Alignement incrémental des données dans des environnements dynamiques

Oumaima El Haddadi[1,2,3], Max Chevalier[1], Bernard Dousset[1], Ahmad El Allaoui[2], Anass El Haddadi[2], Olivier Teste[1]

1. IRIT, SIG, Université de Toulouse, CNRS, France
2. LSA, SDIC, Université Abdelmalek Essaadi Tetouan, Maroc
3. Oumaima.ElHaddadi@irit.fr

RESUME. *De plus en plus de Systèmes d'Informations (SI) sont appliqués dans des environnements dynamiques. Dans de tels environnements, au-delà de la problématique de l'hétérogénéité des données, s'ajoute le problème de la dynamicité des données voire des besoins en informations. Une des réponses apportées pour gérer l'hétérogénéité des données repose sur la notion d'alignement des données. L'objectif de l'alignement est d'identifier des correspondances entre des ensembles de données. Pour considérer ces environnements dynamiques dans leur globalité, une réflexion doit être menée autour des ressources nécessaires pour mettre à jour les correspondances entre les sources de données existantes au fil du temps. Dans cet article, nous nous intéressons particulièrement à la dynamicité des sources afin d'identifier une démarche incrémentale d'alignement des données visant à limiter ces ressources.*

ABSTRACT. *Today, information systems (IS) increasingly applied in dynamic environments. Within such environments, the problem of data heterogeneity is compounded by the problem of dynamic data and information needs. One of the answers to manage data heterogeneity is based on the notion of data alignment. The objective of alignment is to identify correspondences between data sets. To consider these dynamic environments in their entirety, a reflection must be made around the resources needed to update the correspondences between existing data sources over time. In this paper, we focus on the dynamics of the sources in order to identify an incremental approach to data alignment aimed at limiting these resources.*

Mots-clés : Alignement incrémental des données, Graphe, Plongement, Environnements Dynamiques

KEYWORDS: *Incremental Data Alignment, Graph, Embedding, Dynamic Environment*

1. Introduction

Dans des environnements dynamiques, les sources de données sont multiples, variées (les sources de données sont hétérogènes, chaque source a un format et une structure différente) et elles-mêmes dynamiques (les sources de données évoluent avec le temps). D'après (Velegrakis et al., 2003), dans ces environnements « data

sources may change not only their data but also their schemas, their semantics, and their query capabilities ». Ceci implique que les SI doivent considérer les caractéristiques de ces sources de données pour pouvoir répondre au mieux aux besoins en informations qui eux aussi peuvent être dynamiques. Cela signifie que, dans certains contextes (tels que l'aide à la décision, la recherche ou la découverte/exploration de données), les besoins peuvent évoluer au gré des jours et/ou des utilisateurs limitant ainsi l'usage de méthodes manuelles ou semi-automatiques. Ces caractéristiques ont d'autant plus d'impact que les besoins en informations sont critiques comme c'est le cas dans le domaine de l'aide à la décision. Cependant, dans cet article, par soucis de place, nous limitons notre discussion à la dimension dynamique des sources de données. Pour répondre au problème de l'hétérogénéité des données, l'alignement des données (ou « schema matching » en anglais) est essentiel. En effet, l'alignement des données apparait comme un « passage obligé » pour tout processus visant à exploiter et manipuler les données dans de tels environnements. Le résultat de l'alignement est ainsi exploité à la fois dans les approches de stockage centrées besoins – i.e. basées sur l'intégration de données (ou « schema mapping » en anglais) (Miller et al., 2000), mais également dans les approches non centrées besoins telles que les lacs de données (Alserafi et al. 2020) ou encore dans le domaine des ontologies (Aumueller et al., 2005).

De nombreuses méthodes d'alignement des données existent dans la littérature. Cependant, dans le cadre des environnements dynamiques, ces alignements doivent être mis à jour au fur et à mesure de l'évolution des sources. Or, maintenir ces alignements à jour nécessite de nombreux calculs. C'est dans cette optique que nous souhaitons définir une démarche incrémentale d'alignement des données visant à réduire autant que possible les calculs nécessaires. La structure de cet article est la suivante : la section 2 présente l'alignement des données et les évolutions basées sur des graphes dans lesquels la représentation des éléments de ces graphes est sous la forme de plongements (« embedding » en anglais). Sur cette base, nous proposons enfin, en section 3, une démarche générale d'alignement incrémental des données avant de conclure.

2. L'alignement des données : une réponse à l'hétérogénéité

Les données hétérogènes sont toutes les données présentant une variabilité par exemple de types et/ou de formats. Les données peuvent être des tables de base de données, des fichiers XML, des graphes, des ontologies, des séquences de caractères ou tout autre type de données. Au regard de ces multiples exemples, on peut souligner l'hétérogénéité structurelle ou hétérogénéité des schémas. On peut également souligner l'hétérogénéité au niveau de la donnée elle-même (le type, la langue, la valeur, etc.) – via les instances – qui est qualifiée d'hétérogénéité des données. Une classification des approches d'alignement ainsi que la dimension que ces approches considèrent dans le calcul de correspondance est présentée dans (Bernstein et al., 2011). Les approches d'alignement peuvent être notamment linguistiques, syntaxiques, à base de connaissances externes (ontologies,

dictionnaires, …), sémantiques ou encore hybrides. Ces approches peuvent s'appliquer soit au niveau des éléments des schémas (e.g. de la structure) soit au niveau des instances (e.g. des valeurs). Plus récemment, les méthodes d'alignement des données, afin d'améliorer notamment les alignements sémantiques, intègrent les travaux autour de l'apprentissage automatique et notamment le traitement du langage naturel. On peut citer par exemple les travaux de (Zhang et al. 2021, Merieme et al., 2022) qui reposent sur des réseaux de neurones profonds.

On peut également citer les travaux intégrant des approches d'apprentissage de représentation (Bengio et al., 2013) et notamment les plongements (e.g. « embedding »). Un plongement est une représentation vectorielle d'éléments (attributs d'un schéma, instance, …). Ces plongements simplifient le calcul de correspondances puisqu'ils permettent d'utiliser des méthodes classiques de distances car ce sont des vecteurs numériques « classiques ». Dans le cadre de l'alignement des données, ils sont aujourd'hui utilisés pour représenter, souvent dans des espaces réduits et à différents niveaux de granularité, des bases de données ou encore des graphes (Hättasch et al., 2020, Zhao et Castro Fernandez, 2022). Au-delà de l'intégration de ces approches basées sur l'apprentissage automatique, nous pouvons souligner une orientation de certains travaux qui, pour calculer les alignements, transforment, dans un premier temps, les données dans un graphe pour ensuite appliquer des mesures de correspondance sur ce dernier. Dans ce cadre, (Melnik et al., 2002) proposent une approche basée sur les graphes qui intègre à la fois la structure et les instances pour identifier les correspondances.

De manière combinée, des travaux tels que REMA (Koutras et al., 2020) proposent également la transformation des sources de données (structure et instances) sous forme d'un graphe ainsi que l'usage des plongements pour calculer les alignements. Il est à noter que ce type d'approche (graphe + plongements) trouve également des applications dans l'alignement de données via des graphes de connaissances (Ayala et al., 2022). L'intérêt des approches d'alignement des données basées à la fois sur les graphes et les plongements, de notre point de vue, réside tout d'abord dans le fait que le graphe permet d'homogénéiser et peut rendre dynamique la représentation des éléments des différentes sources de données (une base de données relationnelles ou un fichier CSV par exemple). Dans le même temps les plongements permettent d'uniformiser à la fois l'espace de représentation des éléments des schémas et/ou des instances à aligner et le calcul des alignements. Cependant, les méthodes d'alignement des données présentées dans cette section, à notre connaissance, ne prennent pas en compte la notion de dynamicité des sources, c'est-à-dire l'évolutivité des schémas et des instances de manière incrémentale pour optimiser les ressources nécessaires. De ce fait, lors de l'évolution d'une source de données, il sera nécessaire de recalculer les plongements ou stocker les informations qui ont permis de les calculer (les « random walk » par exemple). Cela peut ainsi amener à une utilisation très importantes de ressources (stockage/calcul) surtout lorsque le nombre de sources augmente. Notre objectif est donc de réduire cette utilisation de ressources en mettant à jour autant que possible les plongements de façon incrémentale.

3. Démarche d'alignement incrémental des données

L'alignement incrémental des données vise à intégrer l'évolution des données (schéma/instances) dans le processus d'alignement des données. On doit ainsi considérer l'ajout, la suppression et la mise à jour des éléments décrivant les données (schéma/instances) ainsi que leurs relations. Nous n'avons pas identifié dans la littérature de travaux intégrant la dimension incrémentale pour l'alignement des données tel que défini ci-dessus. À l'inverse, indépendamment de l'alignement des données, nous avons identifié des travaux relatifs au calcul de plongements dans des graphes dynamiques. Deux types d'approches sont identifiées dans (Bielak et al., 2020) : les approches « online » qui mettent à jour les représentations des éléments à chaque modification et les approches « incrémentales » qui font la mise à jour par « lot ». Notre définition d'alignement incrémental des données englobe, à ce niveau de discussion, ces deux temporalités car, souhaitant répondre aux enjeux de mise à jour incrémental des représentations, notre proposition pourrait être utilisée dans ces deux cas. Parmi l'ensemble des approches proposées la plus proche de nos problématiques est sans doute la méthode Online-Node2Vec proposée dans (Béres et al., 2019). Cette méthode vise à mettre à jour le plongement correspondant à chaque nœud d'un graphe au fur et à mesure des modifications. Cependant, la mise à jour des plongements repose notamment sur une fenêtre temporelle excluant les relations existantes dans le graphe les plus « anciennes » ce qui n'est pas pertinent dans le cadre de l'alignement de données. Enfin, Online-Node2Vec vise à faire évoluer la représentation d'un nœud suite à l'ajout d'un nœud et d'une nouvelle relation. Les auteurs ne traitent pas de l'ensemble des mises à jour possibles contrairement à notre objectif.

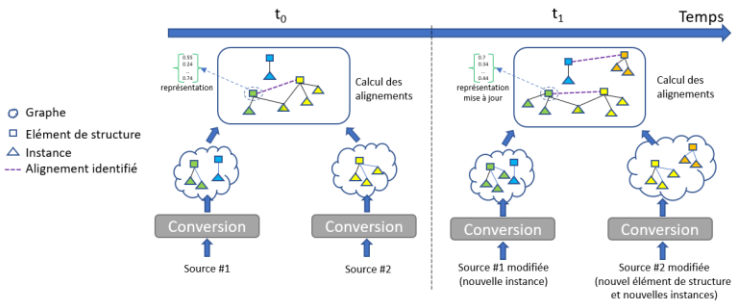

Figure 1 – Illustration générale d'alignement incrémental des données. A t1, une nouvelle instance est ajoutée dans la source #1 impliquant une mise à jour de la représentation de l'élément de structure correspondant et un recalcul des alignements. Par ailleurs, l'ajout d'un nouvel élément de structure et de nouvelles instances dans la source #2 implique le calcul de la représentation correspondante ainsi que le calcul de l'ensemble des alignements possibles (un nouvel alignement été identifié avec un élément de structure de la source #1).

Sur la base de ces travaux, l'illustration d'alignement incrémental des données que nous proposons (Figure 1) repose sur des plongements dans un graphe traduisant

la structure et/ou les instances des sources à aligner. Dans un premier temps, la méthode nécessite la transformation de chaque source de données en un sous-graphe via une méthode de transformation adaptée au format original de la source. Des méthodes de conversion entre différents types de données sont disponibles (https://bit.ly/M2_GM (visité le 03/03/2023)) et peuvent être adaptées. Tout comme dans REMA, nous souhaitons intégrer à la fois la structure et les instances pour identifier les alignements car ces deux niveaux permettent d'augmenter le nombre de nœuds communs potentiels entre les deux sources et ainsi obtenir des représentations plus précises. Avant de calculer les alignements, pour chaque élément de structure, nous calculons leur représentation (i.e. un plongement). Cependant, le défi de notre méthode qu'il nous faut remporter est d'identifier la méthode de plongement la plus adaptée à la dimension incrémentale. Nous recherchons donc actuellement une méthode de représentation des nœuds supportant nativement ou après adaptation l'ensemble des opérations de mise à jour que ce soit en ajout comme c'est le cas dans Online-Node2Vec mais également en suppression/modification. A titre d'illustration, la Figure 1 ne présente qu'un exemple de modifications des sources existantes car l'ajout de nouvelles sources nécessite le calcul exhaustif des représentations correspondantes à cette nouvelle source. Nous avons donc aujourd'hui certaines bases qui malheureusement ne traitent pas, sauf erreur de notre part, l'ensemble du problème. Après avoir calculé ou mis à jour les représentations concernées par les modifications, les alignements sont calculés sur la base de ces dernières. Pour optimiser ce calcul nous privilégions l'usage d'une méthode de hachage, e.g. MinHash (Broder et al., 2000) ou d'une matrice de similarités.

4. Conclusion et Perspectives

L'exploitation des données dans des environnements dynamiques nécessite à la fois la prise en compte de l'hétérogénéité mais également la dynamicité des données et des besoins. Pour optimiser l'alignement des données nous proposons les premières bases d'une démarche incrémentale reposant sur un graphe traduisant les sources de données et dont chaque nœud est représenté par un plongement. Ces plongements doivent supporter une mise à jour incrémentale et/ou ne pas nécessiter trop de ressources pour leur calcul (stockage/calcul). Les objectifs de notre démarche visent à limiter le coût des différentes mises à jour nécessaires pour maintenir les alignements à jour. Notre travail actuel a pour objectif d'affiner cette démarche en identifiant la bonne méthode de représentation des nœuds du graphe afin de la rendre opérationnelle et d'évaluer celle-ci. Concernant la validation, nous souhaitons réaliser des expérimentations sur des jeux de données de différents domaines, ayant différentes caractéristiques, afin d'évaluer les différents cas liés à la dynamique des données. A moyen terme, vu que l'orientation actuelle de nos travaux s'intéresse particulièrement sur les plongements pour représenter les données, nous n'excluons pas d'étudier d'autres approches potentiellement intéressantes et optimisables pour identifier les alignements et ce de façon incrémentale. A plus long terme, nous souhaitons considérer la problématique de la dynamique des besoins.

Bibliographie

Alserafi, A., Abelló, A., Romero, O., Calders T., 2020. Keeping the Data Lake in Form: Proximity Mining for Pre-Filtering Schema Matching. ACM Trans. Inf. Syst. 38, 3.

Aumueller, D., Do, H-H., Massmann, S., Rahm, E., 2005. Schema and ontology matching with COMA++. ACM international conference on Management of data (SIGMOD '05). Association for Computing Machinery, New York, NY, USA, 906–908.

Ayala, D., Hernández, I., Ruiz, D., Rahm, E., 2022. LEAPME: Learning-based Property Matching with Embeddings. Data & Knowledge Engineering, Volume 137.

Bengio, Y., Courville, A., Pascal, V., 2013. Representation Learning: A Review and New Perspectives. IEEE trans. on pattern analysis and machine intelligence. 35. 1798-1828.

Béres, F., Kelen, D.M., Pálovics, R., Benczúr, A.A., 2019. Node embeddings in dynamic graphs. Applied Network Science 4, 64. 10.1007/s41109-019-0169-5

Bernstein, P., Jayant, M., Rahm, E., 2011. Generic Schema Matching, Ten Years Later. PVLDB. 4. 695-701. 10.14778/3402707.3402710.

Bielak, P., Tagowski, K., Falkiewicz, M., Kajdanowicz, T., Chawla, N.V., 2020. FILDNE: A Framework for Incremental Learning of Dynamic Networks Embeddings. Knowledge-Based Systems, 236. doi.org/10.1016/j.knosys.2021.107453

Broder, AZ, Charikar, M, Frieze, AM, Mitzenmacher, M, 2000. Min-wise independent permutations. Journal of Computer System Science 60(3):630–659.

Hättasch, B., Truong-Ngoc, M., Schmidt, A., Binnig, C., 2020. It's AI Match: A Two-Step Approach for Schema Matching Using Embeddings. 2nd International Workshop on Applied AI for Database Systems and Applications (AIDB'20).

Koutras, C., Fragkoulis, M., Katsifodimos, A., Lofi, C., 2020. REMA: Graph Embeddings-based Relational Schema Matching. EDBT/ICDT Workshops.

Melnik, S., Garcia-Molina, H., Rahm, E., 2002. Similarity flooding: a versatile graph matching algorithm and its application to schema matching. Proceedings 18th International Conference on Data Engineering, San Jose, CA, USA, 2002, pp. 117-128.

Merieme, E.A., Mohamed, A., Ali, C., Fakhri, Y., Noreddine, G., 2022. Schema Matching Based On Deep Learning Using LSTM Model, 3rd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), pp. 1–5.

Miller, RJ, Haas, LM, Hernandez, MA, 2000. Schema Mapping as Query Discovery. Very Large DataBase conference (VLDB), pp 77–88.

Velegrakis, Y., Miller, R. J., Popa, L., 2003. Mapping adaptation under evolving schemas. In Proceedings of the 29th VLDB'03, Vol. 29. VLDB Endowment, 584–595.

Zhang, J., Shin, B., Choi, J.D., Ho, J.C., 2021. SMAT: An Attention-Based Deep Learning Solution to the Automation of Schema Matching. Advances in Databases and Information Systems, LNCS. Springer International Publishing, pp. 260–274.

Zhao, Z., Castro Fernandez, R., 2022. Leva: Boosting Machine Learning Performance with Relational Embedding Data Augmentation, in: Proceedings of the 2022 International Conference on Management of Data. ACM, Philadelphia PA USA, pp. 1504–1517.