Vers une mesure de similarité pour les séquences complexes

Elias Egho, Chedy Raïssi, Toon Calders, Nicolas Jay, Amedeo Napoli

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-

tante des séquences S_2 et S_3. Or, la séquences S_1 est quasi identique à la séquence S_2 (hormis l'interversion des deux premiers ensembles). Il est donc important de comparer autrement l'impact des sous-séquences pour réellement mesurer la similarité de séquences d'objets complexes. Il suffit de comparer le nombre de sous-séquences communes qui est de 40 entre S_1 et S_2 alors qu'il est de 14 entre S_1 et S_3. Ce résultat reflète mieux la similarité entre S_1 et S_2 car il prend en compte les différentes structures et combinaisons présentes dans une séquence complexe. La problématique se confronte à la combinatoire associée, l'efficacité computationnelle et nous amène à poser les questions suivantes : (i) Etant donné une séquence complexe, comment compter *sans énumérer* le nombre de sous-séquences distinctes ? (ii) Pour un couple de séquences, comment *efficacement* compter le nombre de sous-séquences communes ? Dans ce contexte, notre contribution est double : un cadre théorique pour définir une *mesure de similarité pour les séquences complexes* basée sur le *nombre de sous-séquences communes* et un algorithme qui met en œuvre de façon efficace la mesure de similarité proposée. Cette approche est basé sur la technique de la programmation dynamique afin de compter efficacement toutes les sous-séquences communes entre deux séquences.

L'article est organisé de la façon suivante. La section 2 présente les définitions préliminaires à notre proposition. Les sections 3 et 4 détaillent notre contribution, présentent de nouveaux résultats combinatoires et discutent la complexité ainsi que la complétude de notre algorithme. La section 5 présente deux études expérimentales. La section 6 fait le bilan et dresse les perspectives associées à ce travail.

2 Concepts préliminaires

Définition 1 (Séquence) *Soit $\mathcal{I}$ un ensemble fini d'* items*. Un ensemble (ou* itemset*) X est un sous-ensemble non vide de $\mathcal{I}$. Une* séquence *S est une liste ordonnée $\langle X_1 \cdots X_n \rangle$ telle que X_i ($1 \leq i \leq n$, $n \in N$) est un itemset. Le l-préfixe, noté S^l, est le préfixe $\langle X_1, \ldots, X_l \rangle$ de la séquence S avec $1 \leq l \leq n$. Le j-ème itemset X_j de la séquence S est noté $S[j]$ avec $1 \leq j \leq n$. La **concaténation** de itemsets Y avec une séquence S, $S \circ Y$, est la séquence $\langle X_1 \cdots X_n\ Y \rangle$. Une séquence $T = \langle Y_1 \cdots Y_m \rangle$ est une **sous-séquence** de $S = \langle X_1 \ldots X_n \rangle$, noté $T \preceq S$, s'il existe $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ telle que $Y_j \subseteq X_{i_j}$, $\forall j = 1 \ldots m$. $\varphi(S)$ indique l'* **ensemble de toutes les sous-séquences** *d'une séquence S et $\phi(S) = |\varphi(S)|$. Pour deux séquences S et T, $\varphi(S,T)$ indique l'ensemble de **toutes les sous-séquences communes entre deux séquences** S et T : $\varphi(S,T) = \varphi(S) \cap \varphi(T)$ et $\phi(S,T) = |\varphi(S,T)|$.*

Définition 2 *La **similarité entre deux séquences** S et T, notée $sim(S, T)$, est définie comme le nombre de toutes les sous-séquences communes entre S et T divisé par le nombre maximal de sous-séquences de S ou T : $sim(S,T) = \frac{\phi(S,T)}{\max\{\phi(S),\phi(T)\}}$.*

De manière usuelle, l'ensemble des parties d'un ensemble X est noté $\mathcal{P}(X)$ et $\mathcal{P}_{\geq 1}(X)$ est l'ensemble des parties de X sans l'ensemble vide (i.e., $\mathcal{P}_{\geq 1}(X) = \mathcal{P}(X) \setminus \{\emptyset\}$).

Exemple 1 *Nous utilisons $\mathcal{D}_{ex}$ dans le tableau 1 afin d'illustrer ces définitions. La base contient 4 séquences construites à partir de $\mathcal{I} = \{a, b, c, d, e\}$. $\langle \{a\}\{b\}\{c, d\} \rangle$ est une sous-séquence de $S_1 = \langle \{a\}\{a, b\}\{e\}\{c, d\}\{b, d\} \rangle$. Le 3-préfixe de S_1, $S_1^3 = \langle \{a\}\{a, b\}\{e\} \rangle$ et $S_1[2] = \{a, b\}$ est le deuxième itemset dans la séquence S_1. L'ensemble de toutes les sous-séquences*

S_1	$\langle\{a\}\{a,b\}\{e\}\{c,d\}\{b,d\}\rangle$
S_2	$\langle\{a\}\{b,c,d\}\{a,d\}\rangle$
S_3	$\langle\{a\}\{b,d\}\{c\}\{a,d\}\rangle$
S_4	$\langle\{a\}\{a,b,d\}\{a,b,c\}\{b,d\}\rangle$

TAB. 1: *Le jeu de données séquentielles utilisé pour les besoins d'illustrations dans notre papier.*

de S_1^2 *est* $\{\langle\rangle, \langle\{a\}\rangle, \langle\{b\}\rangle, \langle\{a,b\}\rangle, \langle\{a\}\{a\}\rangle, \langle\{a\}\{b\}\rangle, \langle\{a\}\{a,b\}\rangle\}$ *et donc,* $\phi(S_4^2) = 15$ *de même le nombre de toutes les sous-séquences communes entre* S_1^4 *et* S_2^3 *est* $\phi(S_1^4, S_2^3) = 13$.

3 Comptage efficace de toutes les sous-séquences distinctes

Dans cette section, nous présentons un ensemble de résultats théoriques ainsi qu'une technique efficace pour calculer $\phi(S)$. Nous esquissons, dans un premier temps, l'intuition qui a permis de développer notre approche. Supposons que nous concaténons la séquence S avec un itemset Y et observons les valeurs de $\phi(S)$ et $\phi(S \circ Y)$. Deux cas peuvent apparaître : (i) Soit Y est disjoint avec tous les itemsets de S et dans ce cas, le calcul de $\phi(S \circ Y)$ est équivalent a $|\varphi(S)| \cdot 2^{|Y|}$ car $\forall T \in \varphi(S)$ et $\forall Y_1, Y_2 \in \mathcal{P}(Y)$, $T \circ Y_1 \neq T \circ Y_2$. Par exemple, $\phi(\langle\{a,b\}\{c\}\rangle \circ \{d,e\}) = 8 \cdot 2^2 = 32$. (ii) Au moins un item de Y apparaît dans un des itemsets de S (i.e., $\exists i \in [1,n] : Y \cap X_i \neq \emptyset$). Dans ce cas, $\phi(S \circ Y) < \phi(S) \cdot 2^{|Y|}$, car l'extension des séquences de $\varphi(S)$ avec les éléments de l'ensemble des parties de Y produira nécessairement des sous-séquences doublons. Afin de calculer correctement $\phi(S \circ Y)$, nous avons donc besoin de définir une méthode pour *supprimer les répétitions au niveau du comptage*. De manière plus formelle, $\phi(S \circ Y) = \phi(S) \cdot 2^{|Y|} - R(S,Y)$ où $R(S,Y)$ représente le terme de correction.

L'idée générale que nous développons s'appuie sur la possibilité de compenser les répétitions de comptages dû à la concaténation de sous-séquences de S avec l'ensemble des parties de l'itemset Y. Afin de détecter où des répétitions de comptages se produisent dans une séquence S donnée, nous introduisons ci-dessous le concept d'ensemble de positions critiques.

Définition 3 (Ensemble de positions critiques) *Pour une séquences* $S = \langle X_1 \cdots X_n \rangle$ *et un itemset* Y*, l'ensemble de positions critiques, noté* $L(S,Y)$ *regroupe toutes les positions d'itemsets de* S *où l'intersection avec* Y *est maximale* [1]*. Plus formellement,*

$$L(S,Y) = \{i \mid (Y \cap X_i \neq \emptyset) \ \wedge \ (\nexists\, j \ , \ j > i \textit{ tel que } \ Y \cap X_i \subseteq Y \cap X_j)\}$$

Cette notion d'ensemble de positions critiques est cruciale dans notre approche puisqu'elle permet de focaliser les calculs uniquement sur les dernières positions où une répétition apparaît pour une séquence S donnée. Le lemme suivant formalise cette intuition.

Lemme 1 *Soient* S *une séquence et* Y *un itemset. Alors,*

$$R(S,Y) \ = \ \left| \bigcup_{\ell \in L(S,Y)} \left\{ \varphi(S^{\ell-1}) \circ \mathcal{P}_{\geq 1}(S[\ell] \cap Y) \right\} \right|$$

1. Selon la relation d'inclusion.

Preuve 1 *Voir le rapport technique Egho et al. (2012).*

Remarque. Les éléments de l'ensemble $\varphi(S^{\ell-1}) \circ \mathcal{P}_{\geq 1}(S[\ell] \cap Y)$ ne sont pas nécessairement disjoints. Pour s'en convaincre, considérerons la séquence $S = \langle\{a,b\}\{b,c\}\rangle$ et $Y = \{a,b,c\}$, alors $L(S,Y) = \{1,2\}$, et la séquence $\langle\{b\}\rangle$ est construite deux fois dans les ensembles $\varphi(S^0) \circ \mathcal{P}_{\geq 1}(S[1] \cap Y)$ et $\varphi(S^1) \circ \mathcal{P}_{\geq 1}(S[2] \cap Y)$.

Afin de prendre en compte ce chevauchement d'éléments, nous nous appuyons pour le calcul de $R(S,Y)$ sur le principe d'inclusion-exclusion.

Théorème 1 *Soient $S = \langle X_1 ... X_n\rangle$ une séquence et Y un itemset. Alors,*

$$\phi(S \circ Y) = 2^{|Y|} \cdot \phi(S) - \sum_{K \subseteq L(S,Y)} (-1)^{|K|+1} \left(\phi(S^{(\min K)-1}) \cdot \left(2^{|(\bigcap_{j \in K} X_j) \cap Y|} - 1\right)\right)$$

Preuve 2 *Voir le rapport technique Egho et al. (2012).*

Exemple 2 *Nous illustrons le processus complet de comptage $\phi(\langle\{a\}\{a,b,d\}\rangle)$.*

$\phi(\langle\rangle) = 1$, $\phi(\langle\{a\}\rangle) = 2^{|\{a\}|} \cdot \phi(\langle\emptyset\rangle) = 2$, $\phi(\langle\{a\}\{a,b,d\}\rangle) = 2^{|\{a,b,d\}|}\phi(\langle\{a\}\rangle) - (2^{|\{a,b,d\} \cap \{a\}|} - 1) \cdot \phi(\langle\emptyset\rangle) = 2^3 \cdot 2 - (2^1 - 1) \cdot 1 = 15$

4 Comptage efficace de toutes les sous-séquences communes

Dans cette section, nous allons étendre les résultats théoriques précédemment énoncés afin de compter toutes les sous-séquences distinctes communes entre deux séquences S et T. Comme pour la Section 3, nous présentons dans un premier temps, l'idée générale avant d'énoncer les résultats formels ainsi que l'algorithme de comptage. Supposons que nous concaténons la séquence S avec un itemset Y et observons la variation des ensembles $\varphi(S,T)$ et $\varphi(S \circ Y, T)$. Deux cas sont possibles : Si aucun item dans Y n'apparaît dans les itemsets des séquences S et T, alors la concaténation de l'itemset Y avec la séquence S n'a aucun effet sur l'ensemble $\varphi(S \circ Y, T)$ (i.e., le nombre de sous-séquences communes ne change pas, $\varphi(S \circ Y, T) = \varphi(S,T)$). Ou si au moins un élément de Y apparaît dans l'une des séquences S ou T (ou les deux), alors de nouvelles séquences communes apparaissent dans $\varphi(S \circ Y, T)$. De la même manière que pour la méthode de comptage des sous-séquences distinctes d'une séquence unique, des répétitions peuvent se produire et il est nécessaire de définir un terme de correction. De manière formelle,

$$|\varphi(S \circ Y, T)| = |\varphi(S,T)| + A(S,T,Y) - R(S,T,Y)$$

et $A(S,T,Y)$ représente le nombre de sous-séquences communes supplémentaires qui devraient être ajoutées au compte après la concaténation de Y et $R(S,T,Y)$ le terme de correction.

De même que pour le problème des sous-séquences distinctes pour une seule séquence, l'ensemble des positions critique joue un rôle dans le calcul de $A(S,T,Y)$ et $R(S,T,Y)$. Le lemme suivant formalise cette observation :

Lemme 2 *Soient* $S = \langle X_1 ... X_n \rangle$, $T = \langle X'_1 ... X'_m \rangle$ *des séquences et* Y *un itemset, alors* $A(S,T,Y) = \left| \bigcup_{\ell \in L(T,Y)} \left\{ \varphi(S, T^{\ell-1}) \circ \mathcal{P}_{\geq 1}(T[\ell] \cap Y) \right\} \right|$ *et*

$$R(S,T,Y) = \left| \bigcup_{\ell \in L(S,Y)} \left\{ \bigcup_{\ell' \in L(T,Y)} \left\{ \varphi(S^{\ell-1}, T^{\ell'-1}) \circ \mathcal{P}_{\geq 1}(S[\ell] \cap T[\ell'] \cap Y) \right\} \right\} \right|$$

Preuve 3 *Voir le rapport technique Egho et al. (2012).*

Comme pour le Lemme 1, le calcul de $A(S,T,Y)$ et $R(S,T,Y)$ implique l'utilisation du principe d'inclusion-exclusion.

Théorème 2 *Soient* $S = \langle X_1 ... X_n \rangle$, $T = \langle X'_1 ... X'_m \rangle$ *des séquences et* Y *un itemset. Alors* $\phi(S \circ Y, T) = \phi(S,T) + A(S,T,Y) - R(S,T,Y)$ *avec*

$$A(S,T,Y) = \sum_{K \subseteq L(T,Y)} (-1)^{|K|+1} \left(\phi(S, T^{\min(K)-1}) \cdot \left(2^{\left| \left(\bigcap_{j \in K} X'_j \right) \cap Y \right|} - 1 \right) \right)$$

et

$$R(S,T,Y) = \sum_{K \subseteq L(S,Y)} (-1)^{|K|+1} \left(\sum_{K' \subseteq L(T,Y)} (-1)^{|K'|+1} \cdot f(K,K') \right)$$

tel que

$$f(K,K') = \phi(S^{(\min K)-1}, T^{(\min K')-1}) \cdot \left(2^{\left| \left(\bigcap_{j \in K} X_j \right) \cap \left(\bigcap_{j' \in K'} X'_{j'} \right) \cap Y \right|} - 1 \right)$$

Preuve 4 *Voir le rapport technique Egho et al. (2012).*

Le théorème 2 permet de concevoir un algorithme simple qui s'appuie sur la technique de programmation dynamique. Pour deux séquences données S et T, de tailles n et m respectivement, l'algorithme produit une $n \times m$-matrice, notée $\mathcal{M}$, tel que la valeur de la cellule $\mathcal{M}_{i,j}$ correspond au nombre de sous-séquences communes entre S^i et T^j (i.e., $\mathcal{M}_{i,j} = \phi(S^i, T^j)$). Considérons les deux séquences S_1 et S_2 dans $\mathcal{D}_{ex}$, alors $\phi(S_1, S_2) = 21$. Le

	$\{\emptyset\}$	$\{a\}$	$\{b,c,d\}$	$\{a,d\}$
$\{\emptyset\}$	1	1	1	1
$\{a\}$	1	2	2	2
$\{a,b\}$	1	2	4	5
$\{e\}$	1	2	4	5
$\{c,d\}$	1	2	10	13
$\{b,d\}$	1	2	12	21

TAB. 2: *Trace matricielle du comptage des sous-séquences communes entre* S_1 *et* S_2.

5 Expériences

Nous invitons le lecteur à lire la section des expérimentations dans le rapport technique Egho et al. (2012). où nous étudions le passage à l'échelle de notre mesure ainsi que son application dans le domaine de regroupement de séquences de protéines dans le domaine biologique.

6 Conclusion

Dans cet article, nous étudions le problème de comptage de toutes les séquences communes entre deux séquences complexes. Nous présentons un cadre théorique et un algorithme de programmation dynamique efficace pour compter efficacement le nombre de sous-séquences communes entre deux séquences. Cette solution nous permet de définir une mesure de similarité entre deux séquences S et T d'une manière simple et intuitive. Un travail théorique en cours consiste à approximer le nombre des sous-séquences communes afin accélérer les calculs sur des séquences extrêmement longues.

Références

Chothia, C. et M. Gerstein (1997). Protein evolution. how far can sequences diverge ? *Nature 6617*(385), 579–581.

Egho, E., C. Raïssi, T. Calders, T. Bourquard, N. Jay, et A. Napoli (2012). On Measuring Similarity for Sequences of Itemsets. Research Report RR-8086, INRIA.

Herranz, J., J. Nin, et M. Sole (2011). Optimal symbol alignment distance : A new distance for sequences of symbols. *IEEE Transactions on Knowledge and Data Engineering 23*, 1541–1554.

Keogh, E. (2002). Exact indexing of dynamic time warping. In *Proceedings of the 28th international conference on Very Large Data Bases*, VLDB '02, pp. 406–417. VLDB Endowment.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*(8), 707–710.

Sander, C. et R. Schneider (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins 1*(9), 56–68.

Wang, H. et Z. Lin (2007). A novel algorithm for counting all common subsequences. In *Proceedings of the 2007 IEEE International Conference on Granular Computing*, GRC '07, Washington, DC, USA, pp. 502–. IEEE Computer Society.

Wodak, S. et J. Janin (2002). Structural basis of macromolecular recognition. *Adv Protein Chem 61*, 9–73.

Summary

Computing the similarity between sequences is a very important challenge for many different data mining tasks. There is a plethora of similarity measures for sequences in the literature, most of them being designed for sequences of items. In this work, we study the problem of measuring the similarity ratio between sequences of itemsets. We present new combinatorial results for efficiently counting distinct and common subsequences. These theoretical results are the cornerstone for an effective dynamic programming approach to deal with this problem.