Vers une mesure de similarité pour les séquences complexes

Elias Egho*, Chedy Raïssi*, Toon Calders**, Thomas Bourquard*,
Nicolas Jay*, Amedeo Napoli*

*LORIA, Vandoeuvre-les-Nancy, France
prénom.nom@loria.fr
**Université Libre de Bruxelles
prénom.nom@ulb.ac.be

Résumé. Le calcul de similarité entre les séquences est d'une extrême importance dans de nombreuses approches d'explorations de données. Il existe une multitude de mesures de similarités de séquences dans la littérature. Or, la plupart de ces mesures sont conçues pour des séquences simples, dites séquences d'items. Dans ce travail, nous étudions d'un point de vue purement combinatoire le problème de similarité entre des séquences complexes (i.e., des séquences d'ensembles ou itemsets). Nous présentons de nouveaux résultats afin de compter efficacement toutes les sous-séquences communes à deux séquences. Ces résultats théoriques sont la base d'une mesure de similarité calculée efficacement grâce à une approche de programmation dynamique.

1 Introduction

Le volume de données numériques actuellement disponible nécessitent de disposer de méthodes efficaces permettant de les structurer, résumer, comparer et regrouper. Dans tous ces cas, il est indispensable de disposer d'une mesure de *similarité* permettant d'évaluer la proximité entre les objets considérés. Les illustrations les plus récentes se situent dans le domaine de la bioinformatique pour l'alignement des sous-séquences d'ADN ou d'acides aminés ((Sander et Schneider, 1991; Chothia et Gerstein, 1997)) ou dans la détection d'intrusion dans les réseaux où les différentes séquences d'accès sont analysées et comparées à une base de signatures de comportements malveillants. En ce qui concerne les données séquentielles, de nombreux travaux ((Levenshtein, 1966; Herranz et al., 2011; Keogh, 2002; Wang et Lin, 2007)) se sont intéressés à des *séquences simples*, c'est-à-dire une liste ordonnée d'éléments atomiques. Or, dès lors que l'on s'intéresse à des séquences d'objets plus complexes, le calcul de similarité se confronte à la nature même des objets comparés. Les trajectoires d'objets mobiles, les informations topologiques en biologie moléculaire ((Wodak et Janin, 2002)) sont des exemples de telles données. Pour illustration, supposons que nous souhaitons comparer les trois séquences complexes suivantes : $S_1 = \langle\{c\}\{b\}\{a,b\}\{a,c\}\rangle$, $S_2 = \langle\{b\}\{c\}\{a,b\}\{a,c\}\rangle$ et $S_3 = \langle\{b,d\}\{a,b\}\{c\}\{d\}\rangle$. Le calcul classique de la plus longue sous-séquence commune entre S_1 et S_2, noté $LCS(S_1, S_2)$, est la sous-séquence $\langle\{c\}\{a\}\{a,c\}\rangle$ de longueur 3. De même, $LCS(S_1, S_3) = \langle\{b\}\{a,b\}\{c\}\rangle$ de longueur 3. La mesure de la plus longue sous-séquence commune nous amène à conclure que la séquence S_1 peut être considérée équidis-