HAL open science

Actes de l'atelier " La place des usagères et usagers dans les outils de fouille et d'exploration de données " (PAUL)

Jérôme Darmont, Cécile Favre, Fabien Labarthe, Gilles Venturini

▶ **To cite this version:**

Jérôme Darmont, Cécile Favre, Fabien Labarthe, Gilles Venturini. Actes de l'atelier " La place des usagères et usagers dans les outils de fouille et d'exploration de données " (PAUL). 24e conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC), 2024. hal-04509930

HAL Id: hal-04509930

https://hal.science/hal-04509930

Submitted on 18 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Actes de l’atelier « La place des usagères et usagers dans les outils de fouille et d'exploration de données » (PAUL)

Conférence EGC 2024 – Dijon, 23/01/2024 – https://eric.univ-lyon2.fr/bi4people/paul/

Sommaire

Jérôme Darmont
Bilan du projet BI4people – Le décisionnel pour tou·tes (résumé)

Raphaëlle Bour (conférence invitée)
Citoyen de l’organisation, citoyen du système d’information – approche pluridisciplinaire (résumé long)

Margaux Coeuret, Joris Falip et Aurélien Bénel
Co-construction d’un outil d’assistance à l’analyse de textes par théorisation ancrée (résumé)

Muhammad Fahad
Une plateforme d’informatique décisionnelle collaborative (résumé)

Olga Cherednichenko et Muhammad Fahad
Towards Collaborative Business Intelligence with Conversational Agents: A User-Centric Approach (article)

Praveen Soni, Cyril de Runz, Fatma Bouali et Gilles Venturini
Un outil de génération automatique de tableaux de bord (démo)

Porteurs

- Jérôme Darmont, ERIC Université Lyon 2 (président de séance)
- Cécile Favre, ERIC Université Lyon 2
- Fabien Labarthe, ELICO Université de Saint-Étienne
- Gilles Venturini, LIFAT Université de Tours

Commité de programme

- Fatma Abdelhédi, Laboratoire CBI², Société TRIMANE
- Tiffany Andry, Université libre de Bruxelles
- Sandro Bimonte, INRAE Clermont-Ferrand
- Julia Bonaccorsi, ELICO Université Lyon 2
- Laurent d'Orazio, IRISA Université de Rennes
- Mohamed-Lamine Messai, ERIC Université Lyon 2
- Veronika Peralta Costabel, LIFAT Université de Tours
- Franck Ravat, IRIT Université Toulouse Capitole

L'atelier PAUL@EGC 2024 est soutenu par le projet ANR-19-CE23-0005 BI4people "Le décisionnel pour tou·tes".

Bilan du projet BI4people – Le décisionnel pour tou·tes

Jérôme Darmont

Université Lyon 2, UR ERIC
jerome.darmont@univ-lyon2.fr – https://eric.univ-lyon2.fr/bi4people/

Les technologies de l'informatique décisionnelle (*business intelligence* ou BI), telles que les entrepôts de données et l'analyse en ligne (*on-line analytical processing* ou OLAP), sont des outils primordiaux dans l'aide à la décision. Il existe des solutions de BI gratuites. Celles qui sont propriétaires se focalisent cependant sur les tableaux de bord et la visualisation et ont tous des fonctionnalités limitées pour intégrer efficacement des données de sources disparates. Les quelques logiciels libres qui proposent des explorations OLAP demeurent techniquement hors de portée des petites entreprises, des associations, des chercheurs, des indépendants comme des journalistes ou des *makers*, et des citoyens actifs, que nous ciblons particulièrement. Finalement, la tendance à déporter la BI dans le nuage rejoint la demande grandissante d'outils collaboratifs qui permettent aux utilisateurs de croiser des données privées, publiques, ainsi que des *self data*, d'effectuer des analyses conjointes, d'annoter des figures ou des rapports et de communiquer via les réseaux sociaux. Les réponses actuelles à cette demande globale restent en deçà des attentes en se limitant au partage en ligne de résultats d'analyse.

L'objectif du projet BI4people est de rendre accessible la puissance de l'analyse interactive OLAP à la plus large audience possible, en mettant en œuvre le processus d'entreposage de données en mode *software-as-a-service*, de l'intégration de données multisource, hétérogènes (typiquement sous la forme de tableaux issus de tableurs, de documents textuels ou semi-structurés, ou encore du Web) à une analyse OLAP et une visualisation très simples (Figure 1).

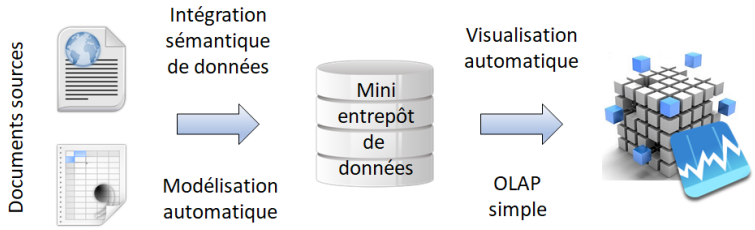

FIG. 1 – *Processus global du projet BI4people*

Pour atteindre ce but, le service de BI doit inclure la *privacy by design*, être autonome, extrêmement simple, ergonomique et intelligible (jargon informatique ou BI interdit !). Dans ce contexte, les étapes classiques de l'entreposage de données s'appliquent, mais doivent être complètement automatisées. Grabova et al. (2014) avaient présenté un premier prototype logiciel très simple, qui a servi de brique de base au projet BI4people (Darmont et Sdiri, 2022) et

a atteint l'automatisation quasi-complète. De plus, notre livrable logiciel permet des analyses collaboratives, malheureusement moins poussées que prévues par manque de main d'œuvre. De même, la prise en charge la confidentialité des données dans toutes les étapes n'a pas pu être assurée. Enfin, nous insistons sur l'importance de l'appropriation des visualisations fournies par l'outil par les utilisateurs, ce qui implique une collaboration interdisciplinaire entre l'informatique et les sciences de l'information et de la communication, bien que la crise COVID ait fortement entravé nos efforts en ce sens.

La conception, l'implémentation, l'évaluation et la valorisation de BI4people inclut 8 *work packages* menés par 5 partenaires (Figure 2) : WP1 Coordination du projet (ERIC Lyon), WP2 Entreposage automatique des données (IRIT Toulouse), WP3 Analyse collaborative (ERIC), WP4 Visualisation assistée (LIFAT Tours), WP5 Confidentialité des données (ERIC), WP6 Validation expérimentale (entreprise TRIMANE Saint-Germain-en-Laye), WP7 Évaluation de l'appropriation par les utilisateurs (ELICO Lyon), WP8 Dissémination et exploitation (tous).

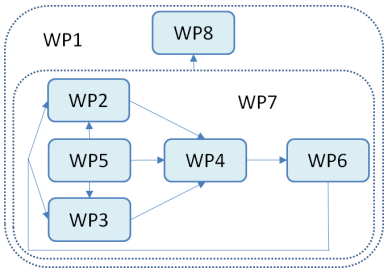

FIG. 2 – *Dépendances des work packages du projet BI4people*

Ce résumé a pour objectif de présenter le bilan du projet ANR-19-CE23-0005 « BI4people : Le décisionnel pour tou·tes », financé par l'Agence nationale de la recherche.

Références

Darmont, J. et B. Sdiri (2022). Projet bi4people : Le décisionnel pour tout.es - saison 2. 18e journées Business Intelligence et Big Data (EDA 2022), Clermont-Ferrand ; RNTI, Vol. B-18, 107-108. Poster.

Grabova, O., S. Sobati-Moghadam, S. Chagheri, et J. Darmont (2014). BI4people : le décisionnel pour tous. 10es journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2014), Vichy ; RNTI, Vol. B-10, 155-158. Démo.

Summary

This abstract depicts the outcome of the BI4people project, which aims to bring the power of OLAP interactive analysis to the largest possible audience, by implementing the data warehousing process in software-as-a-service mode, from multisource, heterogeneous data integration to very simple OLAP-like analysis and dataviz.

Citoyen de l'organisation, citoyen du système d'information – approche pluridisciplinaire

Raphaëlle Bour*

*raphaelle.bour@irit.fr
IRIT UT1 Capitole/société COADA
http://www.si2d.hypotheses.org

La question du citoyen dans l'organisation, et donc de la place de citoyen face aux systèmes d'information numériques dans l'organisation est au cœur des problématiques de l'atelier « Système d'Information et de Décision et Démocratie » (SI2D). L'article propose de présenter succinctement l'approche pluridisciplinaire adoptée pour traiter de cette question.

1. Approche pluridisciplinaire de l'atelier SI2D

L'atelier SI2D est pluridisciplinaire. Il se rassemble autour d'un noyau dur de huit chercheurs en informatique, économie, mathématiques, droit et géographie, exerçant pour certains leur activité hors université : société de conseil, travail dans une administration publique, enseignement dans le secondaire. D'autres chercheurs issus de la sociologie et de la gestion participent à cet atelier de façon plus ponctuelle.

Cette diversité de profils fait la force de l'atelier, qui a choisi de traiter d'une problématique qui ne peut pas être envisagée de façon mono-disciplinaire : les impacts du numérique sur la vie au travail et sur la démocratie dans une organisation.

L'atelier se positionne dans une approche critique au sens de Neumann (2013) : « [Critical approach] *goes beyond surface illusions to reveal underlying structures and conflicts of social relations as a way to empower people to improve the social world* ». En effet, nous souhaitons être utile aux organisations en leur offrant les moyens de comprendre l'impact des Systèmes d'Information Numériques (SIN) sur la démocratie en leur sein, et en leur proposant des leviers d'amélioration.

Pour cela, les missions que nous nous sommes fixées sont les suivantes : analyser, proposer, sensibiliser. L'analyse porte principalement sur l'impact du SIN sur la démocratie dans les organisations. Elle prend sa source dans une bibliographie pluridisciplinaire constituée par tous les chercheurs de l'atelier, par des séminaires avec invités et est formalisée lors de réunions de travail mensuelles nous permettant de construire notre vision commune du sujet traité.

Nos propositions respectent une exigence principale : construire des solutions concrètes et transmissibles hors de l'atelier. Nous avons ainsi proposé une première méthode (ISIDOR) permettant d'évaluer l'impact d'un SIN sur la démocratie dans une organisation, et sommes en début de travail pour la construction d'une méthode de gestion de projet.

La sensibilisation sur la question de l'impact des SIN sur la démocratie des organisations s'adresse à tous les publics : étudiants en informatique ou non, enseignants-chercheurs, grand public, dans le cadre de nos activités d'enseignement ou lors de nos participations à des conférences et ateliers.

2. Citoyen de l'organisation, citoyen du SIN

2.1. La centralité de la question de la citoyenneté dans la démocratie

Il est selon nous indispensable de traiter de la question de la démocratie dans les organisations : « *Du fait que les entreprises emploient des citoyens […] pétris de valeurs démocratiques, l'existence d'une entreprise dans une démocratie pose donc la question de la démocratie dans l'entreprise* » (Salles et al., 2020). Pour autant, la transposition du concept de démocratie dans la société à celui de démocratie dans l'organisation n'est pas aisée. Il demande à redéfinir des dimensions de la démocratie pertinentes (Jardat, 2012), parmi lesquelles la citoyenneté.

L'importance de cette dimension au regard du thème de notre atelier peut être appuyée par cette citation de Ricoeur (1986) : « *la démocratie est le régime dans lequel la participation à la décision est assurée à un nombre toujours plus grand de citoyens* »

2.2. Regards croisés sur la citoyenneté

La question de la citoyenneté s'est rapidement imposée en travaillant sur le champ de la démocratie dans l'organisation. Elle nous a permis de passer d'une vision « informaticienne » prenant en compte l'utilisateur/usager face à son SIN à une vision plus large considérant l'utilisateur avant tout comme un citoyen de son organisation

L'approche pluridisciplinaire sur la question de la citoyenneté nous a amenés à explorer différents questionnements tels que celui portant sur les parties prenantes et constituantes de l'organisation dans le champ de l'économie (Favereau, 2018) ou sur le citoyen éclairé dans le champ du droit. Dans le champ de l'informatique, les apports ont principalement portés sur la question de l'utilisateur acteur de la formation du SIN de l'organisation au travers de méthodes de co-conception ou de co-construction (Halin, 2004, Bour, 2019).

2.3. Construction d'une vision pluridisciplinaire et commune du citoyen

Ces différents apports nous ont permis d'élaborer une vision commune, formalisée bien qu'en perpétuelle évolution, de ce qu'est le citoyen dans l'organisation et face au SIN (Bour et Salles, 2021).

Ainsi selon nous, les salariés sont des parties constituantes de l'organisation (celles sans lesquelles l'organisation n'existe pas). Ces parties constituantes forment le corps des citoyens de l'organisation. En considérant que l'organisation se définit en partie par son projet, les parties constituantes (les citoyens) sont les bâtisseurs de ce projet.

Être citoyen du Système d'Information (SI) signifie donc être partie constituante lors de l'élaboration du projet de l'organisation et de sa représentation dans le SI et être partie constituante des projets informatiques, c'est-à-dire du passage du SI au SIN.

La place du citoyen peut être considérée comme suit au regard des autres dimensions de la démocratie (cf. Fig. 1) :

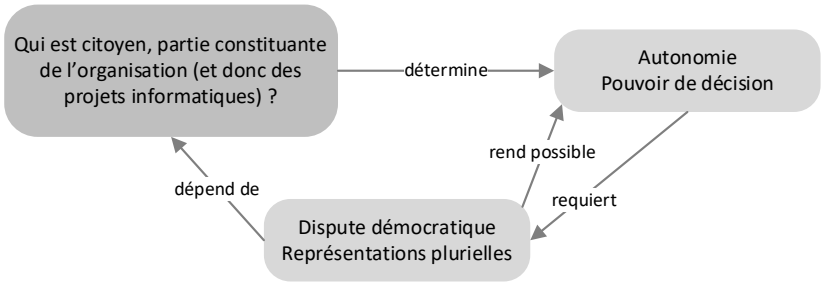

FIG. 1 – Lien entre la dimensions « citoyen » et les autres dimensions de la démocratie dans une organisation

3. Expérimentation de la méthode ISIDOR

Nous avons construit la méthode ISIDOR pour évaluer l'influence d'un SIN sur les différentes dimensions de la démocratie d'une organisation (Salles et al., 2022).

Pour cela, nous avons développé des outils (grilles d'entretien, grilles d'analyse et grilles d'évaluation, structures types), une démarche (phases et étapes) et la structure d'un livrable (rapport). Cela nous permet, pour les trois composants du SIN d'une organisation (des tableaux de bord, un logiciel métier et un projet informatique récent dans l'organisation), de mesurer leur impact sur la place de citoyen, la dispute démocratique, l'existence de représentations plurielles, la prise de décision et l'autonomie.

A ce jour, la méthode a été expérimentée à deux reprises (décembre 2022 – juin 2023, et avril 2023 – décembre 2023). Les expérimentations de la méthode se font dans le cadre d'un projet soutenu par l'ANACT[1], présenté en réponse à son appel « Développer le dialogue social technologique au travail[2] ».

Elles nous permettent, sur des terrains réels de mesurer l'impact du SIN sur le niveau de démocratie dans l'organisation, et de construire ainsi des préconisations adaptées à l'organisation. Les deux expérimentations ont été menées dans deux contextes sensiblement opposés. La première a été menée au service RH d'une grande Université toulousaine, fortement outillé et peu tournée (au départ) ver les questions portant sur la démocratie dans l'organisation. La seconde, quant à elle, s'est déroulée dans une entreprise de fabrication de cadres de vélos d'une vingtaine de salariés, très peu outillée en termes d'informatique et très axée sur les questions de démocratie dans l'entreprise.

[1] Agence Nationale d'Amélioration des Conditions de Travail

[2] www.anact.fr/appel-manifestation-dinteret-developper-le-dialogue-social-technologique-au-travail

Les expérimentations ont confirmé le caractère centrale de la question de la citoyenneté dans la méthode ISIDOR avec des sujets cruciaux tels que :

- La présence (ou non) des utilisateurs/usagers dans les instances des projets informatiques, leur permettant de participer à la dispute démocratique et à la prise de décision sur les sujets fondamentaux d'un projet.
- La garantie d'accès (ou non) aux informations via les outils numériques pour permettre l'exercice plein de la citoyenneté dans l'organisation.

En parallèle des expérimentations d'ISIDOR, l'atelier SI2D développe une méthode de gestion de projet à destination de deux types de publics : informaticiens (ESN ou en interne) et acteurs de l'organisation (dont les futurs utilisateurs) pour permettre de faire de la construction de SIN un levier pour la démocratie dans les organisations.

Références

Bour, R., Soule-Dupuy, C., Vallès-Parlangeau, N., (2019), « *DEMOS: a DEsign Method for demOcratic information System* », dans 13th International Conference on Research Challenges in Information Science (RCIS), Brussels, Belgium.

Bour, R., Salles, M., (2021), « *Ingénierie d'une méthode pour évaluer l'impact d'un SI sur la démocratie d'une organisation* », Actes du congrès Inforsid 2021, Dijon, pp. 87-102.

Favereau, O., (2018), « *Rapport sur les modèles de gouvernance de l'entreprise »*, Évaluation et prospective des modèles actuels, OIT.

Halin, G., (2004), *Modèles et outils pour l'assistance à la conception. Application à la conception architecturale*, HDR, Institut National Polytechnique de Lorraine.

Jardat, R., (2012), « *De la démocratie en entreprise. Quelques résultats empiriques et propositions théoriques* », Revue française de gestion, 2012/9 (N° 228-229), p. 167-184.

Neuman, W. L., (2013), *Social Research Methods: Qualitative and Quantitative Approaches*, seventh edition, Londres (UK), Pearson New International Edition.

Ricoeur, P., (1986), *Du texte à l'action*, Éditions du Seuil, Paris.

Salles, M., Bour, R., Jardat, R., (2020), *« Systèmes d'information numériques : supports ou entraves à la démocratie dans les organisations* ? », Revue Ouverte d'Ingénierie des Systèmes d'Information (ROISI), Numéro spécial Impact des SI sur la démocratie dans les organisations, 20-1, 2.

Salles, M., Bour, R., Colletis, G., Corbion-Condé, L., Fieux, E., Isla, A., (2023), *« Évaluer l'impact d'un système numérique sur le niveau de démocratie d'une organisation »*, Socio-anthropologie, 47, 159-176.

Co-construction d'un outil d'assistance à l'analyse de textes par théorisation ancrée

Margaux Coeuret*, Joris Falip*
Aurélien Bénel*

*LIST3N/Tech-CICO, Université de Troyes, France
margaux.coeuret@utt.fr

Le projet doctoral présenté dans cet article vise à mieux assister le travailleur de la connaissance dans son processus d'analyse qualitative par théorisation ancrée, en plaçant l'humain au cœur du traitement et de l'interprétation des données. Notre question dans ce cadre est : quel type d'automatisation co-construire pour garder l'humain au centre du travail d'interprétation ? Nous commencerons par contextualiser la méthode par théorisation ancrée (GTM), puis exposerons la méthode de terrain employée dans cette recherche, avant de présenter les apports envisagés dans l'élaboration d'un outil d'assistance co-construit soutenant la méthode par théorisation ancrée.

1 Contexte

La méthode par théorisation ancrée (*Grounded Theory Method* ou GTM), formalisée par Glaser et Strauss (1967), vise à élaborer une théorie scientifique à partir des données de terrain, adoptant une approche définie comme inductive et itérative. Utilisée notamment dans l'analyse et la transcription d'entretien en sociologie, la GTM possède des composantes méthodologiques telles que l'échantillonnage théorique, le codage, la méthode de comparaison constante, la saturation théorique et la sensibilité théorique. Cette méthode peut s'inscrire dans une perspective de recherche co-construite (Charmaz, 2014), où la théorie émerge de manière interactive entre les participants et le chercheur. Christophe Lejeune (2017) propose une typologie combinant la GTM à des logiciels semi-automatiques (tels que NVivo, AtlasTI ou MaxQDA) regroupés sous l'appellation "CAQDAS", offrant des fonctionnalités d'étiquetage, de rédaction et de schématisation pour guider efficacement le chercheur dans son analyse. Néanmoins, pour des raisons marketing, des fonctionnalités ont été ajoutées qui s'écartent de la stricte application de la méthode. Se recentrer sur l'outillage de la méthode par théorie ancrée amènera à incorporer des fonctionnalités souvent délaissées tels que le journal de bord, les diagrammes ou un étiquetage réellement ascendant (*bottom-up*).

2 Méthodologie

Pour créer un outil co-construit aligné sur les pratiques de la GTM, notre méthodologie sera la suivante : sélection des participants, recueil des données qualitatives par des méthodes

de terrain, analyse des données et conception d'une contribution informatique. Cette démarche méthodologique sera non linéaire et itérative. Notre étude de terrain s'effectuera auprès des chercheurs et étudiants en sciences sociales de l'Université de Liège, ainsi qu'auprès des chercheurs en *CSCW* de l'Université de Technologie de Troyes. Le recueil des données utilisera diverses techniques, notamment les observations de terrain, les entretiens individuels et les *focus groups*. Ces méthodes permettront d'obtenir des données variées, qui seront comparées dans le but de faire émerger les pratiques et le vécu des utilisateurs d'outils d'analyse qualitative par théorisation ancrée (Mouloudi et al., 2007). Ensuite, nous analyserons ces données collectées et développerons des ateliers de co-conception de *design thinking* pour créer des prototypes fonctionnels et accessibles aux différents profils d'usagers (Caelen et al., 2005). Ces étapes d'élaboration conduiront à une contribution informatique intégrant les apports des techniques d'intelligence artificielle, de lexicométrie ou de visualisation dans l'assistance à l'analyse du corpus d'entretiens.

3 Contributions envisagées

Notre contribution informatique se focalisera sur l'assistance à l'exploration et à l'analyse des données par théorisation ancrée, avec une déclinaison en trois niveaux d'assistance :

— **Assistance à la conduite de l'enquête :** L'objectif principal est de simplifier l'identification des pistes pertinentes pendant le codage du corpus, tout en guidant l'utilisateur dans la décision de clôturer l'enquête. Par la suite, cette assistance permettra d'aiguiller le chercheur dans la sélection des participants.

— **Assistance à la lecture et l'analyse des entretiens :** Ce niveau explorera l'utilisation d'algorithmes de *machine learning* et de lexicométrie pour faciliter l'exploration des données textuelles détecter des tendances, des signaux faibles et faciliter la comparaison des textes.

— **Assistance à l'articulation de cas et de la théorisation :** La contribution finale consistera à soutenir le chercheur dans l'élaboration de son analyse théorique en recourant à des visualisations augmentées (sous forme de graphes ou matrices par exemple). Ces méthodes permettront d'appréhender les gains de la visualisation de données dans le processus d'analyse qualitative et l'émergence théorique (Kuckartz, 2010).

Références

Caelen, J., F. Jambon, et A. Vidal (2005). Conception participative : des moments à leur instrumentation. *Revue d'Interaction Homme-Machine, Vol6 2*.

Charmaz, K. (2014). *Constructing grounded theory*. Sage.

Glaser, B. et A. Strauss (1967). Grounded theory : The discovery of grounded theory. *Sociology the journal of the British sociological association 12*(1), 27–49.

Kuckartz, U. (2010). Innovations dans un logiciel d'analyse qualitative de données : l'intégration d'outils de visualisation. *Recherches Qualitatives 9*, 109–119.

Lejeune, C. (2017). Analyser les contenus, les discours ou les vécus ? À chaque méthode ses logiciels ! In *Les méthodes qualitatives en psychologie et sciences humaines de la santé* (Dunod, Malakoff, France ed.)., Psycho Sup, pp. 203–224.

Mouloudi, A., P. Morizet-Mahoudeaux, C. Lemarchand, et A. Valentin (2007). A proposal of hci model based on user needs, in proceedings of IEEE international human machine interaction conference human. *Timimoun, March*.

Une plateforme d'informatique décisionnelle collaborative

Muhammad Fahad*

* Univ Lyon, Univ Lyon 2, UR ERIC
fahad.muhammad@univ-lyon2.fr
https://eric.univlyon2.fr/bi4people/

Nous présentons dans cet article une plateforme d'informatique décisionnelle collaborative (Collaborative Business Intelligence ou CBI) en libre service, accessible, open source et gratuite, qui met en œuvre le processus d'entreposage de données en mode Software-as-a-Service (SaaS). Notre plateforme de CBI permet l'exploration des données. Ainsi, les utilisateurs découvrent des informations cachées dans les données et présentent leurs découvertes dans des visualisations attrayantes. Elle prend également en charge et assiste le partage d'informations, la prise de décision collaborative et la gestion des annotations au-delà des frontières des individus et des entreprises. Nous concevons un OLAP (OnLine Analytical Processing) collaboratif qui permet d'aider plusieurs collaborateurs à travailler ensemble pour construire une analyse collaborative afin de prendre ensemble de meilleures décisions. Le principal livrable de cette recherche sera un prototype de logiciel prenant en charge toutes les fonctionnalités mentionnées dans le projet BI4People (Darmont et Sdiri, 2022). Il inclut huit tâches numérotées de wp1 à wp8. Notre lot (wp3) se concentre principalement sur l'analyse collaborative. La plateforme de CBI peut fonctionner soit de manière indépendante, soit se « brancher » sur la tâche wp2 « entreposage automatique des données » qui vise à l'intégration de données à partir de plusieurs sources de données hétérogènes dans un entrepôt de données multidimensionnelles. La plateforme de CBI intègre différents types de données de session collaborative (les données d'utilisateur, les données d'annotation, les données de requête, etc.).

La plateforme de CBI démocratise les données hétérogènes en permettant aux utilisateurs d'obtenir facilement et directement des réponses exploitables. Elle stocke également leurs analyses, la structure des tableaux de bord et les graphiques dans une base de connaissances de stockage collaborative. La Figure 1 illustre l'architecture de la plateforme de CBI. Différents utilisateurs (des analystes de données, des scientifiques des données, des utilisateurs novices) peuvent utiliser l'OLAP collaboratif selon leurs besoins et créer une session collaborative pour prendre des décisions. La session collaborative reçoit des contributions de collaborateurs via l'interface graphique d'OLAP. Les entrées peuvent être des annotations de n'importe quel type (une question, une réponse, un commentaire de texte, etc.) et de n'importe quelle forme (une rétroaction générale, une discussion centrée sur le rapport, une analyse des tendances, une coordination des tâches, etc.). En plus de fournir des analyses OLAP collaboratives, notre objectif est de construire un stockage et une récupération efficaces et rapides des informations entre les collaborateurs. Par conséquent, nous visons à incorporer une ontologie dans la plateforme collaborative afin que différents types d'inférences puissent être réalisés sur les données de session. Nous avons conçu des ontologies de session collaboratives (CBIOntology) qui décrivent et conceptualisent formellement les connaissances du domaine et stockent efficacement

les données de session collaboratives. Les données capturées dans CBIOntology deviennent interopérables et traitables par machine pour faciliter le partage des connaissances, avec un vocabulaire commun entre les équipes et les organisations collaboratives indépendantes. Les détails de cet aspect sont élaborés dans une ontologie des données de session collaborative (Fahad M., 2022). Il y a trois composants principaux dans notre plateforme de CBI qui sont responsables du stockage sémantique des données dans la base de connaissances. Tout d'abord, le gestionnaire de session utilisateur (USH) stocke toutes les informations spécifiques à l'utilisateur dans l'ontologie du profil utilisateur. Deuxièmement, le gestionnaire de requêtes d'utilisateur capture toutes les données de requête dans l'ontologie de requête d'utilisateur. Troisièmement, système de gestion des annotations rassemble toutes les annotations et les stocke dans l'ontologie de session de collaboration. Il stocke également l'emplacement et les informations spatiotemporelles sur la collaboration entre les collaborateurs. Ces sous-ontologies font partie de la grande ontologie CBI. Le système de gestion des annotations joue un rôle important dans l'amélioration de l'efficacité et de l'efficience du système CBI, mais il évolue à mesure que le projet grandit avec le temps.

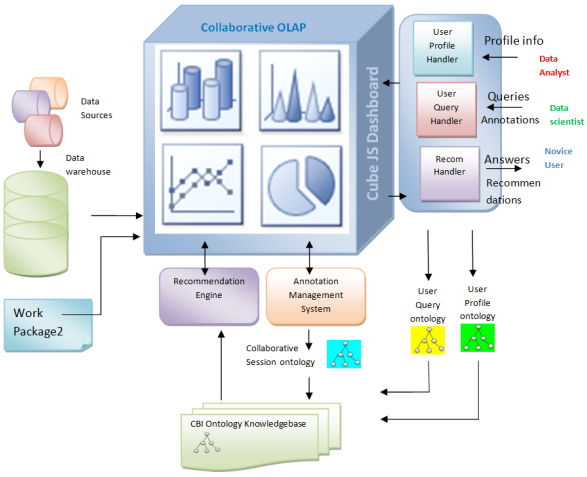

FIG. 1 – *L'architecture de la plateforme de CBI*

Cette article a pour objectif de présenter le projet ANR-19-CE23-0005 « BI4people : Le décisionnel pour tou·tes », financé par l'Agence Nationale de la Recherche (ANR).

Références

Darmont, J. et B. Sdiri (2022). Projet BI4people : Le décisionnel pour tout•es - Saison 2. 18e journées BI et Big Data (EDA 2022), Clermont-Ferrand ; RNTI, Vol. B-18.

Fahad M., Darmont J., F. C. (2022). The collaborative business intelligence ontology (cbiont). 18e journées BI et Big Data (EDA 2022), Clermont-Ferrand ; RNTI, Vol. B-18.

Towards Collaborative Business Intelligence with Conversational Agents: A User-Centric Approach

Olga Cherednichenko* and Muhammad Fahad**

Univ Lyon, Univ Lyon 2, UR ERIC, 5 avenue Mendès France, 69676 Bron Cedex, France
*olga.cherednichenko@univ-lyon2.fr
**fahad.muhammad@univ-lyon2.fr
https://eric.univ-lyon2.fr/bi4people/

Résumé. In the dynamic landscape of contemporary business, the demand for cutting-edge methodologies integrating intelligent technologies to facilitate effective decision-making is on the rise. Business Intelligence systems have become indispensable tools for decision-making across diverse organizations, and the emergence of Collaborative Business Intelligence (CBI) extends the horizons of informed decision-making by tapping into external information from varied sources. The central aim of this research centers on the development of a model tailored for CBI processes within distributed virtual teams. This is achieved through the interaction between users and a CBI Virtual Assistant. The key role of the virtual assistant is to augment the accessibility of data exploration for a wider user base, concurrently minimizing the time and effort required for data analysis. The proposed model, through its integration of intelligent technologies, seeks to propel collaborative decision-making to new heights, emphasizing accessibility, efficiency, and user empowerment.

1 Introduction

The decision-making process is intricately woven, often hinging on the critical information possessed by the decision-maker. In the contemporary world, copious amounts of data are amassed across diverse domains. While these data holds immense potential for informed decision-making, their utility is hindered by disparate systems storing data in varied formats, offering differing levels of access and security. Additionally, the data may suffer from incompleteness, contradictions, and unreliability. The management of vast datasets often falls into the hands of data analysts, and this is where Business Intelligence (BI) becomes indispensable. Business intelligence tools play a pivotal role in extracting valuable insights and aiding in strategic decision-making. These tools sift through historical and current data, presenting results in visually intuitive formats. However, a notable challenge lies in the communication gap between technical specialists, decision-makers, and business process analysts. As underlined in (Cherednichenko et al., 2023), dridging this divide is where Collaborative Business Intelligence (CBI) comes into play. As demonstrated by our analysis, the necessity for data research extends beyond the realms of business profitability or government problem-solving ; it

is equally relevant to society and individual citizens for informed decision-making. However, organizing the collaboration between potential users, decision-makers, and technical specialists in data analysis poses a considerable challenge. The BI4people project [1] has been initiated to address these challenges. Its objective is to democratize Business Intelligence by implementing a data warehousing process in software-as-a-service mode. This spans from integrating data from various sources to intuitive analysis and data visualization. The unique approach of BI4people encourages individuals to contribute their opinions and comments, fostering a collaborative atmosphere and giving rise to what can be termed as general BI or Collaborative BI. Expanding on our previous research, this paper delves into the potential of conversational chatbots in enhancing Collaborative Business Intelligence tools. Building on our earlier work, which introduced a reference model for a virtual collaborative assistant in CBI, comprising conversational, data exploration, and recommendation agents, we now focus on the role of conversational agents in aiding data exploration. This research explores features crucial for converting user queries into executed commands, especially for non-experienced users. Furthermore, the study addresses the interface challenges between conversational and data exploration components, opening avenues for improved collaboration and accessibility within CBI systems.

2 State of the art

The data exploration process systematically scrutinizes and analyzes datasets to elicit insights, identify patterns, and unveil meaningful information (Jebb et al., 2017). It is an iterative and interactive procedure, combining manual exploration with automated analysis techniques. Data exploration assumes a pivotal role in the overarching Business Intelligence (BI) framework, encompassing the examination, scrutiny, and visualization of data to extract significant insights (Çağatay Demiralp et al., 2017). Employing statistical tools and visualizations facilitates the identification of patterns, trends, and relationships within the data, contributing to the formulation of hypotheses and guiding further investigations (Francia et al., 2022). Python, with its comprehensive library ecosystem including NumPy, Seaborn, Matplotlib, and Pandas, provides robust tools for Exploratory Data Analysis (EDA). However, users must possess expertise in data analysis and proficiency in relevant tools and methodologies to effectively manipulate and explore the data.

The current landscape of collaborative business intelligence reflects an expanding field characterized by the continuous development of tools and techniques, all aimed at facilitating effective collaboration within decision-making processes for teams (Muhammad et al., 2022). Notable advancements within the Collaborative Business Intelligence (CBI) domain include the integration of social media functionalities, enhanced accessibility through mobile devices, and the adoption of cloud-based solutions. These innovations empower users to engage in collaborative work, access data from any location, using any device, and at their convenience. Moreover, the integration of Natural Language Processing (NLP) and Machine Learning (ML) simplifies user interactions with data, extracting valuable insights and enhancing decision-making efficiency (Wolf et al., 2019; Otter et al., 2020). An additional benefit lies in the utilization of Natural Language Querying (NLQ), streamlining non-technical users' access to and

1. https ://eric.univ-lyon2.fr/bi4people/index-en.html

analysis of data, thereby enhancing their involvement in decision-making processes (Caldarini et al., 2022; Gamboa-Cruzado et al., 2023).

The incorporation of robotic assistants in e-commerce platforms and web portals with diverse functionalities has become increasingly prevalent (Gamboa-Cruzado et al., 2023). This study seeks to explore the potential advantages offered by conversational chatbots in enhancing the data exploration process and introducing novel prospects within CBI. A chatbot, a computer program designed to simulate human conversation through text or voice-based interactions (citation), offers notable benefits, including round-the-clock availability, expedited response times, personalized interactions, scalability, and cost-effectiveness.

The primary objective of our research centers around developing a model for CBI processes within distributed virtual teams, facilitated through the interaction between users and a CBI Virtual Assistant. The virtual assistant aims to broaden the accessibility of data exploration for a diverse user base while concurrently reducing the time and effort required for data analysis. As a result, the following research questions are posed :

RQ1 : How can pretrained generative language models be effectively utilized for data query construction to support non-expert users ?

RQ2 : What are the essential components of a prompt created for a generative language model to construct data queries effectively ?

To answer the research questions it is needed firstly to investigate and delineate effective methodologies for utilizing pretrained generative language models in the construction of data queries, with a specific focus on enhancing accessibility for non-expert users. This involves assessing the capabilities of generative language models in formulating natural language queries and discerning strategies that cater to the needs of users lacking specialized expertise in data analysis. Then, to identify and define the critical elements integral to the formulation of prompts tailored for generative language models in the context of constructing data queries. This objective involves scrutinizing the components that contribute to the efficacy of prompts, ensuring their ability to guide generative language models in generating relevant and insightful data queries.

By addressing these objectives, the research aims to contribute insights into the practical application of generative language models in the realm of data exploration, particularly in the context of CBI. The overarching goal is to enhance the usage of such models for recommender systems, thereby streamlining and augmenting the data exploration process within non-experts.

3 Materials and methods

Our investigation commences with a focus on data exploration, specifically exploring the advantages of integrating a virtual assistant to aid users in this intricate process. The initial phases involve defining the domain of study, identifying relevant data sources, and consolidating the acquired data. The data loading process entails acquiring and importing pertinent data from diverse sources into the CBI system. Subsequently, the exploration phase is initiated, allowing users to analyze data, discern patterns, and comprehend underlying trends. Metadata, encompassing descriptive information about the data, is pivotal. This includes structural attributes such as format, semantics, and relationships.

User interaction involves formulating queries to extract specific information or perform calculations. Constructing these queries necessitates adept syntactical formulation and the se-

lection of relevant parameters aligned with the desired analytical or informational requisites. Post formulation, queries must be aligned with corresponding commands within the collaborative business intelligence system to ensure accurate interpretation of user intent.

The generation of prompts, integral to user interaction, necessitates an understanding of the user's context, preferences, and analytical objectives. The research question centers on the efficacy of a generative language model in facilitating data exploration for non-expert users. The experimental pursuit seeks to delineate a linguistic framework for prompt construction and the development of a user query reference model.

The selected dataset pertains to traffic accidents in France [2], specifically extracted from the National Interministerial Road Safety Observatory "ONISR." This dataset encompasses information about bodily accidents on public roads, detailing accident characteristics, location, involved vehicles, and victims.

To construct prompts, the GPT 3.5 model [3] in the English language is employed. The focal business process is Exploratory Data Analysis (EDA), comprising common steps applicable across diverse datasets. These EDA steps include Summary Statistics, Univariate Analysis, Bivariate Analysis, and Multivariate Analysis.

Summary Statistics offer a succinct overview of essential dataset features, encompassing central tendency, variability, and distribution. Measures such as Mean, Median, Mode, Variance, Standard Deviation, Range, Skewness, and Kurtosis are employed.

Univariate Analysis scrutinizes individual variables, while Bivariate Analysis explores relationships between two variables. Multivariate Analysis extends this examination to three or more variables, unveiling intricate relationships and patterns. For our experimental endeavors, a deliberate emphasis is placed on univariate and bivariate analyses, predicated on the presumption that the non-expert user is the focal investigator. The intent is to streamline the analysis, making it user-friendly and conducive to the informed inquiry and interpretation by individuals with varying degrees of analytical acumen.

The experiment involves manipulating collocations from EDA steps, generating prompts, user queries, and employing them to interact with the dataset. The generated grades are analyzed to build a prompt framework facilitating non-expert users in initiating EDA.

4 Results

In our experimentation, we focused on utilizing the GPT-3.5 model to generate data exploration prompts for a dataset concerning accidents in France from 2006 to 2015. The dataset contains statistics on accidents categorized by date. Our objective was to understand how effectively we could prompt the model to provide relevant questions for data analysis tasks, given that the user is a non-expert. Our interactions with the model were framed as a task for a data analysis assistant providing questions to a non-expert user about the dataset. The generated prompts included a description of the dataset, the feature to investigate ("trajet"), and a command (e.g., value counts()) to use for analysis. We conducted 114 rounds of experiments, varying feature names, commands, and dataset information, while occasionally introducing specific instructions like "*you cannot name the command or feature*."

2. https ://www.data.gouv.fr
3. https ://platform.openai.com/docs/model-index-for-researchers

We hypothesized that including comprehensive explanations about the dataset, specifying the feature of interest, and providing a command in the prompt would enhance the relevance of the generated questions. The task for generating questions was explicitly defined to guide the model. For example, a typical prompt was : "*You are a data analysis assistant. The user is not an expert. Suggest a question to the data...*"

Our results revealed that, initially, only 30 per cent of the generated answers were deemed relevant to the task. However, upon introducing additional instructions and refining the communication after non-relevant answers, we achieved a notable improvement in the model's response accuracy. Example of a Relevant Answer : "*Question for the data : "Which reasons for travel are most frequently associated with accidents in France from 2006 to 2015 ?" This information may be valuable for understanding the correlation between different reasons for travel and accident frequencies, potentially guiding safety measures or interventions specific to certain types of trips*."

Example of a Non-Relevant Answer : "*Could you please find out the distribution of reasons for travel (feature : "trajet") in the dataset ? I'm interested in understanding how frequently different reasons for travel occur in the reported accidents. By using value counts() on the "trajet" feature, the user can explore the count of each unique value (reason for travel) in the dataset. The result will provide a breakdown of the frequency of each reason for travel, allowing the user to gain insights into the patterns of accidents based on different travel purposes*." This question explicitly names the command and feature, deviating from the experimental protocol. It does not adhere to the constraints specified in the prompt.

Our findings underscore the significance of detailed dataset descriptions and feature names in generating relevant prompts for data exploration with non-expert users. Contrary to our initial hypothesis, the specific command proved less critical. Additionally, it became evident that thoroughly describing the experimental scenario, highlighting limitations, and refining prompts in response to initial non-relevant answers significantly improved the efficacy of the model in generating pertinent questions. We conclude that establishing a well-defined Exploratory Data Analysis (EDA) pipeline in advance and progressively building the dialogue with the user are crucial for successful interaction with generative language models in data exploration scenarios.

5 Conclusion and Discussion

Our investigation into RQ1 centered on the effective utilization of pretrained generative language models to construct data queries tailored for non-expert users. The results of our experiments demonstrated that pretrained models, such as GPT-3.5, can be harnessed to facilitate data query construction for individuals lacking expertise in data analysis. The key factor influencing the success of this utilization lies in providing a dataset description, emphasizing the significance of the feature under investigation, and guiding the model with a clear task formulation. By framing the interaction within the context of a data analysis assistant catering to a non-expert user, we achieved meaningful and relevant prompts that contributed to a more user-friendly data exploration experience.

In addressing RQ2, we delved into the critical components necessary for constructing prompts that effectively guide generative language models in data query creation. Our experiments revealed that the most pivotal elements include a detailed description of the query

and explicit identification of the feature to be analyzed. Surprisingly, the specific command played a less decisive role in comparison to the dataset information and feature name. Additionally, we found that incorporating constraints, such as prohibiting the explicit naming of the command or feature, further enhanced the relevance of the model-generated questions.

In conclusion, our research provides valuable insights into the effective use of pretrained generative language models for supporting non-expert users in data exploration. By understanding the nuanced requirements of constructing prompts, researchers and practitioners can harness the potential of such models to enhance the accessibility and usability of data analysis tools for a broader audience.

6 Acknowledgements

The research study depicted in this paper is funded by the French National Research Agency (ANR), project ANR-19-CE23-0005 BI4people (Business intelligence for the people).

Références

Caldarini, G., S. Jaf, et K. McGarry (2022). A literature survey of recent advances in chatbots. *Information 13*(1), 41.

Cherednichenko, O., F. Muhammad, J. Darmont, et C. Favre (2023). Reference model for collaborative business intelligence virtual assistant. In *CEUR Workshop Proceedings*, Volume Vol-3403, pp. 114–125.

Francia, M., E. Gallinucci, et M. Golfarelli (2022). Cool : A framework for conversational olap. *Inf. Syst. 104*(C).

Gamboa-Cruzado, J. et al. (2023). Use of chatbots in e-commerce : a comprehensive systematic review. *Journal of Theoretical and Applied Information Technology 101*(4).

Jebb, A. T., S. Parrigon, et S. E. Woo (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review 27*(2), 265–276.

Muhammad, F., J. Darmont, et C. Favre (2022). The collaborative business intelligence ontology (cbiont). In *18e journées Business Intelligence et Big Data (EDA-22)*, Volume B-18, Clermont-Ferrand.

Otter, D. W., J. R. Medina, et J. K. Kalita (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 604–624.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et J. Brew (2019). Transformers : State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing*, pp. 38–45.

Çağatay Demiralp, P. J. Haas, S. Parthasarathy, et T. Pedapati (2017). Foresight : Rapid data exploration through guideposts. *ArXiv abs/1709.10513*.

Summary

In the dynamic landscape of contemporary business, the demand for cutting-edge methodologies integrating intelligent technologies to facilitate effective decision-making is on the rise. Business Intelligence systems have become indispensable tools for decision-making across diverse organizations, and the emergence of Collaborative Business Intelligence (CBI) extends the horizons of informed decision-making by tapping into external information from varied sources. The central aim of this research centers on the development of a model tailored for CBI processes within distributed virtual teams. This is achieved through the interaction between users and a CBI Virtual Assistant. The key role of the virtual assistant is to augment the accessibility of data exploration for a wider user base, concurrently minimizing the time and effort required for data analysis. The proposed model, through its integration of intelligent technologies, seeks to propel collaborative decision-making to new heights, emphasizing accessibility, efficiency, and user empowerment.

Un outil de génération automatique de tableaux de bord

Praveen Soni*, Cyril de Runz*, Fatma Bouali* **, Gilles Venturini*

*LIFAT, EA 6300, Université de Tours,
nom.prenom@univ-tours.fr, https://lifat.univ-tours.fr/
** IUT Lille, Université de Lille,
nom.prenom@univ-lille.fr, https://iut.univ-lille.fr/

1 Approche proposée

Nous proposons dans ce résumé un générateur automatique de tableaux de bord et une démonstration d'une première version de ce système. Un tableau de bord (TB) est une interface visuelle multi-vues contenant plusieurs visualisations, éventuellement liées et coordonnées, ayant pour objectif de présenter des données, de surveiller des indicateurs ou encore d'explorer des informations de manière interactive.

Les TBs concernent de nombreux domaines de l'activité humaine et ont fait l'objet de multiples développements et études. Le point spécifique qui nous intéresse est l'aide "automatisée" que l'on peut apporter aux utilisateurs pour construire un TB. Cela consiste à définir des algorithmes de recommandation ou de génération automatique s'adressant en particulier à des utilisateurs novices. On trouve dans la littérature scientifique des approches pour ce problème (en excluant les méthodes qui ne suggèrent qu'une seule visualisation et pas un TB multi-vues) : certaines utilisent des heuristiques simples comme DynSpace (El Meseery et al., 2018), ou des algorithmes d'apprentissage comme par exemple les systèmes DashBot (Da Col et al., 2021) (apprentissage par renforcement, Vizdeck (Key et al., 2012) (arbre de décision) ou DeepEye (Luo et al., 2018) (apprentissage profond). D'autres approches utilisent de l'optimisation sous contraintes comme Draco (Moritz et al., 2018). Egalement, on peut citer des logiciels industriels comme Tableau (Mackinlay et al., 2007), PowerBI ou QlikSense. Toutes ces approches peuvent faire des suggestions de visualisations à l'utilisateur, mais souvent de manière incrémentale (une par une) et sans tenir compte des visualisations déjà présentes dans le TB. Peu d'entre elles prennent en compte le retour utilisateur.

Dans nos travaux, nous avons cherché un algorithme qui puisse faire une optimisation globale des visualisations composant un TB, et ainsi optimiser de manière co-adaptée les visualisations (Soni et al., 2023). Cet algorithme doit aussi tenir compte des compétences de l'utilisateur et de ses objectifs par rapport aux données, et des ses éventuels retours. Pour cela nous proposons une approche utilisant trois modèles (données, utilisateurs, visualisations), une interface et un algorithme génétique. Cet algorithme fait évoluer une population de solutions, chaque solution étant un TB avec un nombre variable de visualisations et un appariement avec les attributs de données. Des opérateurs génétiques permettent de générer des solutions, soit de manière aléatoire (population initiale), soit en modifiant légèrement une solution (mutation

qui ajoute ou supprime une visualisation ou qui vient changer l'appariement avec les attributs de données), ou en combinant deux solutions (croisement qui échange des visualisations entre deux TBs). Trois critères sont utilisés pour évaluer les solutions : la qualité de l'appariement entre les visualisations et les données (il faut trouver les visualisations les plus adaptées aux données), le respect des préférences de l'utilisateur pour certaines données (il faut choisir les attributs de données qui correspondent à ce que veut analyser l'utilisateur), la complexité du TB proposé (il faut tenir compte des compétences de l'utilisateur et sa capacité à comprendre ce que fait l'outil). En ajustant ces trois critères, notre algorithme génétique peut proposer en quelques secondes des TBs à l'utilisateur.

Nous intégrons également dans l'interface graphique une approche semi-interactive dans laquelle l'utilisateur peut, en fonction des suggestions qui lui sont faites par l'algorithme génétique, les accepter en entier ou bien les refuser partiellement ou construire un nouveau TB à partir de morceaux de solution (une ou plusieurs visualisations). Un système de retour utilisateur permet d'aider l'utilisateur à améliorer son usage de l'outil.

Nous avons réalisé une première implémentation de cet outil qui doit encore être améliorée. Nous proposons d'en faire une démonstration lors de notre présentation. Nous utilisons notamment la librairie VegaLite.

Les travaux que nous envisageons ensuite consistent principalement à finir de régler l'optimisation par algorithme génétique (notamment la fonction d'évaluation), puis à réaliser une évaluation utilisateur à plus grande échelle. Nous souhaitons également optimiser l'arrangement 2D des visualisations à l'écran et proposer automatiquement des solutions à l'utilisateur.

Références

Da Col, S., R. Ciucanu, M. Soare, N. Bouarour, et S. Amer-Yahia (2021). Dashbot : An ml-guided dashboard generation system. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 4696–4700.

El Meseery, M., Y. Wu, et W. Stuerzlinger (2018). Multiple workspaces in visual analytics. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, pp. 1–12. IEEE.

Key, A., B. Howe, D. Perry, et C. Aragon (2012). Vizdeck : self-organizing dashboards for visual analytics. In *Proceedings of the 2012 acm sigmod international conference on management of data*, pp. 681–684.

Luo, Y., X. Qin, N. Tang, et G. Li (2018). Deepeye : Towards automatic data visualization. In *2018 IEEE 34th international conference on data engineering (ICDE)*, pp. 101–112. IEEE.

Mackinlay, J., P. Hanrahan, et C. Stolte (2007). Show me : Automatic presentation for visual analysis. *IEEE transactions on visualization and computer graphics 13*(6), 1137–1144.

Moritz, D., C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, et J. Heer (2018). Formalizing visualization design knowledge as constraints : Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics 25*(1), 438–448.

Soni, P., C. De Runz, F. Bouali, et G. Venturini (2023). A genetic algorithm for automatic dashboard generation : first results. In *2023 27th International Conference Information Visualisation (IV)*, pp. 77–82. IEEE.