DAGOBAH : Activités de recherche Orange autour de l'annotation sémantique de données tabulaires

Yoan Chabot, Pierre Monnin

Orange, France
{yoan.chabot, pierre.monnin}@orange.com

Un grand nombre de gisements de données internes aux entreprises ainsi qu'une part non-négligeable des données du Web sont représentés sous forme de tables. La capacité à annoter ces données à l'aide de graphes de connaissances est cruciale et permet d'ouvrir la voie à de nouveaux services basés sur la sémantique (Chabot et al., 2019a).

Dans cet exposé, nous définirons le problème de l'annotation sémantique et dresserons un panorama des approches existantes. Nous structurons cet état de l'art autour d'une décomposition classique en trois étapes :

— CEA (Cell-Entity Annotation) visant à associer, à chaque cellule d'une table, une ou plusieurs entités d'un graphe de connaissances à l'aide de techniques de lookups syntaxiques, d'alignement d'ontologies ou encore de plongements de graphes (Efthymiou et al., 2017; Kilias et al., 2018).

— CTA (Column-Type Annotation) dont le but est d'associer, à chaque colonne de la table, un type issu du graphe de connaissances à l'aide de méthodes telles que le vote majoritaire (Mulwad et al., 2010).

— CPA (Columns-Property Annotation), enfin, permettant d'identifier des propriétés sémantiques entre des paires de colonnes (Ran et al., 2015).

Nous aborderons ensuite les enjeux de l'annotation de données tabulaires pour une entreprise comme Orange. Les efforts de recherche du groupe sur ce sujet, cristallisés au sein d'un projet nommé DAGOBAH (Chabot et al., 2019b, 2020; Huynh et al., 2020), seront présentés avec un focus sur des techniques de plongements de graphes de connaissances pour le typage de colonnes et la désambiguïsation des cellules. Enfin, cet exposé s'attardera sur les efforts en cours au sein de la communauté scientifique autour de ces questions par le biais du challenge ISWC SemTab (Cutrona et al., 2020; Jiménez-Ruiz et al., 2020a,b).

Références

Chabot, Y., P. Grohan, G. L. Calvez, et C. Tarnec (2019a). Dataforum : Faciliter l'échange, la découverte et la valorisation des données à l'aide de technologies sémantiques. In *Extraction et Gestion des connaissances, EGC 2019, Metz, France, January 21-25, 2019*, Volume E-35 of *RNTI*, pp. 441–444. Éditions RNTI.

Chabot, Y., T. Labbé, J. Liu, et R. Troncy (2019b). DAGOBAH : an end-to-end context-free tabular data semantic annotation system. In *Proceedings of the Semantic Web Challenge*

on Tabular Data to Knowledge Graph Matching co-located with the 18th International Semantic Web Conference, SemTab@ISWC 2019, Auckland, New Zealand, October 30, 2019, Volume 2553 of *CEUR Workshop Proceedings*, pp. 41–48. CEUR-WS.org.

Chabot, Y., T. Labbé, J. Liu, et R. Troncy (2020). DAGOBAH : Un système d'annotation sémantique de données tabulaires indépendant du contexte. In *IC 2020 : 31es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 31st French Knowledge Engineering Conference), Angers, France, June 29 - July 3, 2020*, pp. 120–132.

Cutrona, V., F. Bianchi, E. Jiménez-Ruiz, et M. Palmonari (2020). Tough tables : Carefully evaluating entity linking for tabular data. In *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II*, Volume 12507 of *Lecture Notes in Computer Science*, pp. 328–343. Springer.

Efthymiou, V., O. Hassanzadeh, M. Rodriguez-Muro, et V. Christophides (2017). Matching web tables with knowledge base entities : From entity lookups to entity embeddings. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I*, Volume 10587 of *Lecture Notes in Computer Science*, pp. 260–277. Springer.

Huynh, V., J. Liu, Y. Chabot, T. Labbé, P. Monnin, et R. Troncy (2020). DAGOBAH : enhanced scoring algorithms for scalable annotations of tabular data. In *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2020) co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 5, 2020*, Volume 2775 of *CEUR Workshop Proceedings*, pp. 27–39. CEUR-WS.org.

Jiménez-Ruiz, E., O. Hassanzadeh, V. Efthymiou, J. Chen, et K. Srinivas (2020a). Semtab 2019 : Resources to benchmark tabular data to knowledge graph matching systems. In *The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings*, Volume 12123 of *Lecture Notes in Computer Science*, pp. 514–530. Springer.

Jiménez-Ruiz, E., O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, et V. Cutrona (2020b). Results of semtab 2020. In *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2020) co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 5, 2020*, Volume 2775 of *CEUR Workshop Proceedings*, pp. 1–8. CEUR-WS.org.

Kilias, T., A. Löser, F. A. Gers, R. Koopmanschap, Y. Zhang, et M. Kersten (2018). Idel : In-database entity linking with neural embeddings.

Mulwad, V., T. Finin, Z. Syed, et A. Joshi (2010). Using linked data to interpret tables. In *Proceedings of the First International Workshop on Consuming Linked Data, Shanghai, China, November 8, 2010*, Volume 665 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ran, C., W. Shen, J. Wang, et X. Zhu (2015). Domain-specific knowledge base enrichment using wikipedia tables. In *2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015*, pp. 349–358. IEEE Computer Society.