Formulaire de renseignements du mémoire de fin d'études

Année Universitaire : 2022 / 2023

Code Mémoire: 23/1453

IDENTIFICATION DES INTERVENANTS

Titre du mémoire	Détection et réparation des anomalies dans les données et les règles de qualité
Spécialités	Systèmes Intelligents et Données Systèmes Informatiques et Logiciels Systèmes Informatiques Systèmes d'Information et Technologies
Stagiaires	**Matricule:** 18/0130 **Nom:** MEFLAH **Prénom:** WIDED **Email:** iw_meflah@esi.dz
Affiliation du promoteur	**Organisme:** Ecole nationale Supérieure d'Informatique **Adresse:** BP68M Oued Smar, Alger, Algérie **Nom et Prénom:** Sabrina ABDELLAOUI **Tel:** 0558953389, **Email:** s_abdellaoui@esi.dz
Co-encadrants	ABDELLAOUI Sabrina

DESCRIPTION DU PROJET DE FIN D'ETUDES

Résumé	De nos jours, l'importance de la qualité des données est plus reconnue que jamais. En effet, une étude d'Harvard a dévoilé que les données répondant aux standards de qualité dans les entreprises ne dépassent pas 3% . Cette qualité médiocre des données augmente les chances de prendre de mauvaises décisions ainsi que d'avoir des pertes financières estimée à 25% du chiffre d'affaire des entreprises [Haque et chiang, 2019]. Parallèlement aux pertes financières, le cout de la non-qualité en termes de temps est aussi élevé. Environ 60% des data scientistes affirment perdre beaucoup de temps en traitant les données de mauvaise qualité [Ilyas et Xu, 2015]. La gestion de la qualité de données englobe principalement la détection des anomalies et leur réparation. La détection des anomalies repose généralement sur des Règles de Qualité (RQ) qui sont employées comme un moyen déclaratif pour capturer les inconsistance de données. Tout sous-ensemble de données qui n'est pas conforme aux règles définies est considéré comme étant une violation [Bohannon et al., 2005, Cong et al., 2007, Yakout et al., 2011, Dallachiesa et al., 2013, Fan et al., 2014a, Qahtan et al., 2019].

	La réparation de données vise à réparer les anomalies détectées en minimisant la distance entre la base de données originale et la base de données modifiée selon une certaine fonction de coût [Geerts et al., 2013, Fan et al., 2014b, Chu et al., 2013, Berti-Equille et al., 2019]. Afin d'améliorer la précision du processus de réparation, plusieurs travaux impliquent les utilisateurs dans le processus de réparation pour la vérification des réparations générées automatiquement ou bien pour la suggestion des réparations [Yakout et al., 2011, Fan et al., 2012, Mayfield et al., 2010, Abdellaoui et al., 2017, Rezig et al., 2019]. De plus, des études émergentes considèrent des sources externes telles que les master data ou bien les bases de connaissances comme données de référence pour obtenir une plus grande précision des réparations générées [Chu et al., 2015, Abdellaoui et al., 2017]. La majorité des approches de réparation existantes modifient les données de sorte à ceux qu'elles respectent les RQs spécifiées. Cependant, les RQs peuvent être inexactes en raison du bruit inhérent à leur découverte ou bien elles peuvent devenir obsolètes du fait que la logique métier n'est pas statique et peut évoluer avec le temps. Pour pallier à ce problème [Golab et al., 2008] proposent une réparation par ajout, suppression et raffinement de RQs. Néanmoins, ce travail émet l'hypothèse que les données sont propres. Très peu d'approches considèrent que la source d'une violation peut être aussi bien dans les données que dans les RQs [Chiang et Sitaramachandra ,2016, Volkovs et al., 2014, Chiang et Miller, 2011, Beskales et al, 2013]. Nous soutenons dans le cadre de ce PFE que la réparation doit impérativement prendre en compte des erreurs potentielles dans les données ainsi que dans les RQs spécifiées. De ce fait, la réparation doit cibler les données et les RQs simultanément.
Mots clés	qualité de données, nettoyage de données, base de connaissance, détection des anomalies
Objectifs	-Effectuer un état de l'art sur la gestion de la qualité des données. -Identifier l'origine des l'anomalies (données, règles de qualité, données et règles de qualité en même temps) -Proposer une approche de réparation des données et des RQs simultanément. -Réaliser des expérimentations et une comparaison avec les approches existantes en utilisant des données réelles et synthétiques.
Résultats attendus	Conception et réalisation d'une approche pour la détection et réparation des anomalies dans les données et les règles de qualité
Antécédents du travail demandé	Aucun
Bibliographie	[Abdellaoui et al., 2017] Abdellaoui, S., Nader, F., and Chalal, R. (2017). Qdflows : A system driven by knowledge bases for designing quality-aware data flows. Journal of Data and Information Quality (JDIQ), 8(3-4) :14. [Bohannon et al., 2005] Bohannon, P., Fan, W., Flaster, M., and Rastogi, R. (2005). A cost-based model and effective heuristic for repairing constraints by value modification. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 143–154. ACM. [Chu et al., 2013] Chu, X., Ilyas, I. F., and Papotti, P. (2013). Holistic data cleaning : Putting violations into context. In 2013 IEEE 29th International Conference on Data Engineering, pages 458–469. IEEE. [Chu et al., 2015] Chu, X., Morcos, J., Ilyas, I. F., Ouzzani, M., Papotti, P., Tang, N., and Ye, Y. (2015). Katara : A data cleaning system powered by knowledge bases and crowdsourcing. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pages 1247–

1261. ACM.
[Cong et al., 2007] Cong, G., Fan, W., Geerts, F., Jia, X., and Ma, S. (2007). Improving data quality : Consistency and accuracy. In Proceedings of the
33rd international conference on Very large data bases, pages 315–326. VLDB Endowment.
[Dallachiesa et al., 2013] Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A., Ilyas, I. F., Ouzzani, M., and Tang, N. (2013). Nadeef : a commodity data cleaning system. In Proceedings of the 2013 ACM SIGMOD international conference on management of data, pages 541–552. ACM.
[Fan et al., 2014a] Fan, G., Fan, W., and Geerts, F. (2014a). Detecting errors
in numeric attributes. In International Conference on Web-Age Information Management, pages 125–137. Springer.
[Fan et al., 2012] Fan, W., Li, J., Ma, S., Tang, N., and Yu, W. (2012). Towards certain fixes with editing rules and master data. The VLDB Journal—The International Journal on Very Large Data Bases, 21(2) :213–238.
[Fan et al., 2014b] Fan, W., Ma, S., Tang, N., and Yu, W. (2014b). Interaction between record matching and data repairing. Journal of Data and
Information Quality (JDIQ), 4(4) :16.
[Geerts et al., 2013] Geerts, F., Mecca, G., Papotti, P., and Santoro, D.
(2013). The llunatic data-cleaning framework. Proceedings of the VLDB Endowment, 6(9) :625–636.
[Mayfield et al., 2010] Mayfield, C., Neville, J., and Prabhakar, S. (2010).
Eracer : a database approach for statistical inference and data cleaning.
In Proceedings of the 2010 ACM SIGMOD International Conference on
Management of data, pages 75–86. ACM.
[Qahtan et al., 2019] Qahtan, A., Tang, N., Ouzzani, M., Cao, Y., and Stonebraker, M. (2019). Anmat : Automatic knowledge discovery and error
detection through pattern functional dependencies. In Proceedings of the 2019 International Conference on Management of Data, pages 1977–1980. ACM.
[Rezig et al., 2019] Rezig, E. K., Ouzzani, M., Elmagarmid, A. K., Aref, W. G., and Stonebraker, M. (2019). Towards an end-to-end human-centric data cleaning framework. In Proceedings of the Workshop on Human-Inthe-Loop Data Analytics, page 1. ACM.
[Yakout et al., 2011] Yakout, M., Elmagarmid, A. K., Neville, J., Ouzzani, M., and Ilyas, I. F. (2011). Guided data repair. Proceedings of the VLDB Endowment, 4(5) :279–289
[Haque, Enamul, et Fei Chiang. 2019]. « Restoring Consistency in Ontological Multidimensional Data Models via Weighted Repairs ». Procedia Computer Science 159:1085-94. doi: 10.1016/j.procs.2019.09.277.
[Song, S., et al. 2016] Constraint-Variance Tolerant Data Repairing. Proceedings of the 2016 International Conference on Management of Data. San Francisco, California, USA, Association for Computing Machinery: 877–892. https://doi.org/10.1145/2882903.2882955.
[Chiang, F., & Miller, R. J. 2011]. A unified model for data and constraint repair. 2011 IEEE 27th International
Conference on Data Engineering, 446-457. https://doi.org/10.1109/ICDE.2011.5767833
[Beskales, G., et al., 2013]. On the relative trust between inconsistent
data and inaccurate constraints. 2013 IEEE 29th International Conference on Data Engineering (ICDE),
541-552. https://doi.org/10.1109/ICDE.2013.6544854
[Volkovs et al., 2014] Volkovs, M., Fei Chiang, Szlichta, J., & Miller, R. J. (2014). Continuous data cleaning. 2014 IEEE 30th
International Conference on Data Engineering, 244-255.
https://doi.org/10.1109/ICDE.2014.6816655

Echéancier	Etat de l'art sur la qualité des données 2 mois Conception de l'approche proposée 3 mois Expérimentation et prototypage 2.5 mois Finalisation de la rédaction du mémoire 1.5 mois
Moyens informatiques	A la charge de l'école

Projet de recherche	None