MGNLP-Explorer: outil de visualisation pour l'exploration de jeux de données

Stage M2 - Laboratoire I3S / Inria

Mots-clés : visualisation de données, génération et enrichissement de données, graphes

Connaissances nécessaires : Python ou équivalent pour le traitement des données; connaissance ou intérêt par la visualisation de données ; traitement automatique de langues

Contact :
Anaïs Ollagnier, AI Fellow, 3IA, Université Côte d'Azur
Aline Menin, Maître de Conférences, Université Côte d'Azur

Localisation :
Laboratoire I3S, Campus SophiaTech.
Inria Sophia Antipolis, Wimmics team (http://wimmics.inria.fr/)

Date et durée :
6 mois, de Mars à Septembre 2024

Financement du stage. Le montant de gratification est celui en vigueur à UniCA, soit approximativement de ~580 €/mois

Candidatures. Les candidatures (CV, notes et lettre de motivation) doivent être adressées **avant le 15 janvier 2024** à Anaïs Ollagnier (ollagnier@i3s.unice.fr) et Aline Menin (aline.menin@inria.fr).

Contexte et Objective :

Parvenir à localiser et examiner la pléthore de jeux de données disponibles en libre accès est devenue de plus en plus complexe, nécessitant l'utilisation d'outils de navigation

adaptés. Cette complexité étant alimentée par l'évolution continue des systèmes d'IA – dont le développement repose sur l'utilisation de jeux de données dans le but de les former, les tester et les évaluer – nous faisons aujourd'hui face à une abondance de corpus à la pluralité aussi vaste que le nombre de tâches à effectuer. L'approche prédominante pour la recherche et la navigation dans les ensembles de données repose sur des requêtes basées sur des mots-clés, fournissant aux utilisateurs une liste brute de suggestions. Ces approches présentent deux limitations principales : la nécessité de connaissances préalables pour la rédaction de requêtes correspondant aux besoins des utilisateurs et la dissimulation des éventuelles interconnexions entre les ensembles de données. Ces deux points sont cruciaux car ils impactent directement l'accès à l'information. Afin de répondre à ces limitations, nous proposons d'explorer l'utilisation de techniques de visualisation dans le but de supporter les tâches de recherche, d'exploration mais également de découverte de jeux de données.

Nous souhaitons mettre à profit l'utilisation de vues chaînées (séquence de représentations visuelles interconnectées) mise en place par MGExplorer dans le cadre du développement d'une application de navigation de jeux de données appliquée au domaine du Traitement Automatique des Langues (TAL). Cette réalisation requiert donc une expertise dans le domaine du TAL en vue de répondre aux attentes dans ce domaine, il est également question de mettre l'expérience utilisateur au cœur du développement de cette application via des techniques issues du domaine de l'interaction homme-machine.

Tâches Associées :

1. Génération de métadonnées. La recherche de jeux de données consiste généralement en une requête basée sur des mots-clés ou une expression en langage de requête contextuel utilisée pour explorer les métadonnées publiées. En conséquence, des suggestions classées sont produites en fonction de la similarité des métadonnées par rapport à la requête de l'utilisateur. Cependant, ces méthodes sont entravées par la disponibilité de métadonnées appropriées et par leur qualité médiocre. L'une des solutions envisagées pour pallier ces limitations consiste à générer automatiquement des métadonnées. Cela permettrait à la fois d'unifier mais également d'enrichir la description déjà existante des jeux de données. Les enjeux à ce stade sont multiples. Il est tout d'abord question d'intégrer des normes de métadonnées spécifiques au domaine combinées avec des vocabulaires contrôlés, comme suggéré dans le modèle de maturité des données de la RDA (https://www.rd-alliance.org/). Secondement, générer des métadonnées plus riches à partir des informations descriptives des données

2. Prise en compte de l'environnement académique Comme la plupart des ensembles de données sont introduits à travers un article académique, l'intégration d'une dimension d'analyse liée à l'environnement académique de ces ensembles serait une source d'information enrichissante notamment dans leurs représentations. Il s'agirait d'éléments sur lesquels générer de nouvelles topologies de graphes permettant la visualisation des données via de nouvelles perspectives. Par exemple, l'extraction du réseau de citations lié à un jeu de données permettrait de cartographier l'utilisation scientifique de cet ensemble en

fonction des pratiques de citation. De plus, cela contribuerait à révéler l'utilisation “cachée” de l'ensemble de données donné dans les publications de recherche et peut ainsi aider les utilisateurs à savoir si une ressource spécifique est populaire/influente et dans quelle sous-communauté. Il peut également être question de dériver une fonctionnalité de chronologie qui se concentre sur les benchmarks récents, car plus pertinents en raison de leur alignement avec les normes et les tendances actuelles du domaine. Les enjeux sont ici de parvenir à dériver des connaissances à partir de l’exploitation de l’environnement académique et de les adjoindre dans la représentation des datasets en tant que support à la sélection.

3. Valorisation des retours d’expérience. Une fois les données collectées et enrichies, la tâche consiste à générer des visualisations pertinentes pour mettre en évidence les connaissances sous-jacentes aux données, mais aussi de guider les utilisateurs dans la découverte des jeux de données pertinents à leur recherche. L’objectif dans un premier temps est d’intégrer ces données dans l’outil MGExplorer permettant ainsi une exploration visuelle des données. Dans un second temps, il s’agit d’explorer à travers des expériences utilisateurs l’efficacité et utilisabilité de cette solution pour soutenir les usagers du domaine. Les résultats obtenus de ces expérimentations permettront de faire évoluer l’approche en termes de collecte, enrichissement, et visualisation de données.

Références

[1] D. Maier, V. M. Megler, and K. Tufte, “Challenges for dataset search,” in Database Systems for Advanced Applications - 19th International Conference, DASFAA 2014, Bali, Indonesia, April 21-24, 2014. Proceedings, Part I, ser. Lecture Notes in Computer Science, S. S. Bhowmick, C. E. Dyreson, C. S. Jensen, M. Lee, A. Muliantara, and B. Thalheim, Eds., vol. 8421. Springer, 2014, pp. 1–15. [Online]. Available: https://doi.org/10.1007/978-3-319-05810-8 1

[2] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. Ibáñez, E. Kacprzak, and P. Groth, “Dataset search: a survey,” VLDB J., vol. 29, no. 1, pp. 251–272, 2020. [Online]. Available: https://doi.org/10.1007/s00778-019-00564-x

[3] N. W. Paton, J. Chen, and Z. Wu, “Dataset discovery and exploration: A survey,” ACM Comput. Surv., oct 2023, just Accepted. [Online]. Available: https://doi.org/10.1145/3626521

[4] A. Menin, R. Cava, C. M. D. S. Freitas, O. Corby, and M. Winckler, “Towards a visual approach for representing analytical provenance in exploration processes,” in 25th International Conference Information Visualisation (IV), 2021.