Zoé Faget

Docteur en Informatique de l'Université Paris-Dauphine
Docteur en Mathématiques de l'Université Paris VI
née le 29 aout 1977
nationalité française

Tel : 06 76 60 71 76
Email : zoe@armadillo.fr
Adresse : 1 rue des Gâtines, 75020 Paris

Sujets de recherche

Bases de données, modèles de données, séquences temporelles, recherche par contenu, indexation, masses de données, données hétérogènes, recherche d'information musicale.

Diplomes Universitaires

2011 : Thèse de Doctorat d'Informatique

Titre :	Un modèle pour la gestion de séquences temporelles synchronisées. Application aux données musicales symboliques.
Directeur :	Philippe Rigaux
Date :	6 décembre 2011
Lieu :	Université Paris Dauphine (laboratoire LAMSADE, École Doctorale EDDIMO)
Rapporteurs :	Dominique Laurent (Univ. Cergy Pontoise) Bruno Defude (Télécom Paris Sud)
Jury :	Cédric du Mouza (CNAM) Geneviève Jomier (Univ. Paris Dauphine), président du Jury Philippe Rigaux (CNAM), directeur de thèse

L'École Doctorale de Dauphine ne délivre pas de mention.

2002 : Thèse de Doctorat de Mathématiques

Titre :	Meilleures constantes dans les inégalités de Sobolev pour des fonctions invariantes par un groupe d'isométries.
Directeur :	Michel Vaugon
Date :	11 avril 2002
Lieu :	Université Paris VI (Institut Mathématiques, Équipe Géométrie et Dynamique)
Mention :	Très Honorable
Rapporteurs :	Frédéric Helein, président du Jury Franck Pacard (Univ. Paris XII)
Jury :	Emmanuel Hebey (Univ. Cergy Pontoise) Gennadi Henkin (Univ. Paris VI) Harold Rosenberg (Univ. Paris VII) Michel Vaugon (Univ. Paris VI), directeur de thèse

1999 : DEA Analyse Géométrie et Modélisation de l'Université Paris VI, mention Bien (directeur de mémoire Michel Vaugon)
1998 : Maîtrise de Mathématiques Pures de l'Université Paris VI, mention Assez Bien
1997 : Licence de Mathématiques Pures de l'Université Paris VI, mention Assez Bien
1996 : DEUG Mathématiques et Informatique (MIAS) de l'Université Paris VI, mention Assez Bien

Publications

Conférences nationales ou internationales avec comité de lecture (informatique)

- Indexing Symbolic Music Score (avec C. Constantin, C. du Mouza et P. Rigaux), *Bases de Données Avancées (BDA'11)*, 2011
- The Melodic Signature Index For Fast Content-Based Retrieval of Symbolic Scores (avec C. Constantin, C. du Mouza et P. Rigaux), *Proc. Intl. Society for Music Information Retrieval (ISMIR'11)*, 2011
- Modeling Synchronized Time Series (avec D. Gross-Amblard, P. Rigaux et V. Thion-Goasdoué), *Proc. Intl. Databases Engineering and Applications Symposium (IDEAS'10)*, 2010.
- A Database Approach to Symbolic Music Content Management (avec P. Rigaux), *Proc. Intl. Symposium on Computer Music Modeling and Retrieval (CMMR'10)*, 2010.
- The Design and Implementation of Neuma, a Collaborative Digital Score Library (avec L. Abrouk, H. Audéon, N. Cullot, C. Davy-Rigaux, D. Gross-Amblard, P. Rigaux, A. Tacaille, E. Gavignet et V. Thion-Goasdoué), *Int. Jour. of Digital Libraries (IJDL'10)*, 2010.
- The Neuma Project : towards On-line Music Score Libraries (avec L. Abrouk, H. Audéon, N. Cullot, C. Davy-Rigaux, D. Gross-Amblard, H. Lee, P. Rigaux, A. Tacaille, E. Gavignet et V. Thion-Goasdoué), *Proc. Intl. Workshop on Exploring Music Information Spaces (WEMIS'09)*, 2009.

Revues d'audience internationale avec comité de rédaction (mathématiques)

- Optimal constants in the exceptional case of Sobolev inequalities *Transactions of the American Mathematical Society* vol 360, 2303–2325, 2008.
- Best constants in the exceptional case of Sobolev inequalities, *Mathematische Zeitschrift* 252, 133–146, 2006.
- Second best constant and extremal functions in Sobolev inequalities in the presence of symmetries, *Advances in Differential Equations* 9, 745–770, 2004.
- Optimal constants in Sobolev inequalities in the presence of symmetries, *Annals of Global Analysis and Geometry*, 161–200, 2003.
- Best constants in Sobolev inequalities on Riemannian manifolds in the presence of symmetries, *Potential Analysis*, 17, 105–124, 2002.

Ouvrage collectif

Mathématiques L3 - Analyse (Chapitres Topologie et Calcul Différentiel), Pearson Education, 2009.

Postes

Depuis 2008 : Société Armadillo
Projet ANR Neuma : Projet ANR Neuma, liaison université-entreprise.
Thèse d'informatique au Lamsade.
Recherche et développement : Approfondissement de la technologie d'indexation Armadillo. Conception et implantation de nouveaux index.
Autre responsabilités : Encadrement et formation stagiaires et apprentis en interne.
2005-2006 : Post Doc à l'ENS ULM
Equipe Biologie et Mathématiques.
2004-2005 : Post Doc à l'ETH Zürich
Département Mathématiques.
2002-2004 : ATER à l'Université Paris XII
1999-2002 : Allocataire recherche de l'Université Paris VI, Monitorat à l'Université Paris XII
Thèse de mathématiques à l'Institut Mathématiques, équipe Géométrie et Dynamique.

Enseignement

2009-2010 : Arithmétique L2 à l'Université Paris VI *(vacation)*
Cours magistraux et TD en L2, organisation du contrôle continu.
2007-2008 : Géométrie Différentielle L3 à l'Université Paris VI *(vacation)*
Travaux dirigés en L3, organisation du contrôle continu.
2004-2005 : Assistante à l'ETH Zürich *(post-doctorat)*
Travaux dirigés Analyse III, élèves ingénieurs en électronique 2ème année.
Travaux dirigés Analyse I, élèves ingénieurs en électronique 1ère année.
2003-2004 : ATER à l'Université Paris XII
Cours magistraux et TD en DEUG MASS deuxième année (algèbre), responsable de tout l'enseignement, organisation du contrôle continu et des examens finaux, rédaction des sujets et corrections de copies.
2002-2003 : ATER à l'Université Paris XII
Enseignement en DEUG MASS première et deuxième année (analyse), organisation des TD, encadrement des moniteurs, rédaction des sujets et contrôle continu, corrections de copies.
1999-2002 : Monitorat à l'Université Paris XII
Enseignement en DEUG MASS première et deuxième année (analyse), TD et examens oraux (colles).

Conférences et Séminaires

Conférence d'audience nationale ou internationale avec comité de sélection (orateur)

– Octobre 2011, Bases de Données Avancées (BDA'11), Rabat, Maroc
– Aout 2010, Int. Database Engineering and Application Symposium (IDEAS'10), Montreal, Canada
– Juin 2010, Computer Music Modeling and Retrieval (CMMR'10), Malaga, Espagne

Séminaires internes (orateur)

– Octobre 2011, Séminaire du Lip6, Paris VI
– Septembre 2011, Séminaire de l'Eddimo, Paris Dauphine
– Octobre 2010, Séminaire de l'Eddimo, Paris Dauphine
– Décembre 2005, Séminaire de Géométrie, Université de Tours
– Avril 2005, Séminaire de Mathématiques, Université de Reims
– Avril 2005, Séminaire d'Analyse Complexe et Différentielle, Université de Lille
– Mars 2005, Séminaire d'Analyse Appliquée, Université d'Amiens
– Février 2005, Séminaire de Géométrie et d'Analyse, Université de Nice
– Novembre 2004, Séminaire d'Analyse, ETH Zürich, Suisse
– Juillet 2004, Séminaire de Géométrie, Universität Frieburg, Allemagne
– Juin 2003, Séminaire d'Analyse et Géométrie, Université Paris VI
– Janvier 2003, Séminaire de Géométrie Spinorielle, IEC Nancy
– Mai 2001, Séminaire d'Analyse et Géométrie, Université Paris VI

Congrès et écoles d'été (participation sans exposé)

– Janvier 2011, Conférence Web Sémantique SemWeb Pro, Paris
– Mai 2010, École d'été BDA Masse de données distribuées, Les Houches

Résumé des activités de recherche et d'enseignement

Zoé Faget

Mon travail de thèse, réalisé sous la direction de Philippe Rigaux, s'est déroulé dans le cadre du projet NEUMA, projet ANR impliquant plusieurs partenaires universitaires pluri-disciplinaires (LAMSADE et CNAM : base de données, Le2i : web collaboratif, ontologies et protection des données, IRPMF et Sorbonne : musicologues) et une entreprise (Armadillo). Ce projet d'une durée de trois ans avait pour but de mettre en place une plateforme collaborative de gestion de contenu musical symbolique destinée à des communautés d'utilisateurs experts (musicologues, historiens de la musique...) [1, 2]. Cette plateforme est aujourd'hui opérationnelle[1].

Dans le cadre de ce projet, je me suis intéressée au problème de la représentation et de l'interrogation des séquences temporelles, et plus précisément aux données musicales. Mon premier ensemble de résultats consiste en la définition d'un modèle de données pour la gestion des séquences temporelles synchronisées, ainsi qu'un langage d'interrogation de données reposant sur ce modèle. Un second résultat est la définition d'un index permettant de réaliser plusieurs types de recherches sur du contenu musical symbolique.

1 Un modèle de données pour la gestion des séquences temporelles synchronisées

Les travaux exposés dans cette partie ont fait l'objet de deux publications dans des conférences internationales avec comité de lecture [6, 7].

La première partie de mon travail de thèse consiste en la définition d'un modèle pour la gestion des séquences temporelles synchronisées. Ce modèle, initialement prévu pour la gestion du contenu musical symbolique, se révèle dépasser largement le domaine musical et peut être utilisé dans de nombreuses applications faisant intervenir des données temporelles hétérogènes, synchronisées ou non.

1.1 Définition du modèle

Le modèle présenté est une extension du modèle relationnel classique permettant d'inclure un type nouveau, le type séquence temporelle. Pour cela on définit tout d'abord formellement la notion de séquences temporelles synchronisées, qui peuvent être vues comme des fonctions d'un domaine temporel (isomorphe à $\mathbb{N}$) vers un domaine d'application qui peut être simple ou complexe. Une relation du modèle est ensuite définie comme une combinaison d'attributs classiques et de séquences temporelles. On définit de plus une algèbre pouvant opérer sur ce modèle, c'est à dire un ensemble d'opérateurs qui consomment une ou plusieurs instances du modèle et produisent des instances du

[1] http ://www.neuma.fr

modèle. L'algèbre opérant de manière fermée, il est ainsi possible de composer les opérateurs et d'obtenir une grande expressivité. De nouveaux opérateurs, répartis en deux algèbres, sont présentés : les opérateurs relationnels classiques qui sont étendus au type séquence temporelle (algèbre relationnelle), et trois nouveaux opérateurs spécifiquement destinés à la manipulation des séquences temporelles (algèbre temporelle). Précisément, ces opérateurs sont :

- la composition $\circ$, qui permet de manipuler le domaine temporel,
- l'addition $\oplus$, qui permet de propager au niveau de la séquence une fonction utilisateur définie au niveau de l'élément,
- la dérivation/aggrégation **A**, qui permet de travailler sur des vues locales d'une séquence temporelle.

Ces opérateurs sont paramétrables par des fonctions utilisateurs, ce qui assure la stabilité du système, les opérations spécifiques n'ayant pas besoin d'être redéfinies au cas par cas.

1.2 Langage de requête

L'algèbre décrite dans la partie précédente étant une extension de l'algèbre relationnelle classique, il est possible de définir un langage de requête associé en s'inspirant du SQL. Une requête se présente sous la forme :

from	liste de tables
let	variable := expression
construct	expressions
where	condition de sélection

On démontre que chaque expression du langage peut être traduite par une expression algébrique équivalente, en construisant l'algorithme de traduction par induction sur les sous-expressions. La sémantique des opérateurs ayant déjà été définie, ceci nous donne la sémantique du langage.

Réciproquement, les opérateurs de notre algèbre pouvant s'intervertir et se rassembler de la même manière que les opérateurs classiques, toute expression algébrique peut se ramener à une expression de la forme

$$\Pi_{e^\star[b^\star]}\sigma_{P[b^\star]}(T^\star),$$

que l'on sait traduire dans le langage. On a donc l'équivalence du langage et de l'algèbre.

Pour valider l'utilisabilité du langage, une implantation a été réalisée en OCaml. Par implantation on entend l'écriture d'un interprète lisant en entrée des requêtes et des définitions de fonctions, effectuant le travail demandé et retournant à l'utilisateur le résultat (éventuellement vide) ou un message d'erreur en cas de problème.

1.3 Conséquences

L'introduction du type séquence temporelle et des nouveaux opérateurs nous permet d'exprimer simplement des requêtes qu'il eût été plus difficile d'exprimer en algèbre relationnelle classique, comme illustré dans l'exemple suivant.

Exemple de requête : On représente les variations heure par heure d'un indice boursier par une séquence temporelle. On veut déterminer l'heure à laquelle vendre une action achetée le jour même permet de réaliser le meilleur profit (la vente précédant l'achat étant interdite). Ceci revient à déterminer pour chaque instant t la différence entre le prix à l'instant t et le prix minimal de l'action jusqu'à l'instant t. La requête s'exprime très simplement dans notre langage :

from Trades
let $inv := **comp**(stock, inv^{24})
let $run_diff := **derive**($inv, *shift*, first()-min())
construct $run_diff
where date = 12/6/2011

2 Un index pour l'interrogation du contenu musical symbolique

Les travaux exposés dans cette partie ont fait l'objet de deux publications dans des conférences nationales et internationales avec comité de lectures [3, 4].

Le deuxième ensemble de résultats de ma thèse est l'extension aux séquences musicales d'un index existant (AS-index [5]), repris et adapté aux spécificités des données musicales symboliques. Cet index, rebaptisé MS-index (*Music Signature-index*), permet de réaliser plusieurs types de recherche sur du contenu musical symbolique (exacte, transposée, avec ou sans rythme, approchée). La particularité de cet index réside dans l'encodage d'informations multiples pouvant intervenir dans les différents types de recherche. Il repose entre autre sur le concept mathématique de signature algébrique.

2.1 Types de recherche

Dans une collection de partitions symboliques, il est possible de réaliser plusieurs types de recherches sur du contenu musical : recherche exacte (succession des notes), transposée (succession des intervalles), avec ou sans rythme, approchée (pour un motif P donné, trouver les fragment f_p d'une partition tels que $d(P, f_p) < \tau$ où d est une distance et τ un seuil de tolérance définis par l'utilisateur). Ces différents types d'interrogation sont intéressants d'un point de vue musical comme d'un point de vue scientifique. Le MS-index permet d'indexer tous ces types de recherche grâce à une seule structure d'index.

2.2 Indexation et recherche

Le contenu musical est représenté par une séquence de couples (note-durée), vus comme des éléments d'un corps fini. Le principe de l'index est le suivant : une table de hachage dont les clés sont des fragments musicaux de taille constante n appelés n-grams, et dont la fonction de hachage fait intervenir la signature algébrique de ces n-grams.

Les signatures algébriques des n-grams extraits successivement d'une séquence musicale se calculent de manière incrémentale et vérifient des propriétés mathématiques très élégantes sur lesquelles reposent l'efficacité de l'index.

Pour chaque n-gram indexé, on enregistre un sextuplet d'information le décrivant (identifiant de la séquence dont le n-gram est extrait, position dans la séquence, note la plus basse, plusieurs signatures). Ces informations sont suffisantes pour réaliser tous les types de recherche pré-cités.

Pour réaliser une recherche exacte ou transposée avec ou sans rythme, on montre que deux appels à l'index suffisent, quelle que soit la taille du motif recherché. Ceci est du aux propriétés mathématiques des signatures algébriques.

Pour réaliser une recherche approchée, on choisit d'utiliser la distance n-gram [8] comme approximation de la distance d'édition. Le principe de la recherche approchée est de rechercher dans la base en utilisant l'index tous les n-grams (et leurs positions) qui sont présents dans le motif cherché. On parcourt ensuite la liste des n-grams et de leurs positions en appliquant une fenêtre glissante dont la taille peut être bornée. Pour chaque position de la fenêtre, on calcule la plus longue sous-séquence réalisant la meilleure distance. L'intérêt d'utiliser la distance n-gram comme approximation est qu'elle peut être indexée par le MS-index, en revanche, elle renvoie quelques faux-négatifs (*i.e.* des fragments musicaux vérifiant $d_{edit}(P, f_p) < \tau$ sont écartés à tort).

Ce mécanisme d'indexation propre aux données musicales est donc flexible (il permet de réaliser plusieurs types de recherches), compact (il y a peu d'information à enregistrer), et efficace (notamment dans la recherche de motifs longs).

3 Perspectives et travaux futurs

Les travaux initiés dans mon travail de thèse peuvent se prolonger dans différentes directions, aussi bien théoriques que pratiques.

Propriétés des opérateurs et optimisation : Les opérateurs de l'algèbre conservent les mêmes propriétés d'interversion et de regroupement que les opérateurs de l'algèbre relationnelle classique, la seule différence étant que les formules de manipulation de données ont été enrichies (notamment puisque l'on fait intervenir des séquences temporelles). Il serait intéressant d'étudier plus en détail les propriétés de ces opérateurs dans le but d'optimiser les requêtes.

Indexation : L'index introduit permet de conduire des recherches exactes sur des données symboliques. Différents paramètres influencent les performances de l'index, tels que la taille de l'alphabet dans lequel les données prennent leurs valeurs. Sur différents jeux de données réelles (alphabet court : ADN ; alphabet long : texte), l'étude des performances de l'index peut mettre en évidence l'influence de ce facteur.

Par ailleur, cet index permet de faire de la recherche approchée en utilisant la distance n-gram comme approximation de la distance d'édition. On a pu vérifier de manière empirique que cette technique ne provoquait qu'un nombre négligeable de faux-négatifs. Une caractérisation de ces faux-négatifs permettrait une utilisation plus maîtrisée de cette approximation.

Recherche par similarité : La recherche par similarité dans une grande collection permet de retrouver des familles de séquences similaires entre elles (la notion de similarité dépendant du contexte). Cette recherche qui présente un intérêt évident n'est pour le moment pas supportée par notre index.

Applications aux domaines non-musicaux : Enfin, quoique développés dans le cadre d'une application traitant des données musicales, tous ces travaux s'appliquent à d'autres domaines. Le modèle de données et le langage associé permettent d'interroger simplement des masses de données complexes, hétérogènes et atypiques. Les applications financières, boursières, scientifiques, etc. qui manipulent des grands volumes de données évoluant dans le temps peuvent bénéficier de l'utilisation d'un tel modèle. De plus, les différents types de recherches supportées par le MS-index (transposées, avec ou sans rythme) n'ont pour l'instant qu'un sens musical. Il est néamoins possible de leur donner un sens dans d'autre contexte, notamment en biologie, où la recherche de séquence ADN (mutation de gènes) est fréquente et cruciale.

4 Enseignement et autres activités

Ayant commencé ma carrière universitaire par une première thèse en mathématiques, j'ai occupé successivement un poste de monitrice pendant trois ans puis d'ATER pendant deux ans à l'Université Créteil-Paris XII, en première et deuxième année de DEUG MASS. Durant ces cinq années, mes responsabilités ont augmenté progressivement et j'ai pu explorer les divers aspects de cette activité : chargée de TD, cours magistraux, correction de copies, rédaction de sujets, encadrement des moniteurs... Le contenu des cours était l'analyse en DEUG1, l'algèbre linéaire en DEUG2.

Lors de mon premier post-doctorat, j'ai pu m'adresser à un public d'élèves d'un niveau excellent (élèves ingénieurs de l'ETH-Zürich) pour donner des cours d'analyse en première et deuxième année.

J'ai par la suite poursuivi cette activité en tant que vacataire pour assurer les travaux dirigés de géométrie différentielle en L3 à l'Université Paris VI.

Ma dernière expérience d'enseignement était le cours d'arithmétique L2 à l'Université Paris VI en horaire aménagé, qui m'a donné l'occasion d'assurer cours et travaux dirigés à un public d'un niveau très hétérogène (adultes en reprise d'étude, étudiants suivant un double cursus ou étudiants en difficulté).

Ma thèse d'informatique s'étant déroulée en entreprise, je n'ai pas enseigné à l'Université pendant ces trois ans. Cependant, l'une de mes responsabilités au sein de mon entreprise était l'encadrement et la formation d'étudiants en apprentissage (Licence MIAGE). Ceci consistait entre autre à une introduction théorique et pratique aux bases de données.

Références

[1] L. Abrouk, H. Audéon, N. Cullot, C. Davy-Rigaux, Z. Faget, D. Gross-Amblard, P. Rigaux, A. Tacaille, E. Gavignet, and V. Thion-Goasdoué. The design and implementation of NEUMA, a collaborative digital score library. In *Int. Jour. of Digital Libraries (IJDL)*, 2010.

[2] L. Abrouk, H. Audéon, N. Cullot, C. Davy-Rigaux, Z. Faget, D. Gross-Amblard, H. Lee, P. Rigaux, A. Tacaille, E. Gavignet, and V. Thion-Goasdoué. The neuma project : Towards cooperative on-line music score libraries. In *Workshop on Exploring Musical Information Spaces (WEMIS)*, 2009.

[3] C. Constantin, Z. Faget, C. du Mouza, and P. Rigaux. Indexing symbolic music scores. In *Bases de Données Avancées (BDA)*, 2011.

[4] C. Constantin, Z. Faget, C. du Mouza, and P. Rigaux. The melodic signature index for fast content-based retrieval of symbolic scores. In *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, 2011.

[5] C. du Mouza, W. Litwin, P. Rigaux, and T. J. E. Schwarz. AS-index : a Structure for String Search Using n-grams and Algebraic Signatures. In *Proc. Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 295–304, 2009.

[6] Z. Faget and P. Rigaux. A database approach to symbolic music content management. In *Int. Symp. on Computer Music Modeling and Retrieval (CMMR)*, pages 303–320, 2010.

[7] Z. Faget, P. Rigaux, D. Gross-Amblard, and V. Thion-Goasdoué. Modeling synchronized time series. In *Int. Database Engineering and Applications Symp. (IDEAS)*, pages 82–89, 2010.

[8] E. Ukkonen. Approximate String Matching with q-grams and Maximal Matches. *Theoretical Computer Science*, 92 :191–211, 1992.