

# Mining Users Skills Development From Interaction Traces : an exploratory study

Amaury Belin  
University of Lyon, CNRS  
LIRIS, UMR5205, F-69622  
amaury.belin@liris.cnrs.fr

Grigory Antipov , Julien Blanchard,  
Fabrice Guillet and Yannick Prié  
University of Nantes, CNRS  
LINA UMR CNRS 6241,  
antigregory@gmail.com, {julien.blanchard,  
fabrice.guillet, yannick.prie }@univ-nantes.fr

## ABSTRACT

In this paper, we report an exploratory work related to the detection of user skills development from large interactions traces. We present a statistical approach to tackle the development of user's low level skills by identifying patterns of action and studying their progressive automations (i.e. as a user develops skills, they execute common subsequent actions faster) as well as their evolutions (i.e. user elaborating innovative ways to achieve a common task).

## Mots Clés

Interaction trace; user skills; data mining; frequent episode mining.

## ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

## CONTEXT

Our work is based on low-level interaction traces of a student using the video annotation *Lignes de Temps* (LDT) to carry out a research project focusing on 15 films. It was the first time she was using the software and her work spanned over 3 months. In addition to interaction traces, on-field observations and interviews allowed us to assess different phenomenon related to skills acquisition and progressive mastering of LDT by the user.

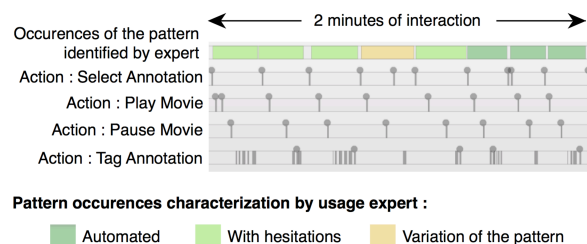
## DATA: INTERACTION TRACE

An interaction trace is composed of *obsels* representing events occurring during the interaction between a user and a particular software [3]. Each *obsel* has timing attributes (i.e. the date and duration of the represented event), as well as attributes related to its type, different *obsel types* being defined to represent the variety of events that could occur in the interaction (e.g. "playVideo", "createAnnotation", etc.). In our experiment, the complete trace contains 50,000 obsels, representing 50 hours of work spanned over the first 15 days the user spent with the software. These obsels

belong to 89 obsel types. We only considered the 34 types representing user actions, the other types being dedicated to the description of system processing in reaction to user actions (e.g. "videoStartPlaying").

## PATTERN ELICITATION

At first, we analysed interaction traces with simple visualization tools. Guided by the other materials we collected (interviews, screen recordings, etc.), we managed to detect some phenomenon related to skills acquisition in the trace [1]. In particular, we identified some *schemes of action* by the user, i.e. systematic ways to achieve a goal by reproducing a similar construction in a set of articulated actions. Some of these schemes were easily noticeable in the visualization of the interaction traces. In Figure 1, we can see that the user is reproducing the same scheme: in order to characterize a part of the movie she is working on, she selects the corresponding annotation, plays and then pauses the movie to watch the segment corresponding to the annotation, and finally enters different tags in a *text field*. We see that the user reproduces the same pattern successively, for each part of the movie she has to deal with.



**Figure 1. Visualization of an interaction trace showing a repeated pattern of actions. Each pattern occurrence is characterized by an expert using a color codification**

The apparition of such pattern is a clear hint of user appropriation of the software. However identifying every pattern is a tedious work that could be facilitated by a statistical approach. With this perspective in sight, we have chosen to use a *frequent episode mining* algorithm [4] to analyse the interaction traces. Frequent episode mining is an essential task in data mining, which allows for detecting frequent patterns (called episodes) in a single sequence of events. This first attempt provided us

with numerous patterns that we analysed given the two approaches presented below.

### EVOLUTION OF PATTERN DURATION

Analysing the occurrences of the annotation-tagging pattern (see last section) reveals that the user is progressively getting faster at chaining-up the different actions, and seems to execute the pattern in a more systematic and automatized way (see pattern characterization in Figure 1). Therefore, we studied the evolution of the duration of the occurrences of the patterns we detected, with the idea that a monotonous decreasing of an occurrence's duration of a particular pattern might point out skill acquisition regarding this pattern. We refined this approach by discarding the part of the duration that could be dependant on the content the user is dealing with. For instance, in the pattern consisting of tagging an annotation, execution time is dependant of the length of the movie segment being annotated (as the pattern involves the user to play this segment) as well as the complexity of the tags being edited. Therefore, in our measure of pattern occurrence duration, we discarded the part of the pattern that can be dependant of the content being addressed.

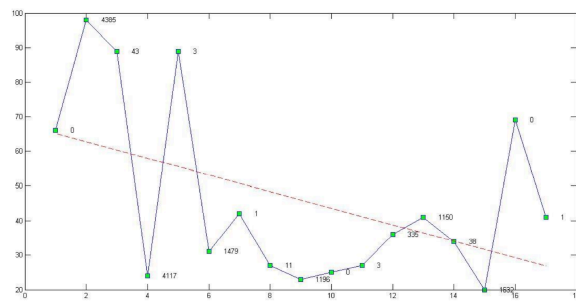


Figure 2. Evolution of the duration of a particular pattern, with regression coefficient line (in red).

For each identified patterns, we measured the *content-independent* duration of each occurrences, and calculated the regression coefficient of the distribution of these durations (see Figure 2). We analysed 60 patterns having regression models presenting potentially interesting coefficients. We did not found straightforward evident results, but indirectly this first attempt guided us to investigate part of the practice where skills acquisition might occur.

### PATTERN VARIATIONS

A pattern evolution is an interesting phenomenon as it can be a cue that the user has learned a more elaborated way to accomplish a common task. For instance, we noticed an evolution of the *annotation-tagging pattern* related above: as the user experimented crashes of the software and lost some work, she was systematically saving her work after the edition of an annotation tag, leading to a noticeable evolution of the pattern. We did therefore look for a statistical mean that would help to address such pattern evolutions. Our approach was to study the distribution of the occurrences of similar

patterns. Here we made two assumptions: if two patterns are close in their composition, they may be used for the same task; and if two patterns occurrences are grouped in different periods of the activity, it could be the expression that one more elaborated version of a pattern has replaced a former version.

In order to measure similarity between patterns, we used the Hamming distance [2]. We refined such distance by valuating substitution between actions composing the pattern based on the similarity of these actions. We then plotted occurrences of similar patterns (see Figure 3) and tried to interpret if one pattern could have replaced another in the user's practice, and if we can speak of a pattern evolution. However we didn't find clearly interesting results yet with this approach.

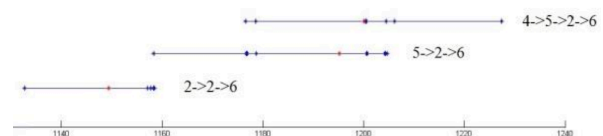


Figure 3. Distribution of the occurrences of similar patterns over several days.

### CONCLUSION

Using interaction traces as a material to study processes occurring over long period of time (like skill development in our case) is challenging. The volume of data is large and rich, and there is a real difficulty to synthetize automatically collected low-level interaction events to higher level information about user activity [3].

We presented an overview of our exploratory attempt to address such challenge, and we introduced some of the ideas we came up with. Our approach should now be refined with some heuristics to pre-filter the amount of detected patterns and focus the analysis on more fertile part of the interaction trace. It also raises numerous questions on how to interactively visualize and mine interaction traces to keep track of user skills development and, more generally, other slow processes underlying human-computer interaction.

### BIBLIOGRAPHIE

1. Belin, A. and Prié, Y. DIAM: towards a model for describing appropriation processes through the evolution of digital artifacts. *Proceedings of the Designing Interactive Systems Conference*, ACM (2012), 645–654.
2. Hamming, R.W. *Coding and information theory*. Prentice-Hall, Inc., 1986.
3. Hilbert, D.M. and Redmiles, D.F. Extracting usability information from user interface events. *ACM Computing Surveys (CSUR)* 32, 4 (2000), 384–421.
4. Mannila, H., Toivonen, H., and Verkamo, A.I. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery* 1, 3 (1997), 259–289.

# Flux de Données contre Séries Temporelles

**Pierrick Bruneau**

CRP-GL

Département ISC

41, rue du Brill

L-4422, Belvaux

bruneau@lippmann.lu

**Olivier Parisot**

CRP-GL

Département ISC

41, rue du Brill

L-4422, Belvaux

parisot@lippmann.lu

**Benoit Otjacques**

CRP-GL

Département ISC

41, rue du Brill

L-4422, Belvaux

otjacque@lippmann.lu

## RÉSUMÉ

Les usages informatiques actuels génèrent des séquences de données complexes et volumineuses (e.g. capteurs, réseaux sociaux). Pour les traiter, deux paradigmes s'imposent naturellement : considérer ces données comme des séries temporelles, et/ou comme des flux de données (i.e. *data streams* dans la littérature). Bien que semblables, ces deux principes diffèrent de manière assez sensible. Nous pensons que leur combinaison avec des techniques de visualisation et d'interaction fait émerger de nouvelles problématiques, qui sont autant de pistes de contributions intéressantes.

## Mots Clés

data streams; séries temporelles; data mining; visualisation

## ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous. Voir <http://www.acm.org/about/class/1998/> pour la liste complète des catégories ACM.

## PRÉSENTATION

Au sein du département "Informatique, Systèmes, Collaboration" (ISC) du Centre de Recherche Public - Gabriel Lippmann (CRP-GL), notre équipe dédie sa recherche au développement d'approches *visual analytics*, i.e. de combinaisons pertinentes entre méthodes de visualisation d'information et de data mining. Ces deux communautés sont historiquement assez décorréées, et notre objectif est de proposer des travaux prenant en compte l'état de l'art joint de ces deux perspectives.

Cet axe est développé dans le contexte de partenariats avec des organismes publics ou privés, voire de la participation à des projets de recherche nationaux et internationaux (e.g. FNR, FP7). En parallèle, le CRP-GL dispose de fonds publics pour son développement stratégique, dont la prospection dans notre axe de recherche fait partie intégrante.

Le CRP-GL a vocation à la recherche appliquée, aussi nous sommes toujours à l'écoute de cas d'utilisation potentiels pour notre thématique de recherche ; toutefois une partie de notre effort est également consacré à une réflexion plus fondamentale.

Notre équipe travaille sur des contributions aux domaines des *visual analytics* et du *data mining* visuel, avec en ligne de mire l'adaptation au traitement de flux de données. En parallèle d'une réflexion fondamentale sur ces sujets, nous sommes impliqués dans divers projets collaboratifs, dont un projet européen CHIST-ERA, visant à la conception d'une plateforme web collaborative de traitement et de visualisation d'annotations multimédia (Camomile), ainsi qu'une collaboration avec le département "Environnement et Agro-Biotechnologies" du CRP-GL, pour l'analyse de données réelles issues de la création d'énergie renouvelable.

## PROBLÉMATIQUES D'INTÉRÊT EN DATA MINING

Les flux de données engendrés par certaines plateformes web (e.g. Twitter, métriques Google), voire par des capteurs individuels désormais peu onéreux soulèvent de nombreuses problématiques scientifiques. En particulier, un des objectifs stratégiques de notre département est de faire avancer l'état de l'art sur le traitement de ces flux dans un contexte *visual analytics*.

L'analyse des séries temporelles est utilisée classiquement dans de nombreux domaines (e.g. traitement du signal, économétrie, météorologie). Ici, une série de valeurs (ou de vecteurs) va être utilisée "hors ligne" pour apprendre un modèle (e.g. de prédiction, ou de détection de motifs dans des cours de bourse) [4, 3]. Une fois appris, le modèle est figé, et ne peut pas tirer parti des nouvelles données arrivant dans le flux sans un réapprentissage complet. Pire, la complexité d'apprentissage de ces modèles dépend souvent de la taille du flux, au mieux linéairement. Le réapprentissage serait donc de plus en plus coûteux au fil du temps. Notons que des travaux considèrent le problème "dual", i.e. compte-tenu d'un ensemble de séries temporelles en flux, caractériser des similarités entre elles, e.g. rechercher des motifs temporels partagés avec l'algorithme *Dynamic Time Warping* (DTW) [1], ou encore estimer les similarités deux-à-deux de séries temporelles à plusieurs niveaux de résolution [8]. Cependant les problèmes de complexité évoqués plus haut sont patents dans ce cas également.

Certains de ces algorithmes sont incrémentaux, donc capables en théorie de s'adapter à un flux de taille

croissante en utilisant un minimum de ressources calculatoires supplémentaires [8, 12]. Toutefois, comme cela a été constaté dès la dérivation de la version incrémentale de l'algorithme classique k-means [9], l'ajout perpétuel de nouveaux éléments fait que l'importance des derniers arrivants décroît graduellement, jusqu'à devenir négligeable : on aurait alors un modèle progressivement figé *de facto*. Des auteurs ont proposé de restreindre le calcul de statistiques exhaustives dans une fenêtre glissante dans le contexte d'un algorithme EM [11], mais la question du paramétrage de la taille de cette fenêtre n'a pas été abordée, et semble peu évidente. Des études théoriques sur les facteurs d'oubli ont été également proposées [10], mais des contributions restent à faire pour en évaluer la valeur pratique.

Le paragraphe précédent a effleuré un dilemme qu'on peut caractériser comme suit : quel est le meilleur compromis entre la stabilité d'un modèle, et sa capacité à s'adapter ? Certains travaux récents traitent de cette question, par exemple [2], en tentant de caractériser et détecter le changement de manière non-paramétrique, s'affranchissant ainsi notamment du problème de taille de fenêtre glissante évoqué précédemment.

## TRANSCRIPTION EN TERMES DE VISUALISATION ET D'INTERACTION

Les questions présentées dans la section précédente, plutôt relatives au *data mining*, ont des transcriptions diverses dans le contexte de la visualisation d'information, et de l'interaction avec cette dernière.

Le dilemme classique *biais-variance* [6] peut être transposé à la visualisation des séries temporelles numériques : celles-ci peuvent présenter des profils de courbe très bruités, rendant leur visualisation difficile, à petite ou à grande échelle. A contrario, un lissage trop important (e.g. via une fenêtre de Parzen) risque de détruire l'information intrinsèque d'une courbe temporelle.

Un flux de données est potentiellement infini, et il faut pourtant parvenir à le représenter en utilisant une bande passante visuelle finie : des moyens de résumer visuellement les données "périmées" du flux doivent être trouvés, faisant ainsi écho au facteur d'oubli de la section précédente. Des travaux traitent de la visualisation efficace de séries temporelles [5], mais en considérant rarement un tel facteur d'oubli.

Maintenant imaginons qu'il existe un modèle capable de s'adapter de manière satisfaisante à la nature locale du flux. Comment le visualiser, et interagir avec lui, sachant qu'à l'instar du flux de données, il évoluera également dans le temps ? Ces problèmes se rapprochent également du domaine du *storytelling*, pour lequel il existe un compromis à réaliser entre la liberté d'interaction et une restitution efficace de l'information [7].

## BIBLIOGRAPHIE

1. Berndt, D., and Clifford, J. Using dynamic time warping to find patterns in time series. *AIII Workshop on KDD* (1994), 359–370.

2. Bondu, A., and Boullé, M. A supervised approach for change detection in data streams. *International Joint Conference on Neural Networks* (2011), 519–526.
3. Durbin, J., and Koopman, S. J. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.
4. Gardner, G., Harvey, A. C., and Phillips, G. D. A. Algorithm AS154. An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. *Applied Statistics* (1980), 311–322.
5. Heer, J., Kong, N., and Agrawala, M. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2009), 1303–1312.
6. James, G. Variance and bias for general loss functions. *Machine Learning* 51 (2003), 115–135.
7. Kosara, R., and Mckinlay, J. Storytelling: The next step for visualization. *IEEE Computer* 46, 5 (2013), 44–50.
8. Lin, J., Vlachos, M., Keogh, E., and Gunopulos, D. Iterative incremental clustering of time series. *Advances in Database Technology (EDBT 2004). LNCS 2992* (2004), 106–122.
9. McQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (1967), 281–297.
10. Sato, M. Online model selection based on the variational bayes. *Neural Computation* 13, 7 (Jul 2001), 1649–1681.
11. Smidl, V., and Quinn, A. *The variational Bayes method in signal processing*. Springer, 2006.
12. Toyoda, M., Sakurai, Y., and Ishikawa, Y. Pattern discovery in data streams under the time warping distance. *The VLDB Journal* 22, 3 (2013), 295–318.

# Visualisation de flux de données pour le grand public

Samuel Huron

IRI, INRIA

4 rue Aubry le Boucher,

75 004, Paris, France

samuel.huron@cybunk.com

## INTRODUCTION

Notre capacité à collecter, gérer et stocker des données ne cesse d'augmenter depuis des décennies. Ce phénomène impacte tous les domaines de notre société, y compris nos vies personnelles. Pour mieux comprendre et manipuler ces données la recherche fait d'incroyables avancées dans des domaines tels que les moteurs de recherches, la fouille de données, l'apprentissage par la machine, et la visualisation d'information.

Malgré les avancées considérables en visualisation d'information, peu de travaux ont été effectués dans le domaine de la visualisation temps réel d'information. Néanmoins ces techniques peuvent s'avérer utiles dans de nombreux scénarios tels que le monitoring pour l'analyse des flux de presse ou de bourse, la sécurité intérieure, ou encore le crowdsourcing de tâches en temps réel.

Concevoir, déployer et analyser les visualisations temps réel d'un flux de données pose de multiples problèmes de conceptions pour le designer comme de compréhension pour l'utilisateur. Mon travail de recherche a consisté à (1) développer plusieurs études de cas où ces techniques se révèlent utiles dans des situations réelles, (2) designer une nouvelle approche de mise à jour de visualisation basée sur un système token+constraint[6], (3) créer un toolkit afin d'explorer cet espace de design.

## PARTICIPANT ET EQUIPE DE RECHERCHE

Samuel Huron est designer et chercheur dans le domaine de la visualisation d'information, des interfaces homme-machine, et des nouveaux médias. Il est employé à l'Iri<sup>1</sup> en tant que responsable du design, et est doctorant en CIFRE à l'Inria dans l'équipe Aviz sous la supervision de Jean Daniel Fekete.

Avant cela, Samuel a pris part au programme de recherche de l'ENSAD<sup>2</sup>. Il a aussi obtenu un Master en Design et nouveaux médias de l'université Paris 1. Samuel a développé depuis 10 ans une expérience dans l'industrie du Web, où il a été en charge de design, du prototypage, de l'intégration et du développement pour un vaste groupe d'institution publique et privée. Enfin, dernièrement il a cofondé le hackerspace Fabelier.

Aviz<sup>3</sup> est une équipe de recherche interdisciplinaire de l'Inria qui cherche à faire évoluer l'analyse et la visualisation de larges et complexes groupes de données en étudiant et en concevant des méthodes d'analyse visuelles et interactives.

L'Iri<sup>4</sup> est un institut de recherche attaché au Centre Pompidou dont la mission est d'anticiper, accompagner, et analyser les mutations des pratiques culturelles permises par les technologies numériques, et pour contribuer parfois à les faire émerger.

## ACTIVITES DE RECHERCHE

Mon travail de recherche s'est inscrit dans une démarche de recherche action. À l'Iri je conçois et développe des prototypes qui sont déployés dans des situations réelles. À Aviz mon travail consiste à évaluer et valider ces travaux au regard des tâches qu'ils adressent.

Le travail séminal de ma thèse a été Polemic Tweet [3], un système d'annotation et d'analyse de vidéo. Au travers d'une visualisation temps réel des microposts émis par le public durant l'événement. Ce système (Fig.1a) a permis d'engager l'audience lors d'une conférence dans un processus d'annotation temps réel d'un flux vidéo [1]. Ce travail est une première étape dans l'exploitation de visualisation temps réel pour le crowdsourcing de tâches. Par la suite deux scénarios proches ont été déployés. Le premier (Bubble t) était une visualisation à destination du grand public qui permettait de voir la distribution des Tweets par candidats pendant les élections présidentielles (Fig.1b). Le second déployé sur une chaîne de télévision nationale [4] permettait de voir l'évolution d'un vote sur Twitter en temps réel (Fig.1c). Ce dernier travail a permis de mettre en avant les possibilités de mise à l'échelle de ce type de dispositif.

Le succès d'audience et la réception de ces travaux par le grand public nous ont motivés à généraliser la technique que nous avions conçue. Pour cela nous avons défini une métaphore de conception, la sédimentation visuelle (Fig.2d) [2,5] et créé une boîte à outils afin d'explorer son espace de design. Cela nous a permis d'appliquer cette technique à un ensemble de diagramme de surface (Fig.2e) et de générer de multiples nouvelles visualisations (Fig.2f).

<sup>1</sup> Institut de Recherche et d'Innovation-Centre Pompidou

<sup>2</sup> Ecole Nationale supérieure des arts décoratifs

<sup>3</sup> [www.aviz.fr](http://www.aviz.fr)

<sup>4</sup> [www.iri.centrepompidou.fr](http://www.iri.centrepompidou.fr)

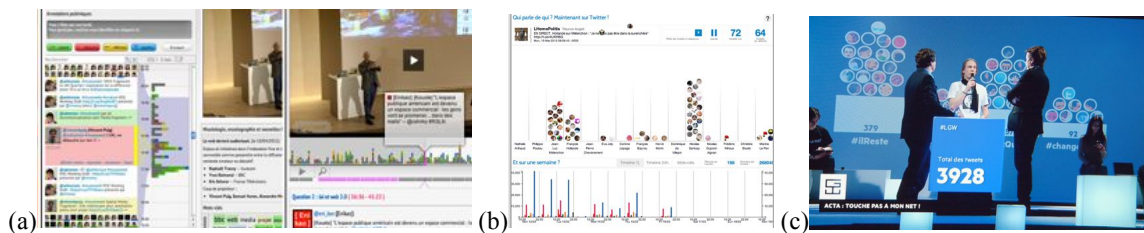


Figure 1. Différentes techniques de visualization de tweets en temps réel. (a) Polemic tree (b) Bubble t (c) BubbleTv

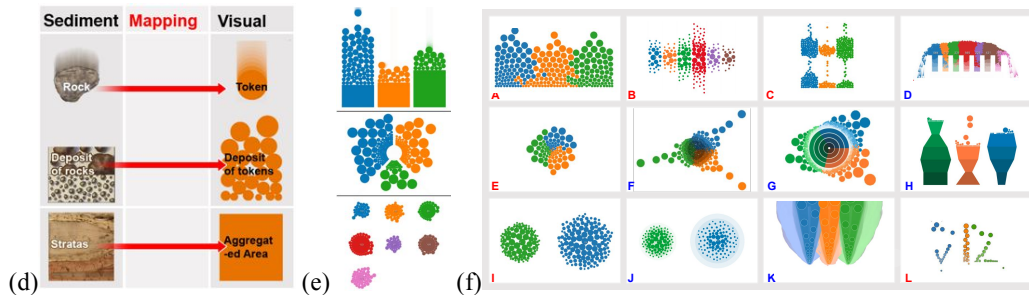


Figure 2. Visual Sédimentation, (d) présentation de l'analogie, (e) présentation de l'application a un diagramme en barre, un diagramme circulaire, (f) diagramme en barre, un diagramme circulaire.

## PROBLEMATIQUES

La visualisation de flux donné temporalisé mis à jour de manière dynamique à destination du grand public est intéressante pour de nombreuses tâches. Elle génère néanmoins de multiples problèmes, comme : la simplicité du diagramme, les propriétés dynamiques du diagramme, le niveau d'expression du diagramme.

Par simplicité j'entends, réduire la courbe d'apprentissage nécessaire à l'utilisateur pour créer un modèle mental qui lui permet de comprendre, lire et agir sur la visualisation rapidement. Cela est particulièrement important dans les cas d'usages précédemment cités où l'attention de l'utilisateur et son engagement dans la tâche dépend de la simplicité et de l'attractivité de la visualisation.

Par propriétés dynamique j'entends, concevoir un diagramme dont la mise à jour est simple à comprendre au cours du temps. Cela implique autant des problèmes liés aux changements d'échelle, que des problèmes liés au passage d'un état d'une visualisation à un autre sans action préalable de l'utilisateur.

Par niveau d'expressivité j'entends, la possibilité d'exprimer de multiples dimensions dans un diagramme. Des données tel que les Tweets ont beaucoup de dimensions différentes. Le diagramme doit pouvoir s'adapter afin d'exprimer à la demande ces différentes dimensions tout en gardant un modèle mentale cohérent.

## INTERETS

Ce groupe de travail me paraît avant tout une fabuleuse opportunité d'échange autour de la thématique de visualisation de données temporelles. J'espère aussi rencontrer des personnes avec un parcours différent du

mien qui pourront donner un éclairage nouveau à mon travail. Je suis aussi intéressé pour échanger sur différents problèmes liés à mon activité de recherche comme les méthodologies de design et d'évaluation. Cette rencontre me paraît aussi une fabuleuse occasion pour mieux connaître les équipes françaises intéressées par la visualisation de données temporelles et envisager avec elles des travaux ou des collaborations à cours ou long terme.

## BIBLIOGRAPHIE

1. Huron, S. and Fekete, J.-D. User-Supplied Sentiments in Tweets. *IEEE Visweek: 2nd Workshop on Interactive Visual Text Analysis*, (2012).
2. Huron, S., Vuillemot, R., and Fekete, J.-D. Visual Sedimentation. *IEEE Transactions on Visualization and Computer Graphics*, (2013).
3. Huron Samuel, Isenberg Petra, and Fekete Jean Daniel. PolemicTweet: Video Annotation and Analysis through Tagged Tweets. *Proceedings of the IFIP TC13 Conference on Human-Computer Interaction (INTERACT)* (2013), (2013).
4. Huron Samuel, Vuillemot Romain, and Fekete Jean Daniel. Bubble-TV: Live Visual Feedback for Social TV Broadcast. *CHI 2013 Workshop: Exploring and enhancing the user experience for television* (2013).
5. Huron Samuel, Vuillemot Romain, and Fekete Jean-Daniel. Towards Visual Sedimentation. *VisWeek 2012 Electronic Conference Proceedings* (2012), (2012).
6. Ullmer, B., Ishii, H., and Jacob, R.J.K. Token+constraint systems for tangible interaction with digital information. *ACM Transactions on Computer-Human Interaction* 12, 1 (2005), 81–118.

# Parcours dans les archipels musicaux

**Éric Languénou**

Lina/Université de Nantes  
2, rue de la Houssinière, BP 92208  
44322 Nantes Cedex 03, France  
eric.languenou@univ-nantes.fr

**Pascale Kuntz**

Lina/université de Nantes  
2, rue de la Houssinière, BP 92208  
44322 Nantes Cedex 03, France  
pascale.kuntz@univ-nantes.fr

## RÉSUMÉ

Cet article décrit un travail en cours sur l'analyse et la représentation des pratiques musicales actuelles.

## Mots Clés

Données musicales, carte, genres, représentation.

## ACM Classification Keywords

H.5.m. Information Interfaces and Presentation: Miscellaneous.

## INTRODUCTION

"43% des 12-25 ans citent la musique parmi leurs principaux intérêts ; 68 % en écoutent régulièrement ; 91% des 11-25 ans déclarent discuter régulièrement de la musique". Ces chiffres issus d'une enquête spécifique sur les pratiques de consommation musicale [4] ont des ordres de grandeur qui corroborent d'autres enquêtes ciblées sur les loisirs [2] ou les problématiques liées aux usages du numérique. Ils rappellent l'importance de la musique dans notre vie quotidienne [1] et en particulier celle des jeunes. Selon S. Frith [3], spécialiste de l'expérience musicale, "la musique est la bande son de notre vie quotidienne" qui impacte consciemment ou inconsciemment nos expériences sociales et émotionnelles. L'enquête précitée précise également que "83% des 12-25 ans écoutent de la musique via leur matériel informatique". Et la dernière enquête Calysto (2012) indique ce même chiffre pour les téléchargements musicaux chez le jeune public des 11-13 ans. Ils interrogent donc sur l'impact du virage du numérique et de la dématérialisation de la musique sur les pratiques quotidiennes. Certains parlent de "numérimorphose" en référence à la "discomorphose" [5].

Etrangement, il existe à notre connaissance peu de travaux sur la musicalisation du quotidien à l'ère du numérique eu égard à son importance. Bon nombre de recherches académiques sur la consommation musicale commencent à dater et n'intègrent donc pas à son échelle réelle le passage au numérique ou bien elles produisent des statistiques générales qui ne permettent guère de comprendre les évolutions sous-jacentes des pratiques sociales. Cependant, les enjeux technologiques associés suscitent

un regain d'intérêt notable pour l'analyse des comportements [6, 7]. Ces travaux permettent d'appréhender plus finement l'impact du virage numérique. Cependant, ils rendent peu compte des pratiques dans leur temporalité. Dans notre projet (ANR Musimorphose), l'objectif est d'analyser des parcours individuels d'écoute sur des supports variés (fixes et mobiles) sur plusieurs mois.

## VISUALISATION DE TRACES COMPORTEMENTALES

De nombreux travaux associant des approches des champs disciplinaires complémentaires (fouille de données, ingénierie des connaissances, IHM) ont été consacrés ces dernières années à l'analyse de traces d'usage sur la Toile (*web usage mining*). Sur la Toile, ces analyses s'opèrent à différentes échelles : en particulier, à l'échelle d'un site où il s'agit d'analyser les parcours de pages et d'utilisation des ressources ou à l'échelle de la Toile lorsqu'il s'agit d'analyser des parcours entre sites (e.g. Liu, 2011[9], Velingiri *et al*, 2011 [16] pour des récents états de l'art). Quelle que soit l'échelle, la plupart des approches de traitement sont basées sur des automates ou des algorithmes de fouille de règles d'association et de motifs temporels qui permettent d'extraire automatiquement des séquences caractéristiques et des relations entre ces séquences qui doivent être ensuite interprétées par les experts. En visualisation, des propositions très variées ont été proposées ces dernières années pour représenter des traces d'activités et mettre en évidence des changements temporels (e.g. *Annual Visual Analytics Science and Technology Contest*). Et de façon générale, le verrou majeur abordé par ces recherches est celui de la volumétrie des traces à traiter et à représenter.

Dans ce projet, la perspective est différente. Le recueil des données ne portera que sur un nombre restreint d'individus volontaires (quelques dizaines). Mais, les informations recueillies pour chacun seront fortement hétérogènes, et la fenêtre temporelle du recueil sera longue (plusieurs mois). Le verrou majeur est donc ici la complexité des traces. Le défi à relever est la prise en compte dans l'analyse et les restitutions visuelles de la multiplicité des points de vue et des échelles temporelles.

## TRAVAIL EN COURS

Nos premiers travaux s'orientent vers la représentation des traces de parcours d'écoute sur une restitution cartographique bi-dimensionnelle des "genres musicaux". Ce choix a été guidé par une question posée par les sociologues impliqués dans le projet : les pratiques individuelles sont-elles restreintes à quelques sous-classes ou bien y a-t-il errance ? La métaphore de la carte géographique [11, 13] où les zones géographiques sont ici associées



à des genres musicaux semble appropriée pour fournir une visualisation synthétique. Elle sous-tend l'existence d'une typologie en genres. Malheureusement, si de très nombreuses propositions ont été faites dans la littérature [8, 10, 17, 14], il n'existe pas à notre connaissance de consensus. Dans notre travail nous expérimentons différentes classifications. Leurs représentations visuelles peuvent être abordées par des démarches différentes (e.g. cartes auto-organisatrices, graphe de similarités, *multidimensional scaling*) qui toutes nécessitent la construction d'une mesure de similarité entre genres. Pour affiner la compréhension et contourner l'absence de consensus sur la définition des genres, nous pouvons associer à chaque genre une carte 2D des artistes associés [12]. Les centres d'intérêts musicaux [15], pour une personne donnée, peuvent alors être représentés par des pics, localisés sur les emplacements 2D des genres et dont la hauteur est une fonction du nombre de morceaux pour ce genre dans la bibliothèque musicale. L'objectif est ensuite de représenter sur cette carte les traces d'écoutes annotées de données temporelles. Les changements de ces traces au cours du temps (p.ex. d'un jour à l'autre) pourront se mesurer et donner des indications quantitatives sur l'évolution de l'écoute et l'exploration de nouveaux genres.

## BIBLIOGRAPHIE

1. DeNora, T. *Music in Everyday Life*. Cambridge University Press, 2000.
2. Donnat, O. *Les Pratiques culturelles des Français à l'ère numérique*. La Découverte, Ministère de la Culture et de la Communication, 2009.
3. Frith, S. *Performing Rites: Evaluating Popular Music*. Oxford University Press, 1996.
4. Granjon, F., and Combes, C. La numérimorphose des pratiques de consommation musicale - le cas de jeunes amateurs. *Réseaux* 25, 145-146 (2007), 291-334.
5. Hennion, A. *Les professionnels du disque. Une sociologie des variétés*. Paris, Métailié, 1981.
6. Kamalzadeh, M., Baur, D., and Moller, T. A survey on music listening and management behaviors. In *(ISMIR) 13th Int. Society for Music Information Retrieval Conference* (2012), 373-378.
7. Leong, T., and Wright, P. Revisiting social practices surrounding music. In *CHI'13*, ACM Press (2013), 951-960.
8. Li, J., Shao, B., Li, T., and Ogihara, M. Hierarchical co-clustering: A new way to organize the music data. *IEEE Transactions on Multimedia* 14, 2 (2012), 471-481.
9. Liu, B. *Web Data Mining, second edition*. Springer, 2011.
10. Lukashevich, H., Abeßer, J., Dittmar, C., and Grossmann, H. From multi-labeling to multi-domain-labeling: A novel two-dimensional approach to music genre classification. In *Proceedings of the 10th International Society for Music Information Retrieval Conference* (Kobe, Japan, October 26-30 2009), 459-464.
11. Pampalk, E. Islands of music: Analysis, organization, and visualization of music archives. Master's thesis, Vienna University of Technology, Austria, December 2001.
12. Rauber, A., Pampalk, E., and Merkl, D. Content-based music indexing and organization. In *SIGIR*, ACM (2002), 409-410.
13. Rauber, A., Pampalk, E., and Merkl, D. The som-enhanced jukebox: Organization and visualization of music collections based on perceptual models. *Journal of New Music Research* 32, 2 (2003), 193-210.
14. Reed, J., and Lee, C. A study on music genre classification based on universal acoustic models. In *Proceedings of the 7th International Conference on Music Information Retrieval* (Victoria (BC), Canada, October 8-12 2006).
15. Torrens, M., Hertzog, P., and Arcos, J. Visualizing and exploring personal music libraries. In *ISMIR* (2004).
16. Vellingiri, J., and Pandian, S. A survey on web usage mining. *Journal of Computer Science and Technology* 11, 4 (2011), 67-72.
17. Wang, F., Wang, X., Shao, B., Li, T., and Ogihara, M. Tag integrated multi-label music style classification with hypergraph. In *Proceedings of the 10th International Society for Music Information Retrieval Conference* (Kobe, Japan, October 26-30 2009), 363-368.



# GT1 - Visualisation d'Information, Interaction, et Fouille de Données

Charles Perin

INRIA, Univ. Paris-Sud, CNRS-LIMSI  
Bat 650, Univ. Paris-Sud  
91405, Orsay, France  
charles.perin@inria.fr

## ACTIVITÉS DE RECHERCHE

Je suis doctorant en dernière année dans l'équipe INRIA de Visualisation Aviz<sup>1</sup>, encadré par Jean-Daniel Fekete. L'équipe se compose de deux directeurs de recherche, Jean-Daniel Fekete (chef d'équipe) et Tobias Isenberg, deux chargés de recherche, Pierre Dragicevic et Petra Isenberg, trois post-doc, sept doctorants, un nombre variable de stagiaires, et plusieurs collaborateurs externes. Je suis co-encadré par Frédéric Vernier, au CNRS-LIMSI.

Les différents axes de recherche de l'équipe sont:

1. les nouvelles représentations visuelles. Par exemple des techniques de rendu non photo-réaliste [5] et des visualisations tangibles/physiques [7].
2. les nouvelles interactions. Par exemple des représentations interactives hybrides pour les réseaux [1, 4] et des contrôleurs tangibles à distance pour des murs d'écrans [6].
3. l'infrastructure logicielle. Par exemple Evo-GraphDice [3], un outil de visualisation basé sur les algorithmes génétiques.
4. les méthodes d'évaluation. Par exemple les méthodes d'évaluation par CrowdSourcing sur Amazon Mechanical Turk [2, 8].

Je travaille sur la visualisation de données multidimensionnelles et temporelles. J'ai présenté/publié des travaux sur la visualisation compacte et l'interaction avec des séries temporelles [10, 11].

Je présente aussi un article à IHM'13 sur l'interaction collaborative avec des données multidimensionnelles et temporelles, dans le contexte de la création de contenus cinématographiques [9].

J'ai travaillé récemment—et continue aujourd'hui encore—avec des données sportives, de football plus précisément [12]. Ces données, hautement multidimensionnelles et intrinsèquement temporelles offrent un domaine d'application nouveau, avec un large public. Je continue dans cette voie en travaillant actuellement sur les tableaux de classement temporels pour CHI'14. Je présente aussi à un workshop à VIS'13 en Octobre sur la collecte de données temporelles en temps-réel par des spectateurs de rencontres sportives et les problématiques associées.

## PROBLÉMATIQUES

Le problème principal lié aux données temporelles concerne l'échelle à laquelle les données doivent être visualisées. Selon le contexte, l'utilisateur final et les tâches qui doivent être accomplies, à la fois la granularité de la dimension temporelle et le niveau de détail des visualisations et interactions est à déterminer ; il n'existe pas de règle générale pour une présentation efficace.

La dimension temporelle est assurément une dimension qui doit être traitée différemment des autres dimensions. Lorsque des données ont une composante temporelle, le temps est la dimension principale. Le défi à relever est alors de représenter des données multidimensionnelles sur un axe temporel. Cette problématique est le sujet de nombreux travaux et reste largement ouverte. Par exemple, la Figure 1 illustre le dernier projet pour lequel j'ai travaillé, où nous proposons des nouvelles techniques d'interaction pour naviguer dans la dimension temporelle de données multidimensionnelles représentées par des tableaux de classement.

Un problème plus général du domaine de la visualisation concerne les connaissances de l'utilisateur final en termes de lecture de visualisations. J'ai pu, au cours de différents projets, réaliser la disparité qui existe à ce niveau et la difficulté à proposer des visualisations et interactions nouvelles à des utilisateurs non instruits.

La figure 2 illustre un projet [12] où nous avons été confrontés à ces trois problématiques. L'échelle à laquelle visualiser les données a fait l'objet de longues recherches, d'interviews avec des journalistes et analystes. La conception de l'interface finale a été itérative, le niveau de détail à présenter étant le majeur problème ; notre réponse apportée étant finalement la *phase* de jeu. Plusieurs prototypes ont été successivement rejetés et ce projet aura duré plus de deux ans au total. La communication avec le public—ou l'utilisateur—est une étape à la fois difficile et enrichissante. Nous sommes parvenus à proposer des visualisations qu'il est capable de comprendre et de s'approprier et surtout d'utiliser dans un but précis afin d'améliorer son exploration, ses analyses, et la communication de ces analyses.

La dernière problématique à laquelle je suis confronté est l'analyse de données temporelles massive, un problème à la fois d'actualité et complexe. L'un des derniers projets de ma thèse aborde la visualisation de quantités de données trop grandes pour être stockées en mémoire.

<sup>1</sup><http://www.aviz.fr>

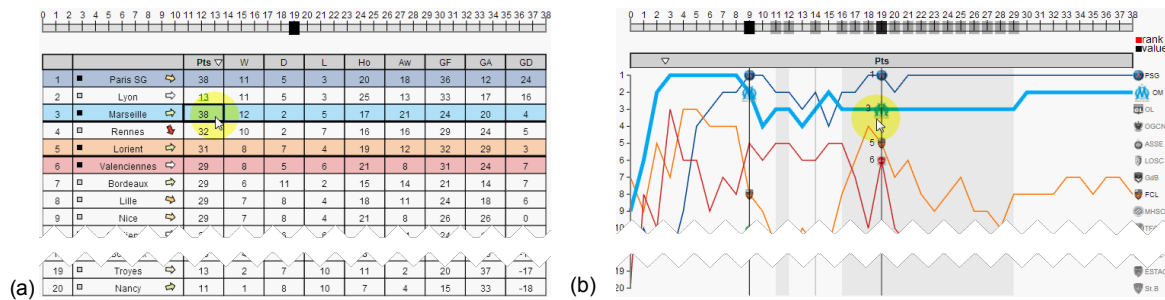


Figure 1. Nouvelles techniques de navigation dans les tableaux de classement pour des données multidimensionnelles et temporelles.

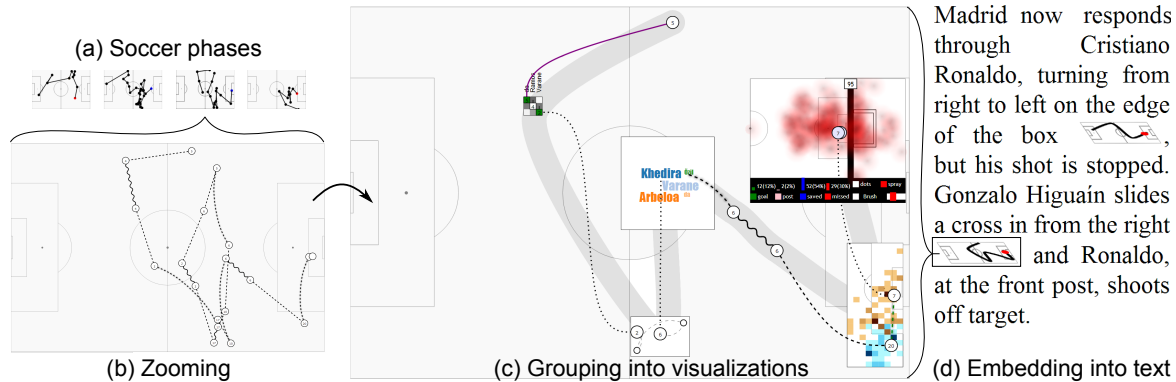


Figure 2. Interface de SoccerStories: (a) navigation dans des phases de jeu ; (b) projection d'une phase sur un terrain de football ; (c) exploration de la phase en regroupant les actions dans des visualisations dédiées et adaptées ; et (d) communication des phases en utilisant des SportLines à inclure dans du texte.

Les problèmes soulevés par l'utilisation de données temporelles en masse trop importante pour être visualisées avec les outils traditionnels sont nombreux et les pistes à explorer le sont plus encore. Par exemple, comment représenter en temps réel un aperçu des données ? Quelles règles d'aggrégation sont pertinentes pour telle ou telle donnée ? Quelles ancres—visuelles ou mentales—offrir à l'utilisateur afin de lui permettre de naviguer efficacement dans des grandes quantités de données ? Faut-il exploiter par exemple la mémoire spatiale ? visuelle ? épisodique ?

## BIBLIOGRAPHIE

1. Bezerianos, A., Dragicevic, P., Fekete, J.-D., Bae, J., and Watson, B. Geneaquilts: A system for exploring large genealogies. *TVCG* 16, 6 (2010), 1073–1081.
2. Boukhelifa, N., Bezerianos, A., Isenberg, T., and Fekete, J. Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *TVCG* 18, 12 (2012), 2769–2778.
3. Cancino Ticona, W. G., Boukhelifa, N., and Lutton, E. EvoGraphDice: Interactive Evolution for Visual Analytics. In *Proc. CEC*, IEEE (2012), 2286–2293.
4. Henry, N., Fekete, J.-D., and McGuffin, M. J. Nodetrix: a hybrid visualization of social networks. *TVCG* 13, 6 (2007), 1302–1309.
5. Isenberg, T. Visual Abstraction and Stylisation of Maps. *The Cartographic Journal* 50, 1 (2013), 8–18.
6. Jansen, Y., Dragicevic, P., and Fekete, J.-D. Tangible remote controllers for wall-size displays. In *CHI'12*, ACM (2012), 2865–2874.
7. Jansen, Y., Dragicevic, P., and Fekete, J.-D. Evaluating the efficiency of physical visualizations. In *CHI'13*, ACM (2013), 2593–2602.
8. Micallef, L., Dragicevic, P., and Fekete, J. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *TVCG* 18, 12 (2012), 2536–2545.
9. Perin, C., Christie, M., Vernier, F., and Lino, C. Collastar : Interaction collaborative avec des données multidimensionnelles et temporelles. À paraître à IHM 2013.
10. Perin, C., Vernier, F., and Fekete, J.-D. Progressive Horizon Graphs: Improving Small Multiples Visualization of Time Series. In *VisWeek 2012 Electronic Conference Proceedings*, IEEE, Ed., IEEE (2012).
11. Perin, C., Vernier, F., and Fekete, J.-D. Interactive horizon graphs: improving the compact visualization of multiple time series. In *CHI'13*, ACM (2013), 3217–3226.
12. Perin, C., Vuillemot, R., and Fekete, J.-D. SoccerStories: A Kick-off for Visual Soccer Analysis. *TVCG* (2013).

# Analyse de graphes dynamiques issus d'un processus de réécriture de graphes

Bruno Pinaud

Université Bordeaux 1, LaBRI  
Bât A30, 351 Cours de la Libération  
33405 Talence Cedex, France  
bruno.pinaud@labri.fr

## INTRODUCTION

Je suis membre de l'équipe Modèles et Algorithmes pour la Bioinformatique et la Visualisation (MaBio-Vis) du Laboratoire Bordelais de Recherche en Informatique (LaBRI) et plus précisément du thème "Visualisation de grandes masses de données". La vision scientifique de l'équipe s'inspire de la formidable collaboration qui se met en place aujourd'hui entre les sciences de l'information, les sciences de la vie, les sciences sociales et les sciences économiques. La combinaison de techniques expérimentales à haut-débit et des méthodes computationnelles avancées permet de traiter des problèmes de taille sans précédent, et a donné naissance à des questions émergentes, par exemple en biologie ou en géographie quantitative. Dans les différents domaines que nous étudions, les données peuvent se modéliser par des systèmes complexes, eux mêmes définis par un ensemble constitué d'un grand nombre d'entités en interaction.

Je m'intéresse plus particulièrement à la modélisation, la simulation et l'analyse de systèmes complexes obtenus par de la réécriture de graphes. Imaginez un jeu dans lequel des règles de transformation sont successivement appliquées sur un graphe jusqu'à atteindre une condition d'arrêt. Une règle décrit un motif local (un sous-graphe) qui doit être identifié dans le graphe et comment transformer ce motif. Le formalisme de la réécriture de graphes est à la fois très riche et complexe rendant l'étude d'un système utilisant ce formalisme difficile. Par exemple, prédire si une suite de règles est applicable dans n'importe quel ordre est bien souvent un problème difficile. Pour la modélisation de systèmes complexes, les formalismes graphiques ont des avantages certains : ils sont intuitifs et rendent plus facile le raisonnement sur le système. Nous avons donc développé la plate-forme PORGY<sup>1</sup> (Figure 1) [3]. Elle est le résultat de trois ans de travail collaboratif avec des experts des systèmes de réécritures. La méthodologie suivie est largement inspirée du travail de

Munzner sur le développement et la validation de plate-formes de visualisation [2].

Dans la suite, le terme "graphe" désigne le graphe sur lequel les règles de réécritures sont appliquées. Après avoir rapidement introduit notre plate-forme et la réécriture, je ferais le lien avec les données temporelles et les problèmes posés par le formalisme que nous utilisons.

## LES BASES DE LA RÉÉCRITURE DE GRAPHE

Chaque modification autorisée (topologie, attributs des sommets/arêtes) sur un graphe donné  $G$  est appelée une *règle de réécriture*. Elles servent à modéliser les connaissances de l'expert sur le système étudié et décrivent comment un (petit) sous-graphe de  $G$  doit être modifié ou *réécrit*. La figure 1, partie 2 est un exemple de règle. Le dessin permet de déduire immédiatement l'effet de cette règle : si un graphe contient un sous graphe avec cinq sommets dans la configuration donnée par la règle, alors la connexion entre les sommets bleu et vert est supprimée. Des modifications plus complexes sont réalisées en combinant des séquences de règles avec des opérateurs spécifiques dans une *stratégie* qui décrit qui/quand/où/comment appliquer les règles [1].

## LIEN AVEC LES DONNÉES TEMPORELLES

Porgy permet dans un premier temps de concevoir et d'éditer graphiquement des règles. Le challenge principal est alors de savoir si la règle ainsi construite est un modèle correct du système étudié. Porgy permet alors de simuler l'évolution du système après avoir défini une stratégie de réécriture (voir la figure 2). L'historique des calculs est géré par l'arbre de dérivation (figure 1, partie 4). Ses sommets sont les différents états pris par le graphe en cours de réécriture. Une arête noire indique une application de règle alors qu'une arête verte indique le point de départ et l'état final d'une stratégie. Chaque sommet de l'arbre de dérivation peut servir de point de départ pour l'application d'une nouvelle stratégie et ainsi créer une nouvelle branche.

La complexité du système de réécriture est d'une certaine façon capturée par l'arbre de dérivation, qui est donc un objet central pour l'étude d'un système de réécriture. Si on considère le graphe sur lequel les règles sont appliqués comme un graphe dynamique, l'arbre de dérivation regroupe alors l'ensemble des transformations possibles de ce graphe (une sorte de *story-board*). Les branches de

<sup>1</sup>Voir <http://tulip.labri.fr/TulipDrupal/?q=porgy>.

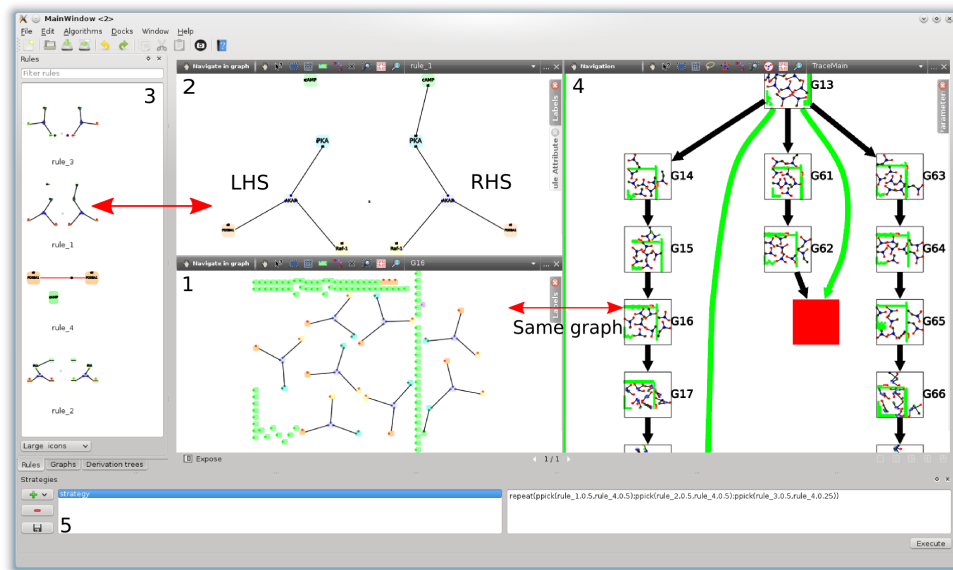


Figure 1. Vue d'ensemble de PORGY: (1) édition d'un état du graphe qui est en cours de réécriture ; (2) édition d'une règle ; (3) les règles de réécriture disponibles ; (4) une partie de l'arbre de dérivation, un historique complet des calculs effectués ; (5) l'éditeur de stratégie.

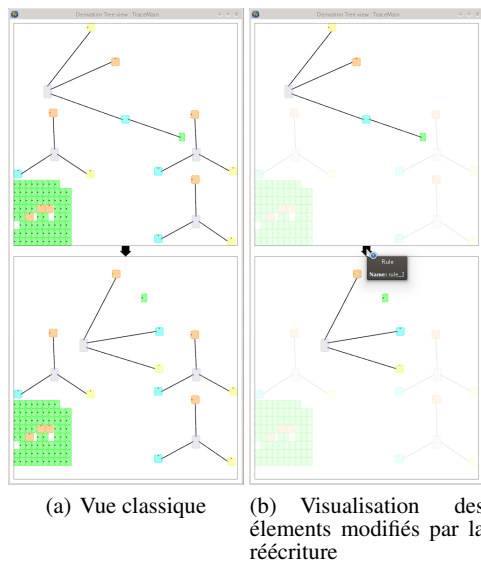


Figure 2. Une étape de réécriture (règle de la figure 1) avant et après avoir passé le pointeur de la souris sur l'arête qui relie le graphe d'origine (celui du haut) au graphe résultat (bas).

l'arbre (qui peut être infini) sont à considérer comme autant de monde parallèles rendant l'étude et la visualisation du système très difficile.

L'étude d'un système de réécriture nécessite de passer en permanence de vues locales (les règles, les graphes) à une vue globale de l'arbre de dérivation. Le principal enjeu est alors de comprendre comment le comportement global du système est dirigé par les règles qui définissent des modifications uniquement locales et une stratégie qui pilote l'application de ces règles. C'est en regardant précisément les graphes dans l'arbre de dérivation tant globalement

que localement qu'un expert du domaine peut apprécier l'adéquation du modèle avec le système.

A notre connaissance, il n'existe aucun algorithme de tracé de graphe qui produise de bons tracés quelle que soit les données utilisées. De plus, cet algorithme de dessin devrait garantir une stabilité relative du dessin entre toutes les branches de l'arbre de dérivation pour conserver au moins à minima la carte mentale de l'utilisateur. De plus, l'évolution du graphe est difficile à prédire car les modifications sont pilotées uniquement par l'application des règles.

## CONCLUSION

La combinaison à notre connaissance unique de la réécriture de graphe et de la visualisation soulève donc de nombreux problèmes. Il nous reste à traiter notamment le cas de règles de taille importante (problème isomorphisme graphe/sous-graphe) ou encore les problèmes de stabilité dans le tracé des graphes dynamiques. Nous avons surtout utilisé pour le moment des exemples basés sur la bio-informatique mais nous sommes à la recherche de nouveaux domaines d'applications (nous pensons aux sciences sociales notamment).

## BIBLIOGRAPHIE

1. Fernandez, M., Kirchner, H., and Namet, O. A strategy language for graph rewriting. In *Logic-Based Program Synthesis and Transformation*, vol. 7225 of *LNCS*, Springer (2012), 173–188.
2. Munzner, T. A nested process model for visualization design and validation. *IEEE Trans. on Visualization and Computer Graphics* 15, 6 (2009), 921–928.
3. Pinaud, B., Melançon, G., and Dubois, J. PORGY: A Visual Graph Rewriting Environment for Complex Systems. *Computer Graphics Forum* 31, 3 (2012), 1265–1274.