Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Fonctionnalités utiles au couplage d'enregistrements

Martin Lachance[1]

Résumé

Dans le domaine du couplage d'enregistrements, il existe un large éventail de comparateurs de chaînes de caractères. Les difficultés lors des comparaisons surviennent lorsque des facteurs influent sur la composition des chaînes (ex. : emploi de surnoms pour les prénoms de personnes, erreurs typographiques, etc.). Il faut alors faire appel à des comparateurs plus sophistiqués. De tels outils permettent de réduire le nombre de liens potentiellement manqués. Malheureusement, une partie des gains peuvent s'avérer de faux liens. Afin d'améliorer les appariements, trois comparateurs de chaînes sophistiqués ont été développés et sont présentés dans cet article. Ce sont le comparateur *Lachance*, ainsi que ses dérivés, les comparateurs *multi-mots* et *multi-types*. Cette gamme d'outils est présentement disponible dans un prototype de couplage d'enregistrements déterministe, *MixMatch*. Ce logiciel permet de faire appel à des connaissances a priori afin de réduire le volume de faux liens générés lors des appariements. Un indicateur de force de liens est également proposé.

Mots Clés : Prétraitement; comparateur de chaînes; faux liens.

1. Introduction

Que ce soit en vue de réduire le fardeau des répondants aux enquêtes, réduire les coûts de collecte, etc, les données administratives sont appelées à prendre une place grandissante dans le cadre des enquêtes au cours des prochaines années. Ceci dit, le couplage d'enregistrements (probabiliste, déterministe ou autre) a vu naître de nombreuses fonctionnalités au fil des ans. Le présent article ne fait que complémenter celles-ci. De fait, les besoins identifiés et fonctionnalités présentées dans cet article sont le fruit d'une dizaine d'années d'expérience acquise soit en tant qu'utilisateur, principalement dans l'optique d'améliorer la couverture de bases de sondage, ou résultent d'une interaction constante avec d'autres utilisateurs du couplage d'enregistrements. Les cinq besoins majeurs identifiés consistent à : (1) simplifier les formats des données, (2) établir de puissants comparateurs de chaînes, (3) améliorer le prétraitement, (4) limiter les erreurs de couplage, (5) fournir des indicateurs de qualité. L'auteur présente ici une liste de suggestions pour tenter de répondre à ces besoins, ainsi que les fonctionnalités qu'il a développées, implantées et mises à l'épreuve.

2. Répondre aux besoins

2.1 Simplifier le format des données

Lorsque vient le temps de faire usage de données administratives, un des tout premiers problèmes qui survient consiste à surmonter les difficultés rencontrées avec le format des données. Ceci est particulièrement vrai lorsque de multiples sources de données sont employées pour tenter d'améliorer la couverture d'une base de sondage. En effet, il n'est pas certain que toutes les sources épouseront un format similaire pour effectuer le couplage avec la base de sondage. De plus, le logiciel employé pour le couplage d'enregistrements peut imposer des contraintes aux usagers, sur le contenu, le type et la longueur des champs. Le logiciel peut exiger des variables bien définies en entrée : prénom, nom, date de naissance, etc. Il faut alors manipuler les données pour respecter ces contraintes. Une fois ceci fait, il peut s'avérer difficile de considérer des comparaisons croisées telles qu'une inversion nom de famille et prénom versus un nom de personne complet.

[1]Martin Lachance, Statistique Canada, 100, promenade du pré Tunney, Ottawa, Ontario, Canada, K1A 0T6, Martin2.Lachance@statcan.gc.ca.

Qu'il représente une personne, un ménage ou autre chose, pouvoir considérer chaque enregistrement d'une source comme une seule variable, soit une seule chaîne de caractères, apporte beaucoup de flexibilité. L'usager peut partitionner cette chaîne en autant de « morceaux » qu'il/elle le désire et ensuite les combiner à sa guise afin d'identifier des appariements potentiels. Le tableau 2.1-1 illustre un exemple de partitionnement. Le partitionnement des enregistrements présentés permet la comparaison entre le mois et le jour de l'année, en cas d'inversion dans les données, de même que la gestion des inversions entre nom et prénom. Les comparaisons sont possibles grâce au partitionnement et à un simple opérateur de concaténation '+'.

Tableau 2.1-1
Comparaisons de chaînes de caractères uniques partitionnées

<table>
<tr><th colspan="3"></th><th colspan="5">Fichier 1</th><th colspan="4">Fichier 2</th></tr>
<tr><td colspan="3">Enregistrements</td><td colspan="5">Martin Lachance 2014 - 10 - 30</td><td colspan="4">Lachance Martin 2014 / 30 / 10</td></tr>
<tr><td colspan="3">Partitionnements</td><td>prénom1</td><td>nom1</td><td>An1</td><td>M1</td><td>J1</td><td>nom2</td><td>An2</td><td>M2</td><td>J2</td></tr>
<tr><td rowspan="4">Comparaisons successives entre les fichiers</td><td>1</td><td colspan="5">(prénom1+nom1) et (An1+M1+J1)</td><td rowspan="4" colspan="4">(nom2) et (An2+M2+J2)</td></tr>
<tr><td>2</td><td colspan="5">(nom1+prénom1) et (An1+M1+J1)</td></tr>
<tr><td>3</td><td colspan="5">(prénom1+nom1) et (An1+J1+M1)</td></tr>
<tr><td>4</td><td colspan="5">(nom1+prénom1) et (An1+J1+M1)</td></tr>
</table>

2.2 Établir de puissants comparateurs de chaînes

De nombreux comparateurs de chaînes de caractères existent pour tenter de surmonter les erreurs phonétiques, typographiques, etc. La plupart considèrent chacune des chaînes de caractères à comparer comme une seule séquence de caractères, i.e. un seul mot une fois les espacements entre les mots ignorés. Par exemple, « Martin Lachance » serait transformé en un seul mot « MARTINLACHANCE ». Ce sont les comparateurs sophistiqués *à mot unique*. C'est ainsi que nous les distinguerons des comparateurs sophistiqués *multi-mots*, capables d'effectuer des comparaisons individuelles sur chaque mot d'une chaîne de caractères. Les comparateurs *multi-mots* apportent non seulement plus de flexibilité dans les comparaisons, mais peuvent également permettre d'établir des liens plus forts qu'avec l'emploi de comparateurs à mot unique. Qui plus est, pouvoir gérer les nombres dans les chaînes de caractères (exemple : « 100 » dans « 100, Avenue Symposium ») à l'aide de comparateurs multi-mots apporte d'autant plus de flexibilité. Nous parlerons dans ce dernier cas de comparateur sophistiqué *multi-types*.

Trois comparateurs de chaînes de caractères sont présentés dans cette section. Ils font partie d'une gamme de comparateurs disponibles sous le prototype de couplage d'enregistrements déterministe *MixMatch* (Lachance, 2014). Ce sont le comparateur *Lachance* et les comparateurs multi-mots et multi-types. Les deux derniers font appel au comparateur Lachance.

2.2.1 Le comparateur Lachance

D'abord, il est important de définir ce que nous appellerons un appariement *légitime*. Un appariement légitime est un appariement non parfait, selon des critères préétablis, entre deux chaînes de caractères considérées chacune comme un mot unique. Au-delà des appariements parfaits, le comparateur Lachance permet d'identifier des appariements légitimes qui respectent des paramètres de tolérance imposés. L'algorithme associé au comparateur Lachance est basé sur un arbre de recherche. Il consiste à identifier un nombre maximum de caractères en commun entre deux mots uniques comparés, mais tout en respectant l'ordre dans lequel les caractères sont listés dans chacun des deux mots. À partir de ce nombre de caractères en commun, un degré de similitude entre les deux mots comparés est calculé à l'aide d'une formule simple. Ainsi, le degré de similitude, ou score, correspond au ratio entre deux fois le nombre de caractères en commun respectant l'ordre selon lequel les caractères sont listés dans chacun des deux mots et la somme des longueurs des deux mots. À titre d'exemple, « MARRTINLACHANCE » versus « MARTINLCHANCE » donnerait un ratio de 26/28, ou 0,93. Entre autres, le tableau 2.2.1-1 reprend quelques exemples de comparateurs existant dans la littérature (Porter et Winkler, 1997) et permet de comparer les scores obtenus avec le comparateur Lachance.

Les paramètres de tolérance accompagnant le comparateur Lachance comportent deux composantes : le nombre minimal de caractères requis pour obtenir un appariement (parfait ou légitime) et le degré de similitude exprimé par

un pourcentage. Le comparateur Lachance est employé depuis plusieurs années à Statistique Canada et un nombre minimal de quatre caractères, combiné à un degré de similitude de 85%, offrent généralement les meilleurs gains, tout en limitant le nombre de faux liens. Enfin, comme la performance de l'algorithme ralentit avec la longueur des mots comparés, et puisque les comparateurs multi-mots offrent davantage de flexibilité, il est préférable d'employer d'abord un comparateur de chaînes multi-mots lorsque la chaîne de caractères est constituée de plusieurs mots.

Tableau 2.2.1-1
Comparaisons impliquant des prénoms et noms de famille

Chaînes de caractères comparées		**Scores de comparateurs de chaînes**			
		Jaro	Winkler	Lynch	**Lachance**
SHACKLEFORD	SHACKELFORD	0.970	0.982	0.989	0.909
DUNNINGHAM	CUNNIGHAM	0.896	0.896	0.931	0.842
NICHLESON	NICHULSON	0.926	0.956	0.977	0.889
JONES	JOHNSON	0.790	0.832	0.874	0.667
MASSEY	MASSIE	0.889	0.933	0.953	0.833
DWAYNE	DUANE	0.822	0.840	0.896	0.727
SEAN	SUSAN	0.783	0.805	0.845	0.667

2.2.2 Les comparateurs multi-mots, multi-types

Le comparateur de chaînes de caractères multi-mots procède en deux temps. Il est basé sur l'identification successive de mots appariant d'abord parfaitement, puis de mots appariant de façon légitime, suivi de l'identification de mots inclus les uns dans les autres, mais à partir du début (exemple : Martin versus M). Ceci définit la composante « accord » du comparateur. Dans un deuxième temps, tous les mots qui restent sont considérés comme des conflits ou des extras. C'est la composante « désaccord ». Le comparateur forme ainsi des paires avec les mots restants de chaque côté : ce sont les conflits. Après qu'un maximum de paires eurent été formées, où chaque mot restant n'apparaît qu'une seule fois, s'il reste encore des mots d'un côté, ils correspondent aux extras. Le tableau 2.2.2-1 présente un exemple fictif pour lequel la composante légitime est définie en appliquant le comparateur Lachance. Les paramètres de tolérance sont établis par l'usager et permettent d'établir quelles paires de chaînes de caractères seront considérées comme des liens. À noter que si l'exemple présenté mène à un lien, les mêmes paramètres de tolérance auraient menés à un non lien si « M. MARTIN RICHARD LACHANCE » avait été comparé à « LACNANCE MARTIN RENE », en raison du conflit entre « RICHARD » et « RENE ».

Tableau 2.2.2-1
Application du comparateur de chaînes Lachance multi-mots

Chaînes de caractères comparées	**Paramètres de tolérance**				**Statut de la comparaison**
	Établis par l'usager		Composantes identifiées	Statut du paramètre	
« M. MARTIN R. LACHANCE » versus « LACNANCE MARTIN RENE »	Accord	1 appariement parfait, minimum de 4 caractères	MARTIN, MARTIN	Respecté	Lien
		Appariements légitimes : minimum de 4 caractères, 85% en commun	LACHANCE, LACNANCE	1 paire identifiée	
		Inclusions : minimum de 1 caractère	RENE, R	1 inclusion identifiée	
	Désaccord	0 conflit permis	Aucune!	Respecté	
		(aucune contrainte sur les extras)	M	1 mot identifié	

Une variante du comparateur de chaînes multi-mots existe: le comparateur de chaînes multi-types. À l'instar du comparateur de chaînes multi-mots, il possède les mêmes paramètres de tolérance, mais impose une contrainte supplémentaire. Il exige un nombre minimum d'appariements parfaits sur des nombres isolés (exemple : « 100 » dans « 100, Avenue Symposium »). Entre autres, ce comparateur peut s'avérer intéressant pour le traitement des adresses.

2.3 Améliorer le prétraitement

En dépit de comparateurs de chaînes de caractères sophistiqués, le prétraitement des données demeure souvent presque inévitable. De fait, le prétraitement peut être décomposé en deux étapes principales. La première étape est une phase d'épuration des données, où l'on retire essentiellement des éléments indésirables des chaînes de caractères (exemples : les points, les virgules). La deuxième étape consiste ensuite à faire du recodage, où l'accent est mis sur l'uniformisation vers des valeurs communes afin de favoriser les appariements, lesquels pourraient autrement être manqués. Entre autres, l'emploi de conversions de prénoms en surnoms, tels que « William » en « Bill », ou vice versa selon le logiciel employé, est un exemple typique de recodage.

Tableau 2.3-1
Emploi de divers comparateurs de chaînes suite à divers recodages des données

Chaînes de caractères comparées	**Prétraitement**				**Comparateur de chaînes employé**	**Statut de la comparaison**
	Après épuration	**Après recodage**				
		#	Chaîne 1	Chaîne 2		
« WILLIAM R. SYMPOSIUM » versus « RON BILL SYMPOSIUM »	« R. » devient « R »	1	WILLRSYMP	RONBSYMP	Parfait	Non lien
		2	BILL R SYMPOSIUM	RON BILL SYMPOSIUM	Parfait	Non lien
		2	BILL R SYMPOSIUM	RON BILL SYMPOSIUM	Multi-mots (paramètres spécifiés)	Lien

Généralement, une fois les données recodées, on se limite à identifier des appariements parfaits. Néanmoins, considérer des comparateurs de chaînes sophistiqués peut présenter des avantages, dont une plus grande flexibilité et la possibilité de gagner des liens, tel que le montre l'exemple du tableau 2.3-1. À noter que si l'emploi comporte également des risques, il est toutefois possible de limiter les erreurs de couplage.

2.4 Limiter les erreurs de couplage

Le prétraitement et l'emploi de comparateurs de chaînes sophistiqués peut apporter des liens difficilement repérables, par exemple « MARTIN LACHANCE » avec « MARTIN LACAHNCE ». Toutefois, ils peuvent également apporter leur part de faux liens. À titre d'exemple, il suffit de considérer deux noms de personnes parfaitement valides et distincts, mais se retrouvant appariés en raison d'une forte ressemblance, comme « MARTIN LACHANCE » avec « MARTIN LACASSE », ou « ALEXANDRE » avec « ALEXANDRA » représentant deux personnes de sexe opposé. Alors, comment prévenir les mauvais liens? Une option intéressante consiste à faire appel à des connaissances a priori et à les intégrer au processus d'appariement sous forme de liste, que nous appellerons ici *liste d'exclusions*. Le tableau 2.4-1 présente une liste d'exclusions couvrant les prénoms avec une orthographe proche. L'idée consiste à lister des paires de sous-chaînes de caractères pouvant se retrouver chacune de côté opposé, dans l'une des deux chaînes de caractères comparées, tout comme « GÉRARD » serait d'un côté et « GÉRALD » de l'autre. Si les sous-chaînes sont trouvées, il en résulte que les chaînes comparées sont un non lien.

Tableau 2.4-1
Liste d'exclusions pour les prénoms de personnes

Mots dans la première chaîne de caractères	Mots dans la seconde chaîne de caractères
DAVIS	DAVID
MARC	MARCO
GERALD	GERARD
…	…

Les avantages d'établir de telles listes sont non négligeables. Non seulement ces listes aident à prévenir les mauvais liens, mais il devient alors possible de partager diverses listes d'exclusions entre plusieurs projets de couplage d'enregistrements, permettant ainsi de réduire le volume de résolution manuelle de tous et chacun. Par ailleurs, avec l'ajout de telles listes, le processus de comparaison de chaînes de caractères suivant le prétraitement comporte finalement trois composantes : une composante prévention (listes d'exclusions), une composante recodage (listes de conversions) et le comparateur de chaîne employé. Les exemples du tableau 2.4-2 résument cette séquence.

Tableau 2.4-2
Évaluation complète de paires de chaînes de caractères

	Paire #1	**Paire #2**	**Paire #3**
Chaînes d'origine :	MARTIN LACHNACE MARTY LACHANCE	MARCO SYMPOSIUM MARC SYMPOSIUM	M. NOVEMBRE M. DÉCEMBRE
Après l'application d'une liste d'exclusions :	MARTIN LACHNACE MARTY LACHANCE	Exclusion : MARC, MARCO	M NOVEMBRE M DECEMBRE
Après recodage (conversions) :	MARTY LACHNACE MARTY LACHANCE		M NOVEMBRE M DECEMBRE
Après l'emploi du comparateur de chaînes Lachance (4, 85%):	MARTYLACHNACE MARTYLACHANCE		MNOVEMBRE MDECEMBRE
Résultat de la comparaison :	Lien	Non lien	Non lien

2.5 Fournir des indicateurs de qualité

De bas taux de faux liens et de bas taux de liens manqués représentent des indicateurs de qualité globaux favorables dans un processus de couplage d'enregistrements. Au niveau des composantes du processus de couplage, telles que les comparateurs de chaînes de caractères, le résultat d'une tentative d'appariement sera généralement exprimé par un pourcentage pour indiquer la force du lien établi (voir le tableau 2.2.1-1). Il est toutefois possible de fournir une mesure de force de lien plus informative qu'un simple pourcentage et qui peut être déterminée pour chaque comparateur de chaînes. L'indicateur de qualité PLICE a été conçu pour répondre à cette définition. En effet, le PLICE se présente comme l'expression du résultat de plusieurs petits appariements à l'intérieur d'un processus de comparaison entre deux chaînes de caractères. Il est défini au tableau 2.5-1. Il s'applique aussi bien aux comparateurs de chaînes à mot unique (exemple : le comparateur Lachance) qu'aux comparateurs de chaînes multi-mots ou multi-types. À titre d'exemple, les comparateurs comme Jaro et Winkler, similaires au comparateur Lachance, ne contribueraient qu'au « L » du PLICE, soit à la composante « légitime ». En particulier, cet indicateur permet de séparer plus facilement les liens obtenus en les ordonnant simultanément en ordre décroissant pour la partie « accord » et en ordre croissant pour la partie « désaccord ».

Tableau 2.5-1
Définition de l'indicateur PLICE

	Lettre	**Définition**	**Exemples de liens établis**	
			Martin Henry Lachance versus Lachance Martin Harry	John Symposium versus J. R. Symposuim
Accord	**P**	Nombre d'appariements **P**arfaits observés	2	0
	L	Nombre d'appariements **L**égitimes observés	0	1
	I	Nombre d'**I**nclusions observées	0	1
Désaccord	**C**	Nombre de **C**onflits observés (si acceptés lors de l'emploi d'un comparateur multi-mots)	1	0
	E	Nombre de mots en **E**xtra	0	1

3. Mise en pratique

Toutes les suggestions faites à la Section 2 ont été implantées et incorporées sous le prototype MixMatch, incluant l'indicateur de qualité PLICE. Le logiciel effectue du couplage déterministe. Il est interne à Statistique Canada. Il tient son nom depuis toujours de l'anglais « Mix and Match », définissant sa grande flexibilité. Il a récemment été complètement réécrit en langage informatique SAS dans le cadre d'un projet de recherche, sur les bases d'une version antérieure, écrite en langage informatique C et mise à jour plusieurs années durant. La version SAS représente par ailleurs une version nettement améliorée. Assurément, le logiciel a beaucoup évolué au fil des ans,

mais son développement s'est toujours fait via une constante interaction avec les usagers dès les premières versions, afin de s'adapter le mieux possible à leurs besoins réels. Cette approche de l'auteur lui a permis d'identifier trois qualités chères aux usagers : la simplicité, la flexibilité et la performance, gages de succès jusqu'à présent.

Le principe de fonctionnement de MixMatch est relativement simple. D'abord, trois types de couplage sont possibles : (1) lier des enregistrements en provenance de deux fichiers; (2) identifier des enregistrements apparentés entre eux à l'intérieur d'un même fichier; (3) étant donné des paires d'enregistrements déjà formées, identifier les paires qui correspondent à des liens valides. Quelle que soit la façon dont les paires d'enregistrements sont formées, un ensemble de règles logiques sont appliquées en séquence pour les évaluer. À titre d'exemple, une première règle logique pourrait consister à identifier des paires d'individus qui apparient parfaitement sur le nom, prénom, la date de naissance et le code postal. La règle comporterait quatre conditions. Ensuite, une seconde règle pourrait consister à identifier, parmi les paires restantes, les paires qui apparient parfaitement sur la date de naissance et le code postal, mais pour lesquelles le comparateur de chaînes multi-mots présenté dans cet article serait employé sur une combinaison du nom et du prénom, avec des paramètres de tolérance spécifiques. D'autres règles pourraient emboîter le pas par la suite. Fait notoire, l'ordre d'évaluation des conditions d'une règle a été optimisé pour réduire le temps d'évaluation, selon la complexité des comparateurs employés. Mieux encore, le logiciel s'assure également d'évaluer une seule fois chaque condition, qu'elle soit répétée sur plusieurs règles ou non. Au bout du compte, les paires d'enregistrements ne rencontrant aucune des règles énumérées ne sont pas considérées comme des liens par le logiciel. Finalement, la séquence de règles, fournie par l'usager, combinée aux indicateurs de force PLICE de chaque paire liée, permet l'ordonnancement des liens établis.

4. Conclusion

Les options proposées pour répondre aux besoins majeurs identifiés ont été implantées et ont fait leur preuve à Statistique Canada. Le logiciel de couplage d'enregistrements déterministe MixMatch considère les données des enregistrements comme une seule chaîne de caractères que l'usager peut découper à sa guise. De puissants comparateurs de chaînes ont été élaborés, tels le comparateur Lachance et les comparateurs de chaînes multi-mots et multi-types, lesquels suscitent énormément d'intérêt. La structure interne de MixMatch donne aussi beaucoup de flexibilité aux usagers dans le prétraitement des données. Entre autres, il est possible d'effectuer le recodage d'une séquence de mots en une autre, et ce, de manière itérative. Le logiciel permet également la prévention des erreurs à l'aide de listes d'exclusions. Enfin, un indicateur de force de lien détaillé, le PLICE, a été élaboré et mis en place. En résumé, l'approche globale de simplicité, flexibilité et performance dans le développement et l'implantation des outils présentés a joué un rôle clé. Ceci dit, si MixMatch a été construit pour contrôler le temps de traitement, il reste encore place à l'amélioration. Et outre ce prototype, les idées véhiculées dans cet article peuvent également servir au couplage d'enregistrements probabiliste. De fait, l'intégration des caractéristiques de MixMatch au logiciel officiel de couplage d'enregistrements G-Coup de Statistique Canada (Statistique Canada, 2014) a débuté.

Bibliographie

Lachance, M. (2014), « MixMatch 1.2 - User Guide », document non publié, Ottawa, Canada: Statistique Canada.

Porter, E. H., et Winkler, W. E. (1997), « Approximate String Comparison and its Effect on an Advanced Record Linkage System », Record Linkage Techniques - 1997, *Proceedings of an International Workshop and Exposition, Federal Committee on Statistical Methodology*, pp. 190-199.

Statistique Canada (2014), « Manuel de l'utilisateur pour G-Coup version 3.0 », Ottawa, Canada: Statistique Canada.