquence de régions d'intérêts spatiales visitées fréquemment avec l'annotation du temps de déplacement typique. Dans ce même contexte, mais avec d'autres utilisations, un patron de mobilité est destiné dans (Mouza et Rigaux, 2005) à représenter de manière générique le comportement commun d'un groupe d'individus. Si l'on imagine une application permettant de suivre les déplacements et les activités d'un ensemble d'individus dans la région parisienne découpée en 7 zones (a, b, c, d, e, f, g), le patron de mobilité «$f - a - d - c$» permettrait de classer un certain nombre d'individus dont le déplacement correspond à ce patron.
Ces méthodes peuvent être appliquées sur des trajectoires d'objets mobiles contraintes par le réseau. Cependant, les patrons formés à base de ces trajectoires, contrairement à notre travail, ne font pas référence au réseau.

D'autres travaux considèrent qu'un patron peut être n'importe quelle configuration de quelques objets mobiles dans un certain secteur et/ou au cours d'une certaine période de temps

(Benkert et al., 2008). Le $flock$ (Gudmundsson et al., 2004) fait partie de ces travaux. C'est un groupe d'au moins m objets mobiles qui se déplacent ensemble dans une région circulaire de rayon r pendant un intervalle de temps spécifique d'au moins k instants de temps. Tandis que le $flock$ pourrait être sensible à la taille personnalisée par l'utilisateur du disque où la forme circulaire ne pourrait pas toujours être appropriée, Jeung et al. (2010) proposent le concept du $convoy$. Le $convoy$ est un groupe d'objets mobiles, contenant au moins m objets qui sont connectés à base de densité, respectant une distance e parcourue pendant k instants consécutifs de temps. En le comparant avec le $flock$, le $convoy$ a une définition plus flexible pour trouver des clusters d'objets mobiles, mais il est encore limité par une contrainte rigide sur les instants consécutifs. Dans Li et al. (2010), les auteurs proposent un nouveau concept de patron de mobilité, appelé $SWARM$, qui relaxe les contraintes du $flock$ et du $convoy$ et s'avère plus flexible pour l'opération du clustering. $SWARM$ est une paire (O, T) avec O l'ensemble d'objets qui respecte un nombre minimum min_o d'objets et T l'ensemble de points de temps respectant un nombre minimum min_t d'instants de temps éventuellement non consécutifs. $Swarm$ peut avoir une forme géométrique arbitraire des clusters. Pour éviter de trouver des $SWARM$ redondants, Li et al. (2010) proposent le concept du $closedSWARM$. L'idée de base est que si (O, T) est un $SWARM$, il est inutile d'extraire un autre $SWARM$ sous-ensemble (O', T') telle que $O' \subseteq O$ et $T' \subseteq T$.
Bien que la notion de patrons de mobilité vise des objectifs similaires aux notre, la définition de ces patrons diffère des PM proposés. En effet, nous supposons qu'un patron de mobilité est un chemin fréquent sur le réseau routier avec un comportement typique d'objets mobiles. Ainsi, les PM constituent des résumés spatio-temporels et descriptifs de groupes de trajectoires. Notons qu'il existe dans l'état de l'art de nombreux autres travaux relatifs aux patrons de mobilité. En revanche, à notre connaissance, il n'existe pas de travaux sur des patrons spatio-temporels spécifiques aux réseaux routiers décrivant en même temps une information spatiale et une autre descriptive.

3 Préliminaires

Nous présentons dans cette section des notions préliminaires utiles pour le processus de découverte des PM. D'abord, nous définissons la manière de représentation des données. Nous expliquons ensuite une méthode existante pour le découpage de séries temporelles.

3.1 Représentation des données

Contrairement aux trajectoires d'objets mobiles libres, la représentation de trajectoires d'objets mobiles contraints par un réseau s'appuie en premier lieu sur la modélisation du réseau sous-jacent. La forme et les caractéristiques du réseau influencent largement la forme du mouvement des objets et leur vitesse.

Représentation du réseau : La représentation du réseau est donnée par l'ensemble de sections de routes. Nous définissons un réseau routier comme suit :

Définition .1 (Réseau routier)**.** Un réseau routier est un graphe orienté $G(N, S)$ où N est l'ensemble de nœuds correspondant aux intersections de routes et S d'arcs, chacun correspond à une section de route située entre les intersections.

Représentation de trajectoires : Une trajectoire est généralement représentée par une suite de triplets (x, y, t) où x et y sont les coordonnées absolues de la position à l'instant t. Lorsque la mobilité est contrainte par un réseau, il devient possible de simplifier ce modèle - en capturant l'information essentielle pour l'analyse - en une suite (s, t) où s est une section de routes et t une référence au temps. Cette représentation par référence au réseau est dite symbolique et en plus d'être compacte, elle est plus informative que la représentation absolue et plus adaptée à la fouille de données. Une autre spécificité de notre approche est d'étendre les trajectoires par des mesures associées. Ces mesures peuvent être générées par des capteurs embarqués (comme le bruit ou la pollution le long d'un trajet) ou dérivés par calcul de la mobilité de l'objet (par exemple la vitesse). Par conséquent, nous proposons la définition de trajectoire étendue aux mesures comme suit :

Définition .2 (Trajectoire étiquetée)**.** Une trajectoire étiquetée est définie par :
$Tre < \sharp S_1, M_1, t_1, \ldots, \sharp S_n, M_n, t_n >$ où $\sharp S_i$ référence la section traversée par Tre à un instant t_i et étiquetée par la mesure M_i.

Notons que la représentation par référence dépend de la granularité spatiale et temporelle. Le découpage en sections du réseau routier constitue une granularité suffisante pour l'analyse des trajectoires. Cependant, choisir une granularité temporelle n'est pas facile, car la mobilité varie au cours du temps et le patron de mobilité doivent refléter cette variation. Pour y remédier, nous proposons à la section 3.2 une méthode adaptative pour la discrétisation du temps selon la densité.

3.2 Discrétisation du temps

La période de temps total T couverte par les données historiques analysées d'objets mobiles varie. Elle peut être sur plusieurs années, plusieurs mois, plusieurs jours, voire plusieurs heures. Afin d'analyser le comportement des objets mobiles au cours du temps, nous proposons un partitionnement automatique de T en sous-périodes (intervalles) de temps. Ce partitionnement est appliqué sur la série temporelle de la densité globale du réseau. En partant de la représentation symbolique des trajectoires d'objets mobiles, cette série est produite en générant des statistiques sur la densité du réseau sur un temps discret.
L'objectif du partitionnement est de détecter un changement significatif de la densité au cours du temps. Une solution serait d'adopter un découpage à des périodes de temps égales. Cependant, cette solution risque de, soit ignorer un grand changement de la densité, soit détecter un changement minime qui n'a pas d'importance entre deux intervalles de temps adjacents. Il devient, alors, nécessaire d'utiliser une méthode de découpage dynamique et intelligent du temps.

D'un point de vue pratique, nous visons à obtenir une séquence d'intervalles de temps It_i ayant chacun une paire d'attributs : un instant de début et un instant de fin. Chaque It_i se caractérise par une différence de densité remarquable avec ses voisins. Les valeurs temporelles

sont représentées dans un espace discret (cf. figure 1). La largeur d'un intervalle de temps $L(It_i) = n_i$, avec $It_i \in T$

$$\sum_{i=1}^{k} It_i = L(T), \text{ avec } k\text{= nombre d'intervalles}$$

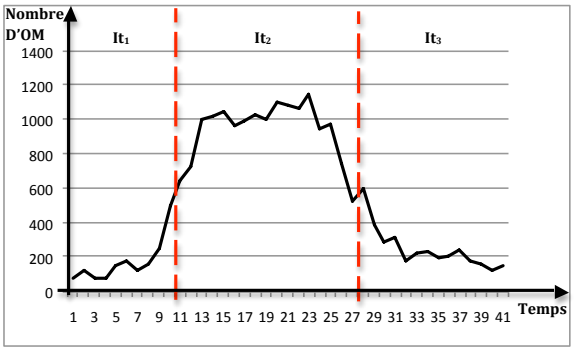


FIG. 1 – *Série temporelle de densité globale partitionnée*

Par ailleurs, le partitionnement consiste à définir les points de ruptures de la série temporelle de densité du réseau routier. Cette problématique a fait l'objet de recherche dans le domaine d'exploration des séries temporelles. L'approche de 'Segmentation' paraît la plus appropriée dans le cas de ruptures multiples dans les séquences. La méthode de 'Segmentation' de Hubert et al. (1989) est basée sur une procédure originale de découpage des séries temporelles. Cette méthode a été appliquée sur T pour un obtenir l'ensemble d'intervalles de temps $\mathfrak{I}$.

4 Découverte de chemins denses

L'algorithme NETSCAN (Kharrat et al., 2008) propose une méthode de classification automatique des trajectoires d'objets mobiles dont le mouvement est contraint par un réseau routier. Cet algorithme est basée sur la densité du réseau. Il se compose de deux phases : la première découvre des chemins denses, puis la seconde s'appuie sur ces chemins denses pour former des groupes de sous-trajectoires qui leurs sont similaires. Ces deux phases sont précédées par une étape de préparation de données. Elle consiste à calculer une matrice de transitions pour chaque intervalle de temps. Nous résumons dans la sous-section suivante cette étape suivie des deux phases de l'algorithme NETSCAN. Se référer à Kharrat et al. (2008) pour les détails de cet algorithme.

4.1 Matrice de transitions par période de temps

Connaissant l'ensemble des trajectoires, nous calculons une matrice de transitions associée au réseau routier pour chaque intervalle de temps. Celle-ci fournit des statistiques sur les passages aux carrefours et les mouvements tournants, en reportant le nombre de trajectoires qui

transitent d'une section à une section adjacente dans chaque intervalle de temps.

Définition .3 (Matrice de transition)**.** Une matrice de transition à l'intervalle de temps It est une matrice pondérée M_{It} telle que $M_{It}(i, j)$ correspond au nombre de trajectoires passant de la section S_i vers la section S_j dans l'intervalle de temps It avec $It \in \{It_1, \ldots, It_k\}$, k étant le nombre d'intervalles de temps. On note S_{ij} la transition de la section S_i à la section S_j.

4.2 Clustering de sections (NETSCAN - Phase I)

L'algorithme NETSCAN effectue le clustering des sections de routes denses et les agrège en formant des chemins denses. Il prend en entrée l'ensemble de sections qui constituent le réseau routier, les matrices de transitions spatiotemporelles associées à chaque intervalle de temps, un seuil de densité et un seuil de similarité entre les densités des transitions. NETSCAN regroupe les sections où transite le maximum d'objets mobiles en premier (les transitions les plus denses). Il regroupe, ensuite, les transitions connexes dans l'espace dont les densités sont similaires, constituant ainsi des chemins denses. Cette procédure est refaite pour chaque intervalle de temps. Tout comme les trajectoires, les chemins denses sont représentés sous forme de séquences de sections. Chaque section est identifiée par un symbole.

4.3 Clustering de trajectoires (NETSCAN - Phase II)

Cette section présente la deuxième phase de l'algorithme NETSCAN correspondant au clustering de trajectoires. Celle-là se base sur le résultat obtenu par la première phase de l'algorithme présentée dans la section précédente. En effet, les chemins denses sont considérés comme des centres naturels de clusters de sous-trajectoires d'objets mobiles. L'algorithme de clustering de trajectoires consiste à grouper les sous-trajectoires selon leurs similarités avec chacun des chemins denses générées lors de la phase I de NETSCAN. Le nombre de clusters retournés à la fin de l'algorithme est égal au nombre de chemins denses.

5 Patrons de mobilité

Plusieurs approches de fouilles de motifs de trajectoires considèrent les patrons de trajectoires comme une suite de points datés dans l'espace adjacent (Benkert et al., 2008) (Jeung et al., 2010). Cependant, cette approche n'est pas appropriée dans notre contexte où nous considérons des objets mobiles se déplaçant dans un réseau routier. Pour nous, un PM dans le réseau routier est un chemin du réseau global regroupant les routes les plus fréquentées par les objets mobiles dans un espace de temps. Ce même PM est également décrit à travers des mesures descriptives, comme la vitesse moyenne des objets mobiles. Nous détaillons dans ce qui suit notre algorithme appelé PATEXTRACT en vue de la découverte des PM. Tout d'abord, nous utilisons l'algorithme NETSCAN pour nous permettre d'obtenir les chemins denses appropriés à chaque intervalle de temps It (Phase I de l'algorithme). Ces chemins denses obtenus représentent des noyaux qui après calcul de similarité généreraient les clusters des sous-trajectoires d'objets mobiles denses (Phase II de l'algorithme). Ensuite, une opération de filtrage est effectuée sur les clusters à base de leur support. Enfin, les clusters obtenus après l'opération de

filtrage seront enrichis avec des informations de type mesures descriptives pour former nos PM. Ces mesures peuvent être liées aux sous-trajectoires qui le composent (exemple : la vitesse moyenne, la consommation de carburant, la moyenne des temps d'arrêts temporaires...). Formellement on peut définir les PM comme suit :

Définition .4 (Patron de mobilité (PM))**.** Un patron de mobilité PM_i est défini dans l'intervalle de temps It_i par le tuple $(\sharp ID, It, S, M)$ où :
- $\sharp ID$ est un identifiant,
- It l'intervalle de temps de PM,
- $S = (S_1, ..., S_n)$ représente une séquence finie de sections de route, $\forall S_i \in S, S_{i+1} \neq S_i$ et $n > 1$ représente la longueur du PM,
- $M = (M_1, ..., M_k)$ représente un ensemble de mesures associées avec PM.

Définition .5 (Support d'un PM)**.** Le support Sup d'un PM est défini comme étant le pourcentage du nombre de sous-trajectoires ayant une intersection non vide avec S (T_patron) par rapport au nombre total de trajectoires (T_total) dans le même intervalle de temps que PM. Le support est donc :

$$Sup = \frac{T\_patron}{T\_total}$$

Comme pour une trajectoire d'objet mobile, notre modèle de PM représente l'information spatiale sous forme symbolique. Pour rester dans le même contexte symbolique, nous avons adopté le même principe pour la représentation des mesures M en l'exprimant avec des symboles d'un alphabet fini. Nous nous basons pour représenter ces mesures sur la méthode SAX. La section suivante explique son adaptation à notre contexte.

5.1 Représentation symbolique des mesures

Les mesures désignent un ensemble d'informations décrivant chacun des PM. Ces mesures peuvent être très utiles pour l'analyse du trafic routier et notamment le comportement des conducteurs. Dans le même contexte de représentation symbolique des PM, nous proposons un modèle de représentation de ces mesures. Nous adaptons la méthode SAX (Symbolic Aggregate approXimation) (Lin et al., 2007) au contexte de trajectoire symbolique au lieu des séries temporelles auxquelles elles sont appliquées habituellement.

SAX permet de réduire une série temporelle de longueur n à une chaine de caractères de longueur w ($w < n$) suivant un alphabet de taille a. Une série temporelle passe par les trois étapes suivantes pour être représentée symboliquement :
- Normalisation de données.
- Transformation de données vers la représentation PAA.
- Symbolisation de la représentation PAA vers une représentation discrète.

SAX effectue des regroupements des points des séries temporelles représentées suivant des épisodes de même taille. Or les séries temporelles que nous étudions sont issues de mesures en rapport avec le déplacement d'objets mobiles sur le réseau routier. Ces séries présentent des séquences liées aux changement de section de route durant le déplacement. Un découpage pertinent en épisodes de même taille de telles séries pourrait ne pas représenter correctement les moments au cours desquels les séries temporelles ne sont pas constantes.

Pour cette raison, nous proposons d'adapter le découpage en épisodes à l'espace. Par conséquent, le découpage s'effectue au moment où l'objet mobile change de section de route. Ce passage d'une représentation non adaptative à une représentation adaptative ne répond plus à la condition qui considère que la longueur de la série temporelle d'origine de longueur n doit être divisible par la longueur de la série réduite w.
Notre proposition considère que chaque Mesure M dépend d'une section de route S. Donc la valeur w est égale au nombre de section de route dans un patron. La figure 2 illustre un exemple de l'utilisation de SAX pour la représentation symbolique de la vitesse. Après la normalisation des données, on calcule pour chaque trajectoire similaire à PM sa vitesse moyenne locale dans chaque section. Ensuite, on calcule une deuxième moyenne de toutes les vitesses moyennes locales des trajectoires. Le résultat obtenu correspond à une représentation PAA adaptative. Enfin et comme illustré par la figure 3, on transforme PAA en une représentation discrète qui donne la séquence d'alphabet suivante : "a b c b"

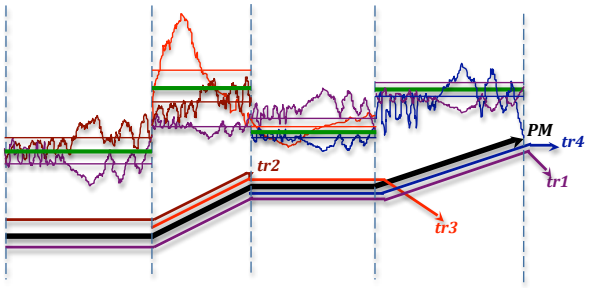


FIG. 2 – *Représentation PAA (Piecewise Aggregate Approximation)*

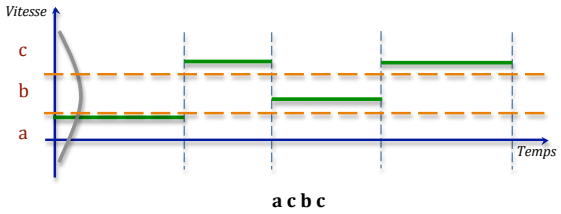


FIG. 3 – *Représentation SAX (Symbolic Aggregate approXimation)*

5.2 La construction des PM

L'algorithme PATEXTRACT dédié à la construction automatique et intelligente des PM comprend trois composantes qui sont :

1. L'application de l'algorithme NETSCAN avec ses deux sous composantes (NETSCAN-Phase I et NETSCAN-Phase II) dans chaque intervalle de temps It.
2. Le processus de filtrage des clusters des sous-trajectoires denses obtenus par la deuxième composante pour obtenir la partie spatiale (S) des PM.

3. L'opération d'enrichissement des PM avec l'ensemble de mesures (M).

L'algorithme PATEXTRACT est donné ci-dessous.

Algorithme PATEXTRACT
Entrée :
– Ensemble de trajectoires d'objets mobiles
$\mathfrak{T} = \{Tre_1, Tre_2, \ldots, Tre_{nb\_trajectoires}\}$.
– Intervalles de temps $\mathfrak{I} = \{It_1, It_2, \ldots, It_k\}$.
– Seuil de densité $MinSup$
Sortie :
– Ensemble de patrons $\mathfrak{P} = \{PM_1, PM_2, ..., PM_n\}$.
Début :
Pour chaque intervalle de temps $It_i \in \mathfrak{I}$ **faire**
  Appliquer l'algorithme NETSCAN
Fin Pour
Pour chaque cluster de sous-trajectoires $Clus$ **faire**
  Calculer Support Sup ;
Fin Pour
Filtrage :
Garder les clusters $Clus$ ayant $Sup > MinSupp$;
Initialisation d'un nouveau PM dans $\mathfrak{P}$ avec chaque $Clus$ sélectionné.
Enrichissement :
Pour chaque $PM_i \in \mathfrak{P}$ **faire**
  Calculer les mesures m_i ;
  $Representation\_Symbolique(m_i)$;
  Ajouter m_i à PM_i
Fin Pour
Fin
Retourner $\mathfrak{P}$;

Algorithme 1: Extraction de Patrons de mobilités

L'algorithme PATEXTRACT prend en entrée l'ensemble de trajectoires d'objets mobiles $\mathfrak{T}$, l'ensemble d'intervalles de temps $\mathfrak{I}$ et un seuil de densité $MinSup$. PATEXTRACT commence par appliquer l'algorithme NETSCAN dans chaque intervalle de temps $It \in \mathfrak{I}$. Ensuite, il applique l'opération de filtrage des clusters denses qui ont un support inférieur à $MinSup$. Le résultat de cette opération permet d'initialiser un nouveau PM, précisément, la partie spatial S de PM reçoit le noyau d'un $clus$. Ce dernier consiste en un chemin dense obtenu par NETSCAN-Phase I. Enfin, PATEXTRACT applique l'opération d'enrichissement en rajoutant les mesures descriptives pour chaque PM. La fonction "Representation_Symbolique" permet la représentation symbolique des mesures selon la méthode SAX.

6 Tests et validation

Dans cette section, nous évaluons l'efficacité de notre algorithme en décrivant les données et l'environnement de travail. Nous discutons et interprétons également dans cette section les résultats de notre expérimentation.

6.1 Configuration et données de base

L'algorithme PATEXTRACT a été implémenté en Java et tous les tests ont été effectués sur un PC fonctionnant sous Windows XP professionnel. La configuration matérielle est comme suit : Processeurs AMD Athlon TM 64 X 2 dual Core 2 GHZ, 1.5 giga octet de mémoire RAM et 80 gigaoctets de disque dur. Nous utilisons Oracle 10g comme gestionnaire de données.

La trajectoire d'un objet mobile peut être obtenue à partir de plusieurs sources, telles que les données de flottes de voitures, mais celles-ci sont généralement des données propriétaires et leur accès est limité. Par conséquent, nous utilisons des données simulées qui permettent de se rapprocher du monde réel dans les tests et les validation et aussi d'évaluer le comportement des algorithmes ou des données dans des situations exactement spécifique ou extrême. Dans notre domaine, le générateur développé par Brinkhoff est le plus utilisé pour les tests et l'évaluation (Brinkhoff, 2002).

6.2 Implémentation

Nous avons appliqué le générateur sur le réseau routier d'Oldenburg (7035 segments de routes et 6105 nœuds) pour produire différents jeux de données de trajectoires d'objets mobiles sur ce réseau routier. Partant de ces données, nous appliquons la méthode de Hubert pour le découpage temporel. Ayant des intervalles de temps bien délimités, nous procédons au calcul des matrices de transitions. Plus précisément, pour chaque transition (i, j) dans cette matrice, nous comptons les occurrences des objets mobiles la traversant. L'algorithme PATEXTRACT a été implémenté et testé selon différentes configurations. Nous avons fait varier le nombre d'objets mobiles entre 1000 et 10000.

6.3 Résultats d'expérimentation

La figure 4 montre les 304 chemins denses découverts pendant trois intervalles de temps. Chaque couleur désigne un intervalle de temps. La couleur jaune désigne It_1 avec ses 250 chemins denses. It_1 montre une densité largement supérieur à It_2 (en bleu) avec 29 chemins denses et It_3 (en noir) avec uniquement 25 chemins denses.
La figure 5 montre les patrons de mobilité des mêmes intervalles de temps avec le même jeu de données. On peut remarquer visuellement que l'on a moins de patrons que de chemins denses (105 patrons de mobilité). Ça s'explique par l'étape de filtrage que subissent les clusters denses par un seuil MinSup = 20. Dans It_1 nous avons 51 patrons, dans It_2 nous avons 29 patrons et dans It_3 25 patrons.

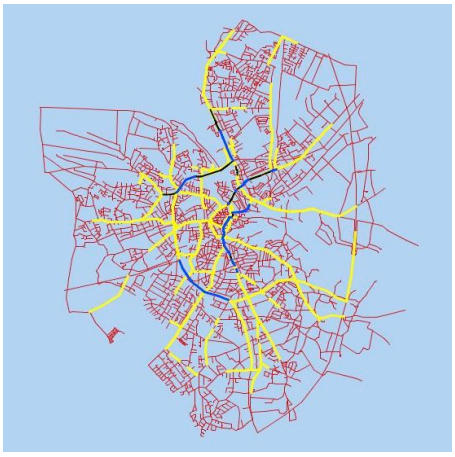


FIG. 4 – *Chemins denses*

FIG. 5 – *Patrons de mobilité*

Le nombre de patrons varie selon la variation du seuil $MinSup$. La figure 6 montre la courbe de cette variation.

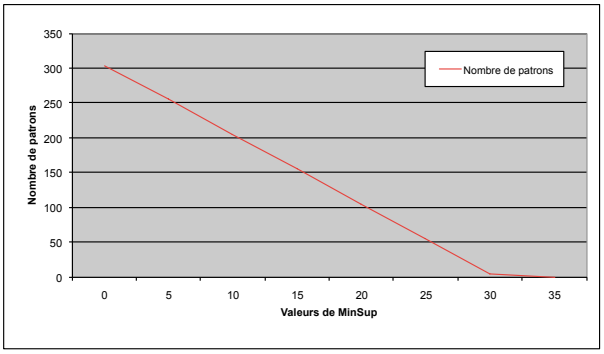


FIG. 6 – *Variation du nombre de patrons*

7 Conclusion et perspectives

Cet article se situe dans le cadre de la fouille de données spatio-temporelles. Plus précisément, il propose une nouvelle approche de découverte des PM.

L'extraction de ces PM à partir des données d'objets mobiles facilite l'analyse et l'exploitation de ces données. Notre méthode génère des PM qui résument les sous-trajectoires fréquentes, les périodes et les informations descriptives associées. En guise de perspective, nous proposons d'utiliser ces résumés pour compresser la représentation des trajectoires de base en substituant des sous-trajectoires entières par certains des PM découverts.

Références

Agrawal, R. et R. Srikant (1995). Mining Sequential Patterns. pp. 3–14.

Benkert, M., J. Gudmundsson, F. Hübner, et T. Wolle (2008). Reporting flock patterns. *Computational Geometry 41*(3), 111–125.
eng

Brinkhoff, T. (2002). A framework for generating network-based moving objects. *Geoinformatica 6*(2), 153–180.

Chardonnel, S., C. du Mouza, M. Fauvet, D. Josselin, et P. Rigaux (2004). Patrons de mobilité : proposition de définition, de méthode de représentation et d'interrogation. *7ème Journées CASSINI*, 19–25.

Giannotti, F., M. Nanni, F. Pinelli, et D. Pedreschi (2007). Trajectory pattern mining. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD San Jose*, 330.

Gudmundsson, J., M. Van Kreveld, et B. Speckmann (2004). Efficient detection of motion patterns in spatio-temporal data sets. *Proceedings of the 12th annual ACM international workshop on Geographic information systems GIS 04*, 250.

Hubert, P., J. P. Carbonnel, et A. Chaouche (1989). Segmentation des séries hydrométéorologiques. Application à des séries de précipitations et de débits de l'afrique de l'ouest. *Journal of Hydrology 110*(3-4), 349–367.

Jeung, H., M. L. Yiu, X. Zhou, C. S. Jensen, et H. T. Shen (2010). Discovery of Convoys in Trajectory Databases. *Proceedings of the VLDB Endowment 1*(1), 1068–1080.

Kharrat, A., I. S. Popa, K. Zeitouni, et S. Faiz (2008). Clustering Algorithm for Network Constraint Trajectories. In *SDH*, Lecture Notes in Geoinformation and Cartography, Berlin, Heidelberg, pp. 631–647. Springer Berlin Heidelberg.

Li, Z., B. Ding, J. Han, et R. Kays (2010). Swarm : mining relaxed temporal moving object clusters. *Proceedings of the VLDB Endowment 3*(1-2), 723–734.

Lin, J., E. Keogh, L. Wei, et S. Lonardi (2007). Experiencing SAX : a novel symbolic representation of time series. *Data Mining and Knowledge Discovery 15*(2), 107–144.

Meratnia, N. et R. A. de By (2002). Aggregation and comparison of trajectories. In *Proceedings of the tenth ACM international symposium on Advances in geographic information systems - GIS '02*, New York, USA, pp. 49. ACM Press.

Mouza, C. et P. Rigaux (2005). Mobility Patterns. *GeoInformatica 9*(4), 297–319.

Summary

In this paper, we are interested in mining trajectories of moving objects such as vehicles in the road network. We propose a method for a Mobility Pattern (MP) discovery which is characterized by taking into account the underlying network and temporal variability of these patterns. In addition, we propose to enrich the discovered MP by measurements or statistical summaries related to the trajectories.

Fouille d'images animées : cinéradiographies d'un locuteur

Julie BUSSET* et Martine CADOT*

*Université de Nancy1/LORIA,
Campus scientifique, BP 239, F-54506 Vandoeuvre-lès-Nancy Cedex
julie.busset@loria.fr
martine.cadot@loria.fr

Résumé. L'analyse de séquences cinéradiographiques d'une personne prononçant plusieurs phrases présente des difficultés. La première, technique, est que ces données proviennent d'images annotées en plusieurs lieux, temps, et de manière semi-automatique ou manuelle. La deuxième, représentationnelle, est que les mouvements des articulateurs pendant la parole (langue, mâchoire, etc.) se situent dans un espace-temps complexe du fait des interdépendances mécaniques multiples et dynamiques. Nous décrivons dans cet article l'extraction d'un modèle articulatoire de la parole à partir des données, sans ajout de connaissances *a priori*, à l'aide d'une méthode de fouille de données. Ce modèle met au jour l'organisation des structures articulatoires du locuteur tant dans la dimension spatiale que dans la dimension temporelle. La confrontation par l'expert de ce modèle aux mouvements attendus des articulateurs est un succès et nous invite à poursuivre dans cette voie.

1 Introduction

La production de parole nécessite l'action de plus de 200 muscles. Pour la synthétiser, on peut s'appuyer sur un modèle mettant en jeu les articulateurs.

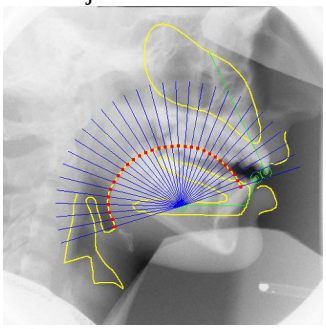


FIG. 1 – *Annotation d'images cinéradiographiques selon Busset (2009), Laprie et Busset (2011) : contours en jaune (articulateurs et régions de référence), grille adaptative en bleu, points du contour de langue en rouge, balises en vert (centre de la grille et incisives).*

Le modèle articulatoire le plus connu est celui de Maeda (1990), comportant 7 paramètres (1 pour la mâchoire, 3 pour la langue, 2 pour les lèvres, et 1 pour le larynx). Il a été cons-

truit à partir de l'observation d'images cinéradiographiques d'un locuteur prononçant de petites phrases. Sur chaque image les contours des articulateurs ont été repérés par les coordonnées d'un certain nombre de points, puis soumis à des analyses factorielles. Par exemple la langue a fourni plus de 100 points, et les 3 paramètres retenus dans le modèle de Maeda correspondent au nombre de dimensions de l'ACP (Analyse en Composantes Principales) expliquant 98% de la variance de cet articulateur. Des variantes de ce modèle ont été proposées depuis, essentiellement obtenues par analyses factorielles sur des points repérés différemment. Un exemple de variante proposée par Busset (2009), Laprie et Busset (2011) est représenté en figure 1 : des balises (en vert) ont été ajoutées, ainsi que des « régions », aux différents articulateurs (en jaune) pour contrôler les mouvements de la tête. Les points de la langue (en rouge) ont été obtenus par intersection de son contour avec une grille (en bleu) différente de la grille semi-polaire initiale de Maeda, puis des ACP ont été appliquées afin d'extraire les paramètres articulatoires.

Nous exposons dans cet article la construction d'un nouveau modèle articulatoire par une méthode de fouille de données appliquée aux données de quatre séquences comportant presque 1000 images annotées (voir figure 3 un exemple d'annotation). Nous décrivons en partie 2 la méthode de fouille de données utilisée. En partie 3 nous détaillons les données et leur transformation. La partie 4 expose les différentes étapes de la construction du modèle à partir des données transformées, d'abord son extraction sur une séquence, puis sa validation par l'expert et enfin sa cohérence avec une autre séquence. La conclusion est en partie 5.

2 La fouille de données

Les méthodes de fouille de données se sont développées depuis quelques décennies suite à l'afflux de grandes quantités de données librement accessibles sur Internet et à l'augmentation de puissance des ordinateurs, ayant une capacité accrue de stockage et une plus grande rapidité de traitement. Elles sont venues compléter les méthodes statistiques déjà utilisées pour analyser les données.

2.1 Fouille de données versus Statistique

Les méthodes de fouille de données sont utilisées pour extraire automatiquement de la connaissance à partir des données à l'aide d'algorithmes, en s'affranchissant de toutes les contraintes imposées par les statistiques *inférentielles*, qui sont : construction d'un échantillon, élaboration d'une hypothèse, recherche du test d'hypothèse approprié, contrôle de toutes les conditions imposées par le test, comme la normalité des résidus, etc. Mais en s'affranchissant ainsi, elles ne peuvent plus prétendre extraire un *modèle* c'est-à-dire de la connaissance généralisable à d'autres données. Pour remédier à cela, les techniques d'apprentissage automatique se sont développées, telles que le *rééchantillonage*, la *validation croisée*, ou tout simplement la division des données en deux parties ou plus, une pour la découverte du modèle (appelée ensemble d'*entraînement*) et les autres pour sa vérification (un ensemble de *test* et éventuellement un ensemble d'*ajustement*).

Comme pour les données traitées en statistique, celles issues de la fouille de données doivent être aussi nettoyées : erreurs de saisie, mauvais encodage, données manquantes. Mais ce nettoyage va bien au-delà de celui pratiqué par les statisticiens dans les décennies précédentes, quand on appelait un *grand échantillon* un ensemble d'une trentaine d'enregistrements,

et qu'on envisageait difficilement plus d'une vingtaine de variables. Pour donner un ordre d'idées, il est connu que la phase de pré-traitement des données représente parfois plus 90% du travail de fouille de données. Notamment, les analyses factorielles interviennent parfois à un moment de cette étape pour permettre de structurer les données, qu'elles soient mal structurées ou que leur structure soit mal adaptée au traitement envisagé.

L'extraction des motifs et règles d'association est une méthode de fouille de données qui s'est largement diffusée suite à son succès dans le traitement des données issues des tickets de caisse de la grande distribution[1] : un *motif* (itemset en anglais) est un groupe d'articles présents simultanément sur plusieurs tickets de caisse, et une *règle d'association* associe deux motifs dans une relation de type « cause à effet » ou apparenté. On extrait ainsi des *pépites de connaissance* au sein de données de grande dimension (Han et al, 1989). Dans la mesure où cette extraction met au jour des informations inconnues auparavant, elle fait partie des méthodes d'apprentissage non supervisées.

2.2 L'extraction de motifs

L'algorithme Apriori a été conçu par Agrawal et Srikant (1994) pour extraire rapidement les motifs les plus fréquents de données volumineuses. Pasquier et al. (1999) en ont proposé une amélioration utilisant les motifs fermés et les treillis. Depuis, de nombreux développements ont permis d'améliorer cette extraction, notamment en ne choisissant que les motifs les plus pertinents selon des points de vue variés. Pour les données qui nous intéressent ici, nous avons privilégié la représentation présente dans les travaux de Guigues et Duquenne (1986) et Luxenburger (1991), qui s'appuyait déjà sur la notion de motif fermé dans une structure de treillis. Ils ont synthétisé ainsi des tableaux booléens de type (sujets X variables), comme celui croisant les clients et les articles qu'ils ont achetés, mais de taille réduite. Les motifs qu'ils définissent, appelés *concepts*, sont formés de deux ensembles, un ensemble A de variables et un ensemble B de sujets tels que tous les sujets de B vérifient toutes les variables de A. Ils sont fermés quand pour toute variable v en dehors de A, il existe au moins un sujet de B qui ne la vérifie pas, et pour tout sujet s en dehors de B, il existe au moins une variable de A qu'il ne vérifie pas. Nous renvoyons à Cadot (2006) pour plus de détails.

3 La problématique et les données

La première partie de cette section décrit le cadre général de notre étude en reprenant de façon condensée le chapitre 3 du mémoire de stage de Master 2 réalisé par Julie Busset en 2006-2007 au sein de l'équipe PAROLE du LORIA et la deuxième partie expose le pré-traitement et la mise en forme des données.

Nous nous sommes limitées dans la première partie aux éléments de connaissance du domaine de la parole nécessaires pour pouvoir mesurer l'apport de la fouille de données à la modélisation des mouvements du conduit vocal lors de la parole. Nous renvoyons le lecteur intéressé par plus de détails au document (Busset, 2007) d'où sont tirés ces éléments, ainsi qu'aux nombreux ouvrages théoriques traitant de ce domaine (Haton et al., 2006).

[1] C'est ce qu'on appelle de façon plus imagée le « panier de la ménagère » en français et « market basket » en anglais.

La deuxième partie représente de façon plus détaillée le travail de mise en forme des données préalable à l'extraction des motifs : acquisition, nettoyage, moyennage par segments et pré-traitement par une méthode factorielle. Ce sont les choix faits lors de cette étape de transformation des données qui vont permettre d'extraire sous forme de motifs un véritable modèle des données.

3.1 Les organes de production de la parole

La parole est une modulation du flux d'air qui provient des poumons. En passant par les cordes vocales, ce débit d'air engendre une onde voisée ou sonore qui est envoyée dans le conduit vocal (Figure 2), avant de sortir à travers les lèvres et le conduit nasal. Les différents sons proviennent essentiellement des déformations du conduit vocal, représenté de façon très simplifiée, pour un adulte, par un tube de 17 cm de longueur avec une section transversale qui varie entre 0 (constriction maximale) et 20 cm². Les principaux articulateurs de ce conduit vocal sont la mâchoire, la langue, les lèvres, le vélum (voile du palais) et le larynx.

En français, la plus petite unité de son est appelée « phonème », et ceux-ci se divisent essentiellement en « voyelles » et « consonnes ».

3.1.1 Production des voyelles et des consonnes du français

Les voyelles se caractérisent par l'absence ou la présence de nasalité (quand le voile du palais s'abaisse, ce qui met en parallèle les cavités buccale et nasale), le degré d'ouverture du conduit vocal, la position de la constriction principale du conduit vocal et la position des lèvres. Il y a 12 voyelles orales émises seulement par la bouche (par ex. dans les mots l**i**t, pr**é**, b**e**lle, p**a**tte, p**â**te, l**u**, p**eu**, l**eu**r, le, l**ou**p, l**o**t, l**o**tte) et 4 voyelles nasales (par ex. dans les mots br**in**, br**un**, l**en**t, l**on**g).

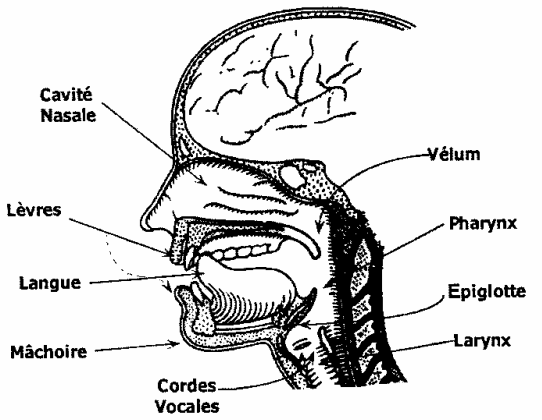


FIG. 2 – *Schéma de l'anatomie du conduit vocal (d'après Flanagran 1972)*

Les consonnes se caractérisent par le « voisement » (sonores ou sourdes), le mode d'articulation (occlusif, nasal, fricatif, glissant ou liquide), la position de la constriction principale (lèvres, dents, voile du palais). Par exemple la consonne **t** du mot « **t**ous » fait partie

de la classe des consonnes occlusives et sonores (Haton et al. 2006), comme le **c** de « **c**as » et le **p** de « **p**ère », ce qui se traduit par la succession des 3 étapes acoustiques suivantes :

1. un silence correspondant au blocage de l'air dans le conduit vocal ;
2. une explosion (« burst »), due au relâchement brusque de l'air ;
3. une transition vers la voyelle qui suit.

3.1.2 La co-articulation

Nous avons vu dans la sous-section précédente que la prononciation de la consonne **p** de « **p**ère » nécessitait trois étapes, la troisième étant une transition vers la voyelle suivante. Cette transition entre 2 phonèmes fait partie d'un phénomène plus général de « coarticulation », qui signifie que deux phonèmes qui se suivent subissent tous deux une distorsion[2]. L'étude de la coarticulation est un domaine de recherche très actif qui a donné lieu à plusieurs articles et thèses au sein de l'équipe Parole du LORIA.

3.2 Les données, leur acquisition et leur transformation

Le corpus a été enregistré dans les années 1990 afin d'étudier la coarticulation de la langue française. Les données sont formées de 4 séquences, que nous appellerons H1, H2, H3 et H4, composés de radiographies successives d'une même personne en train de prononcer des « phrases » courtes de 6 mots. Pour H3 et H4, la séquence est simple[3] (/aku/, /iku/, /uku/, /atu/, /itu/ et /utu/), mais pour H1 et H2 les 6 phrases sont plus complexes ; elles commencent et finissent par les mêmes phonèmes /se dø si/ et /yltɛʁ/ avec au milieu une consonne non labiale[4] de plus à chaque fois. La première phrase est /se dø siyltɛʁ/ et la sixième est /se dø si**kst sk**yltɛʁ/[5]. Les phrases sont prononcées plus vite dans H2 et H4 que dans H1 et H3.

3.2.1 Acquisition

Les images ont été annotées une à une d'abord manuellement puis de façon semi-automatique. Dans un premier temps, les contours des articulateurs ont été indiqués par autant de clics sur la radiographie que de points sur l'image de la figure 3, les coordonnées 2D de ces points étant stockées dans un fichier. Puis un logiciel a été mis au point afin d'aider l'annotation en se calant sur quelques images de référence. Dans la figure 3, on a représenté une image annotée avec une version de ce logiciel, la forme et la couleur des points indiquant leur appartenance à des contours. Pour réaliser, par exemple, le contour de l'os hyoïde, en jaune foncé, les points ont été importés du fichier de l'image de référence, puis le contour a été déplacé avec la souris (rotations et translations) pour l'adapter au mieux à l'image. Et on a procédé de même pour chacun des contours des articulateurs non déformables, comme la mâchoire (*jaw*, en bleu foncé), le haut de la tête (*HeadRegister*, en bleu clair, *upperMandi-*

[2] Termes repris d'une définition de Wikipedia (voir section Références)

[3] Le « slash » indique que la notation n'est pas celle du français, mais de la phonétique internationale : le /u/ est l'écriture phonétique du « ou » du français, alors que /y/ est l'écriture phonétique du « u » du français.

[4] Labiale : qui se prononce avec les lèvres

[5] Le symbole /ø/ désigne le « eu » de « peu », /ɛ/ désigne le « è » de « belle », et /ʁ/ désigne le « r » du français, qui est dans « père »

ble, en rose clair). Les lèvres inférieure et supérieure (levresMod1, levresMod2, en rouge), les 3 parties de l'épiglotte (ep1, ep2, epi3, en vert) étaient, quant à elles, trop déformables pour pouvoir utiliser cette méthode, elles ont donc été saisies point par point directement sur l'image. C'est la langue (*tongue* en rose), plus difficile à repérer car en grande partie cachée, qui a occasionné le plus de difficultés. Plusieurs techniques ont été testées pour aider à la détermination de son contour (Busset, 2011), dont certaines en posant des balises ainsi qu'une grille comme on peut le voir sur la figure 1.

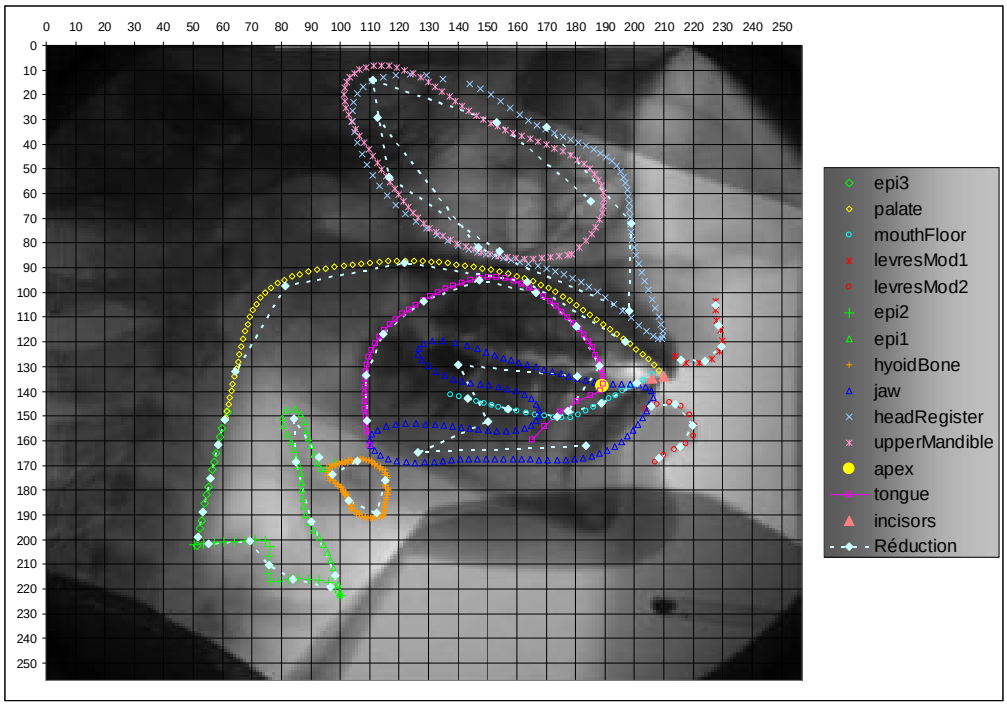


FIG. 3 – *En couleur les points des contours des articulateurs annotés sur la radiographie Hirsch1_156. En pointillés les courbes « réduction » normalisées, formées de 5 points pour chaque contour de Hirsch1_156, sauf la langue (tongue) qui en a 9.*

3.2.2 Nettoyage et normalisation

Nous disposions de 1000 fichiers de coordonnées de contours environ, autant que d'images, dont 671 sans contour manquant. Nous les avons lus un à un pour stocker les informations sur chaque contour (intitulé, coordonnées des points) dans un tableur. Le nettoyage a consisté à ne garder qu'un nombre fixe de points par contour (Figure 3, tracés en pointillés), et à effectuer des rotations et translations pour recaler toutes les images entre elles à partir de points de référence. Nous avons gardé les points de référence dans les données car il leur restait une légère variabilité[6], et avec les points des contours, nous avons obtenu 71 points. Par exemple l'image de la figure 3 correspond à la ligne 91 de ce fichier, et les coor-

[6] Rappelons que ces radiographies de la tête ont été faites en 2D sur des personnes en 3D en train de parler, et qu'il est difficile de parler sans tourner du tout la tête !

données 189 et 138 de l'apex (point jaune situé à l'extrémité de la langue et en partie caché par la courbe rose de celle-ci) y sont écrites en colonnes 19 et 20.

Ce tableau de 671 lignes et 142 colonnes de données représente en fait 4 hyper-matrices (une par phrase prononcée, H1, H2, H3 et H4) formées chacune de 71 colonnes (les points des contours), de p lignes (p étant le nombre d'images conservées pour chaque phrase, compris entre 108 pour H4 et 241 pour H1), et de profondeur 2 (coordonnées x et y).

3.2.3 Pré-traitement avec INDSCAL

Pour pouvoir traiter ces hyper-matrices, il convenait de les transformer en matrices, en essayant de ne pas perdre les liaisons spatiales et/ou temporelles. Pour cela, nous avons choisi l'algorithme 3-modes INDSCAL de Carroll et Chang (1970), décrit dans Schiffman et al. (1981). En fournissant en entrée de l'algorithme les distances euclidiennes entre les points des images prises deux à deux, nous avons obtenu en sortie les coordonnées des points et des images dans un espace de dimension réduite respectant au mieux les distances de départ.

Pour détailler, prenons l'exemple de l'apex de la langue dans l'image H1_156 (Fig. 3, point jaune de coordonnées 189 et 138). Pour chaque séquence (H1, H2, H3 et H4), nous avons créé une matrice de distances correspondant à l'apex. Pour H1, la matrice avait autant de lignes et de colonnes que le nombre d'images présentes dans H1, soit 241. Sa ligne 91 correspondait aux distances euclidiennes entre l'apex sur l'image H1_156 et l'apex sur les autres images de la séquence H1. La matrice correspondant à l'apex pour H1 était donc une matrice carrée de taille 241^2, symétrique et de diagonale nulle,.

Nous avons ainsi obtenu pour H1 une pile de 71 matrices carrées de taille 241^2, une par point de contour, à laquelle nous avons appliqué l'algorithme INDSCAL en fixant le nombre de dimensions à 25. Il nous a fourni en sortie les coordonnées des 241 images dans un espace de dimension 25 (matrice M1 de taille 241x25), et les coordonnées des 71 contours dans cet espace (M2 de taille 71x25). Comme pour l'ACP, il n'y a pas équivalence totale entre les deux représentations de ce même espace : les axes se correspondent (ils ont la même interprétation), mais alors que les coordonnées des images dans le premier espace peuvent prendre des valeurs quelconques, positives ou négatives, les points des contours sont situés sur une hyper-sphère de rayon 1, dans l'octant de coordonnées positives ou nulles. Nous avons opéré de la même façon pour H2, H3 et H4, et nous disposons à la fin du pré-traitement des 2 matrices M1 et M2 pour chacune des 4 séries.

4 Construction du modèle des données de H3 et H4

On élabore le modèle indépendamment sur H3 et H4 afin de confronter les résultats. L'élaboration du modèle se fait en deux temps : on modélise d'abord les liaisons spatiales des contours à partir de la matrice M2, puis on ajoute au modèle la dimension temporelle à partir de la matrice M1, ce qui permet d'en faire un modèle spatio-temporel.

4.1 Extraction des concepts (points, dimensions) de H4

La matrice M2, croisant contours et dimensions, a été rendue binaire en prenant un seuil de 0,5 (toutes les valeurs supérieures ou égales à 0,5 ont été remplacées par 1, les autres par 0). On a extrait tous les motifs fermés (A, B) de cette matrice, avec A et B non vides, et on

les a répartis en classes « indépendantes » selon le principe suivant : si un motif a au moins un élément commun avec un motif d'une classe (point ou dimension) alors il appartient à cette classe. On a ainsi obtenu 28 motifs fermés (voir Annexe 2) répartis en 11 classes (voir Annexe 1), utilisant toutes les dimensions sauf les quatre dimensions vides 7, 15, 24 et 25.

4.2 Interprétation des classes de motifs de H4

On a représenté en figure 4 les 28 motifs en faisant une marque de même couleur sur chaque point d'un même motif. Puis les points ont été joints chaque fois qu'ils correspondaient à un même contour, ce qui s'est produit dans la quasi-totalité des cas, excepté les balises, qui se trouvaient en dehors des contours, et le point epi1_4. On a ensuite indiqué à côté de ces points les numéros des dimensions du motif par des nombres écrits de la même couleur. Un seul numéro de dimension a suffi par motif, sauf pour le motif 26, correspondant au premier point de la lèvre supérieure, indiqué par un carré de 2 couleurs (18 et 22).

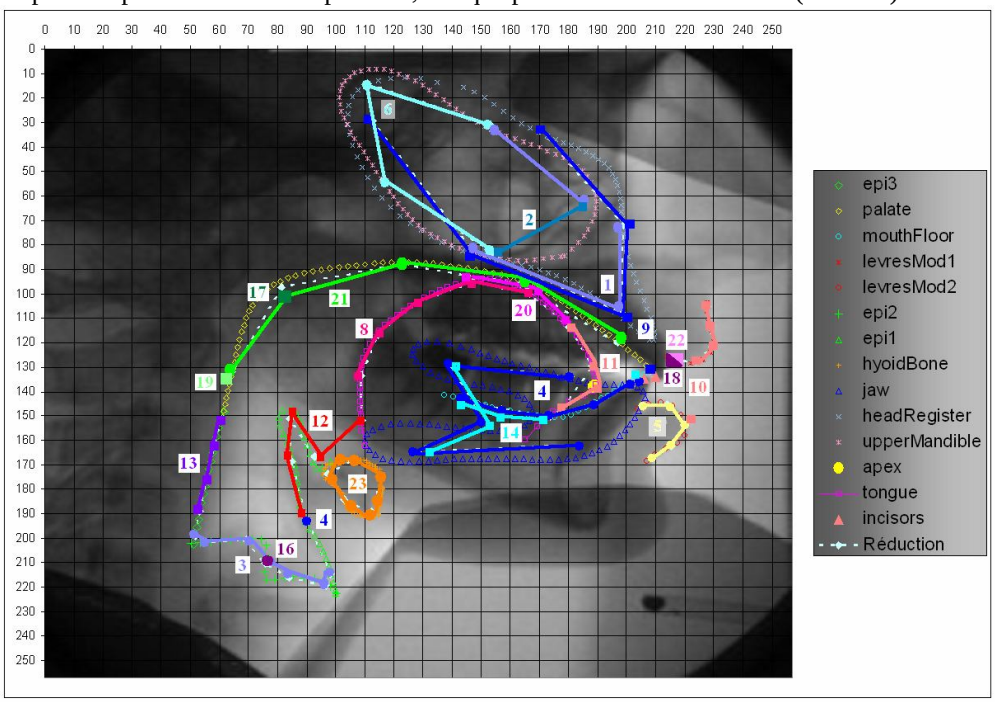


FIG. 4 – *Représentation par des points et traits de couleur des 28 motifs extraits des données de H4. Le numéro à côté des points indique les dimensions des motifs.*

Détaillons la représentation sur ce graphique de la classe 4, intitulée *Mâchoire et partie avant de l'épiglotte*. Elle est formée des 4 motifs, numéros 17, 18, 19 et 20, et des dimensions 4, 12 et 14. La dimension 12 est indiquée en rouge, elle correspond au motif 18, formé des quatre points epi1_(1 2 3 4) situés au début de la première partie de l'épiglotte et du point tong_1 qui leur est contigu, situé à la racine de la langue. La dimension 4 est en bleu foncé, elle correspond au motif 20, formé de 13 points dont les 5 points J_(1 2 3 4 5) du contour de la mâchoire (une sorte de Σ recouvrant les points de « jaw »), les 5 points MtF_(1

2 3 4 5) du plancher de la langue (un contour en forme de coupe aplatie traversant le Σ de la mâchoire), et plus bas, à la suite de la courbe rouge représentant le motif 18, le point bleu de epi1_4. Le point TgO qui fait partie des 13 points de la dimension 4 est une balise fixée dans la mâchoire, qui n'a pas été représentée, et le point LIn, qui est celui de l'incisive inférieure, a été marqué mais il se confond avec le plancher de la langue, qu'il prolonge. On voit que le point epi1_4, marqué par un point rouge et un point bleu, est commun aux dimensions 12 et 4, il forme à lui tout seul le motif 17. La dimension 14, de couleur cyan, correspond au motif 19 formé d'une partie des points du motif 20. Il a deux dimensions, 4 et 14, mais comme 4 a déjà été écrit (en bleu foncé), il ne reste plus qu'à écrire 14 en cyan également. L'inclusion du motif 19 dans le motif 20 est révélée sur le graphique par les points et des traits de couleur bleue doublés par des points et traits de couleur cyan. Ainsi cette classe 4 réunit une partie non déformable formée des points de la mâchoire et une partie déformable qui est l'avant de l'épiglotte. Ces parties seraient séparées sans l'intervention du point epi1_4, commun aux deux.

Les dix autres classes nous paraissent plus cohérentes tant du point de vue spatial qu'articulatoire. La classe 1 contient le haut de la tête, partie rigide qui ne devrait pas bouger. La classe 2, contient les 5 motifs de la langue correspondant à des zones de points contigus plus ou moins grandes en interaction. La classe 3 contient les lèvres, avec un motif pour les 5 points de la lèvre inférieure, interagissant légèrement avec la lèvre supérieure. La classe 5 contient le palais dur, mais pas le voile du palais, plus mobile, qui forme la classe 6 à lui tout seul. L'indépendance entre les deux parties du palais s'expliquerait plus par la respiration que par les phrases prononcées dans H4, qui ne contiennent aucun son nasal. La classe 7 est celle de l'os Hyoïde, séparée des classes de l'épiglotte qui lui sont proches spatialement. La classe 8 et 10 sont respectivement la partie arrière et la partie horizontale de l'épiglotte. Quand aux classes 9 et 11, elles sont réduites à un motif lui-même réduit à 1 point, le premier point de la lèvre supérieure, près de l'incisive et le point du milieu de l'épiglotte horizontale, dans la zone des cordes vocales.

4.3 Projection des classes et motifs de H4 dans l'espace-temps

On utilise la matrice M1 de H4, donnant les coordonnées des images dans les 25 dimensions, et on représente les suites d'images dans les dimensions de chaque classe.

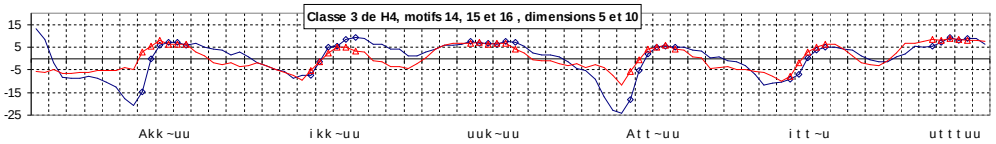


FIG. 5 – *Représentation de la classe 3 de H4 correspondant aux lèvres en fonction du temps. En abscisse les sons correspondant à chaque image et en ordonnée la valeur de la dimension pour l'image. En rouge la lèvre inférieure, en bleu le milieu de celle-ci et la lèvre supérieure.*

La classe 3 des lèvres est donnée en figure 5. On voit que les 3 premières séquences diffèrent très peu des 3 suivantes, alors qu'elles diffèrent entre elles. Cette classe permettrait donc de différencier les voyelles /A , /i/ et /u/, mais pas les consonnes /k/ et /t/. En bleu, la dimension 5 indique le mouvement de la lèvre inférieure pendant l'ouverture de la bouche, grande pour /A/, moyenne pour /i/, et nulle pour /u/. En rouge, la dimension 10 indique le mouvement de la lèvre supérieure et du point de la lèvre inférieure situé en face, pendant

l'étirement de la bouche, vu de profil, légèrement plus grand pour /i/ que pour /A/ et nul pour /u/. Les autres classes s'interprètent de la même façon, mais avec moins de facilité, car il n'y a qu'un petit nombre de phonèmes disponibles et plus de 20 dimensions.

4.4 Confrontation entre H3 et H4

On reproduit pour H3 ce qu'on a fait pour H4, de façon indépendante, et on obtient des résultats de même ordre : 33 motifs, 10 classes, 22 dimensions. Les motifs lient également des points situés sur un même contour, ou sur des contours contigus. Par contre la distribution des points des contours dans les motifs et les classes s'est souvent faite différemment. Par exemple l'os hyoïde est maintenant dans la classe de l'épiglotte, la lèvre inférieure et la lèvre supérieure forment deux classes séparées, ainsi que le haut de la tête. La classe de la mâchoire ne contient plus que les 12 points des structures rigides, elle n'est plus liée à celle du début de la partie 1 de l'épiglotte par le point epi1_4. La confrontation des dimensions les plus interprétables, comme celles des lèvres par exemple, montre que le modèle obtenu à partir des motifs de H3 est très proche de celui obtenu à partir de H4. Pour comprendre l'origine des différences entre distributions des points d'un même contour selon H3 et H4, nous avons étudié le cas du palais, qui a fourni pour H4 trois motifs dont deux liés contre deux motifs liés pour H3. Nous sommes retournées aux données, et nous avons découvert que pour H4, le contour du palais avait été obtenu à partir de 10 images de référence contre 2 pour H3 ce qui suffisait à expliquer la différence de décomposition entre H4 et H3.

5 Conclusions et perspectives

Nous avons montré la construction en deux étapes d'un modèle articulatoire de locuteur à base de motifs de points extraits d'une cinéradiographie : d'abord une analyse 3-modes d'une séquence d'images, puis une extraction de motifs des données résultant de l'analyse..Nous avons commencé à vérifier que ce modèle était opératoire dans les domaines à la fois spatiaux et temporel : il retrouve certaines relations connues comme celles entre mouvements des lèvres et production des trois voyelles /A/, /i/ et /u/ à travers deux des dimensions détectées. Il reste maintenant à explorer les autres dimensions, afin d'ouvrir de nouvelles pistes dans la description de la parole, et à raffiner ce modèle dans plusieurs directions : par ex. utiliser plus de 9 points pour la langue, et plus de 5 points pour les autres contours déformables, construire un modèle sur les séquences H1 et H2, puis sur d'autres cinéradiographies.

Références

Agrawal, R. and Srikant, H, (1994) *Fast algorithms for mining association rules in large databases*, Research Rep. RJ 9839, IBM Almaden Research Center, San Jose, California.

Busset, J. (2007). *Analyse acoustique-articulatoire des fricatives*. Mémoire de Master Ingénierie Mathématique et Outils Informatiques (IMOI). Université Nancy 1.

Busset, J. (2009). Utilisation d'une grille polaire adaptative pour la construction d'un modèle articulatoire de la langue. *Rencontres Jeunes Chercheurs en Parole (RJCP 2009)*, Avignon, France.

Cadot M. (2006). *Extraire et valider les relations complexes en sciences humaines: statistiques, motifs et règles d'association*. Thèse, université de Franche-Comté.

Flanagran J. L. (1972). *Speech Analysis, Synthesis and Perception*. Springer-Verlag. NY.

Guigues J.L. et Duquenne V. (1986) Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Math. Sci. Hum.* n°95, pp. 5-18

Han J. and Kamber M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco.

Haton, J.-P., Cerisara, C., Fohr, D., Laprie, Y., et Smaïli, K. (2006). *Reconnaissance automatique de la parole. Du signal à l'interprétation*. Dunod, Paris.

Laprie, Y., et Busset, J. (2011). Construction and evaluation of an articulatory model of the vocal tract, *19th European Signal Processing Conference - EUSIPCO-2011*, Barcelona, Espagne.

Luxenburger M. (1991). Implications partielles dans un contexte, *Math. Sci. Hum.* n°113, pp. 35-55

Maeda, S. (1990), Compensatory articulation during speech : Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model, dans *Speech production and speech modelling*, W. Hardcastle and A. Marchal, Eds. Amsterdam : Kluwer Academic Publisher, vol. 4, pp. 131–149.

Pasquier N., Bastide Y., Taouil R., Lakhal L. (1999), Efficient mining of association rules using closed itemset lattices, *Journal of Information Systems*, vol. 24(1), pp. 25-46.

Schiffman, S. S., Reynolds, M. L., and Young, F. W. (1981). *Introduction to Multidimensional Scaling*. London : Academic Press.

Article « coarticulation », http://fr.wikipedia.org/wiki/Coarticulation, 14/10/2012, 11h45.

Annexe 1

Liste des 11 classes indépendantes de motifs de H4. Elles ont été nommées en fonction des points qu'elles contiennent.

1. *Haut de la tête*, comportant les 5 points de HeadRegister, les 5 points de UpMandible, l'incisive supérieure (HIn), et une balise (ArO) : motifs 1, 2, 3, 4, 5, 7 et 8, dim. 1, 2, 6 et 9.
2. *Langue*, comportant les 9 points de sauf le premier (Tg_2 à Tg_9, Apex) : motifs 9, 10, 11, 12 et 13, dim. 8, 11 et 20.
3. *Lèvres*, comportant les 5 points de la lèvre inférieure (lev2) et tous les points de la lèvre supérieure, sauf le premier (lev1_2 à lev1_5) : motifs 14, 15 et 16, dim. 5 et 10.
4. *Mâchoire et partie avant de l'épiglotte*, comportant les 5 points de la mâchoire (Jaw), les 5 points du plancher de la langue (MthFloor), les 4 points de l'épiglotte les plus près de la mâchoire (epi1_1 à epi1_4), la racine de la langue (Tg_1), l'incisive inférieure (LIn), et une balise qui est l'origine de la grille (TgO) : motifs 17, 18, 19 et 20, dim. 4, 12 et 14.

5. *Palais dur*, comportant 4 des 5 points du palais (Pal_1 à Pal_5 sauf Pal_4) : motifs 21 et 22, dim. 19 et 21.
6. *Voile du palais*, comportant un seul point, Pal_4 : motif 23, dim. 17.
7. *Os hyoïde*, comportant 7 points, les 5 du contour (Hy_1 à Hy_5), et les 2 extrémités (TopH et BotH) : motif 24, dim. 23.
8. *Partie arrière de l'épiglotte*, comportant les 4 points les plus hauts (epi3_2 à epi3_5) : motif 25, dim. 13.
9. *Dessous de la lèvre supérieure*, comportant le premier point de la lèvre lev1_1 : motif 26, dim. 18 et 22.
10. *Partie horizontale de l'épiglotte*, avec tous les points de la partie horizontale epi2_1 à epi2_5 sauf celui du milieu, et les points les plus proche de chaque partie verticale epi1_5 pour la partie avant et epi3_1 pour la partie arrière : motif 27, dim. 3.
11. *Milieu de la partie horizontale de l'épiglotte*, le point epi2_3 : motif 28, dim. 16

Annexe 2

Liste des 28 motifs de H4. Pour chaque motif, on a indiqué les points des contours entre crochets, puis les dimensions entre parenthèses.

1. [HRR_(2 3 4) HIn] (dim. 1 et 9)
2. [UpM_(4)] (dim. 1 et 6)
3. [UpM_(5)] (dim. 1 et 2)
4. [UpM_(1), ArO] (dim. 2 et 6)
5. [HRR_(1 2 3 4 5) HIn] (dim. 9)
6. [HRR_(2 3 4) UpM_(4 5) HIn] (dim. 1)
7. [UpM_(1 5) ArO] (dim. 2)
8. [UpM_(1 2 3 4) ArO] (dim. 6)
9. [Tongue_(5 6)] (dim. 8 et 20)
10. [Tongue_(7)] (dim. 11 et 20)
11. [Tongue_(2 3 4 5 6)] (dim. 8)
12. [Tongue_(5 6 7)] (dim. 20)
13. [Tongue _(7 8 9) Apex] (dim. 11)
14. [lev2_(3)] (dim. 5 et 10)
15. [lev2_(1 2 3 4 5)] (dim. 5)
16. [lev1_(2 3 4 5) lev2_(3)] (dim. 10)
17. [epi1_(4)] (dim. 4 et 12)
18. [epi1_(1 2 3 4) Tongue_(1)] (dim. 12)
19. [Jaw_(2 3 4) MtF_(3 4 5) TgO (dim. 4 et 14)
20. [Jaw_(1 2 3 4 5) MtF_(1 2 3 4 5) epi1_4 TgO LIn] (dim. 4)
21. [Pal_(5)] (dim. 19 et 21)
22. [Pal_(1 2 3 5)] (dim. 21)
23. [Pal_(4)] (dim. 17)
24. [Hy_(1 2 3 4 5) THy BHy] (dim. 23)
25. [epi3_(2 3 4 5)] (dim. 13)
26. [lev1_(1)] (dim. 18et 22)
27. [epi1_(5) epi2_(1 2 4 5) epi3_(1)] (dim. 3)
28. [epi2_(3)] (dim. 16)

Summary

For several reasons it is difficult to analyze the sequences of radiographs of a person talking. The first is technical: these data are images annotated in several places, times, and semi-automatically or manually. The second is representational: the movements of the articulators during speech (tongue, jaw, etc.) are complex to describe because of multiple mechanical and dynamic interdependencies. We present in this paper the extraction of an articulatory model of speech from the data without adding a priori knowledge, using a method of data mining. This model reveals the organization of articulatory structures in both the spatial dimension and the temporal dimension. The comparison of this model to expected movements of the articulators by the expert is a success and invites us to continue along this path.

Fouille de données spatio-temporelles : des données aux motifs

Maguelonne Teisseire*, Sandra Bringay**

*TETIS, ISTREA **LIRMM, Université de Montpellier

Dans le processus d'extraction de connaissances, traiter à la fois la temporalité et la spatialité augmente la complexité des méthodes à mettre en œuvre. La prise en compte de la spatialité correspond à la localisation précise d'un objet (e.g. coordonnées GPS) mais également à une position relative à un autre objet (e.g. au nord de cet objet). L'objectif de cet exposé est de présenter le processus d'extraction de connaissances à partir de données spatio-temporelles. Dans un premier temps, nous présenterons les typologies existantes de données spatiales et les conséquences sur les motifs à extraire. Dans un second temps, nous nous focaliserons sur la fouille de données d'objets mobiles. Nous montrerons qu'il existe pléthore d'approches et qu'il devient indispensable de proposer une méthode uniforme d'extraction. Nous conclurons cet exposé par les nouveaux challenges auxquels la communauté est actuellement confrontée.

Index

Partenaires :