Afrique SCIENCE *16(6) (2020) 126 - 134*
ISSN 1813-548X, http://www.afriquescience.net

Optimisation de la qualité des données appliquée au domaine du marketing

Chiatsè Mireille Patricia ANASSIN[1*], Boko AKA[1] et Michel BABRI[2]

[1] *Laboratoire de Mathématiques et Informatique, Université Nangui Abrogoua, 02 BP 801 Abidjan, Côte d'Ivoire*
[2] *Laboratoire de recherche en informatique et télécommunication, Institut National Polytechnique, BP 1093 Yamoussoukro, Côte d'Ivoire*

* Correspondance, courriel : *tresichia@gmail.com*

Résumé

Cet article présente un nouveau cadre d'optimisation pour le problème de la qualité des données. Ce cadre combine deux approches. D'une part l'interaction des facteurs de la qualité et d'autre part la personnalisation des analyses. Les requêtes utilisateurs et les propriétés de l'entrepôt sont mises en évidence. Pour résoudre ce problème, nous considérons la théorie de l'incertitude. Il s'agit de faire une étude comparative du nouveau cadre d'optimisation avec la théorie de l'incertitude et proposer des tests sur une application réelle d'entrepôt de données provenant du domaine marketing. L'étude comparative indique que le nouveau cadre proposé est le meilleur en termes de temps d'exécution, de précision, de confiance et de coût. Ces résultats fournissent une très bonne performance aux processus d'entrepôts de données.

Mots-clés : *qualité des données, fraicheur, complétude, exactitude.*

Abstract

Optimization of data quality applicated to the field of marketing

This article introduces a new optimization framework for the data quality issue. This framework combines two approaches. On the one hand the interaction of the quality factors and on the other hand the personalization of analyzes. User queries and warehouse properties are highlighted. For this issue, we consider the theory of uncertainty. We have to do a comparative study of the new evaluation framework with the three baselines propose some tests on a real data warehouse application. The comparative study indicates that the proposed new framework is the best in terms of execution time, accuracy, confidence and cost. These results provide a very good performance for data warehouse processes.

Keywords : *data quality, user queries, freshness, precision, completeness.*

1. Introduction

L'optimisation de la qualité des données [1] a été reconnue comme un problème fondamental dans les systèmes d'information. De nombreux travaux [1 - 5] dans les domaines de la conception du système d'information et du génie logiciel traitent de la qualité des données. Dans un contexte d'entrepôt de données

fournissant un accès à de grandes quantités de données, la qualité est devenue une propriété de première classe de plus en plus exigée par les utilisateurs finaux. Certaines enquêtes et études empiriques [3 - 7] ont montré l'importance de la qualité des données pour les utilisateurs finaux, en particulier, lorsqu'ils traitent avec des données hétérogènes provenant de sources autonomes. En outre, les problèmes de qualité de l'information ont été signalés comme critique [7, 8] dans plusieurs domaines scientifiques et sociaux tels que l'environnement, la biologie, la génétique, le commerce, l'économie et l'informatique dans le Web. En dépit du grand nombre de travaux de recherche [1 - 9] traitant de la qualité des données, plusieurs problèmes restent encore à résoudre. Une analyse en matière d'optimisation a révélé que certains problèmes techniques ne sont pas suffisamment traités [10]. En particulier, nous soulignons : la définition des attentes de l'utilisateur, des propriétés de l'entrepôt, l'analyse des sources et l'interaction des facteurs de qualité. Cet article porte sur ces problèmes et propose un cadre pour l'optimisation de la qualité des données dans les entrepôts de données. Il convient à analyser trois principaux facteurs de qualité : la fraîcheur, la précision et la complétude des données et de les mettre en évidence dans un processus de personnalisation des entrepôts de données. Une taxonomie des techniques de mesure des différents facteurs sur la base de la nature des données, les techniques architecturales et les politiques de synchronisation a été proposée pour les différents facteurs permettant d'identifier les propriétés qui influent sur le processus d'optimisation. L'article est organisé comme suit : la section 2 présente l'état de l'art ; la section 3 intitulée présentation de l'approche, met en relief le modèle proposée, quant à la section 4, elle fait une comparaison de notre approches aux autres ; enfin la section 5 propose une conclusion et les perspectives.

2. Présentation de la problématique

Cette section présente les techniques utilisées pour l'optimisation de la qualité des données. Trois techniques plus utilisées de la littérature ont été présenté : la théorie des jeux, la théorie de possibilité et les techniques de matérialisation de vue. A la fin une étude comparative a été faite suivie d'une discussion sur les problèmes de recherche ouverte.

2-1. La théorie des jeux (TJ)

De nombreuses heuristiques pour résoudre le problème de la sélection de vue matérialisée, basées sur la théorie des jeux, ont été représentées dans la littérature [11]. Certaines de ces méthodes sont basées sur des algorithmes déterministes et produisent des réponses optimisées [12]. D'autres méthodes ont exploité les techniques pour répondre dans un temps d'exécution plus court que la stratégie déterministe [13]. Il existe une troisième catégorie de méthode de sélection de vue matérialisée [14] qui combinent des algorithmes et les deux techniques mentionnés pour résoudre le problème dans un délai raisonnable avec une grande précision. Plusieurs algorithmes déterministes sont utilisés pour résoudre le problème de sélection de vue matérialisée. Beaucoup d'entre eux utilisent des stratégies gourmandes pour résoudre le problème. L'algorithme glouton de [15] est utilisé pour la structure MVPP (multiple views processing plan) afin de sélectionner un ensemble optimal de vues. L'algorithme a minimisé le coût du traitement des requêtes et le coût de la maintenance de la vue, et n'a considéré aucune limitation. Certains auteurs [16] ont fourni un cadre pour optimiser l'ensemble des vues matérialisées. En effet, ils définissent deux concepts contradictoires dans le problème de sélection de vue matérialisée, à savoir le coût de traitement des requêtes comme un joueur et les coûts de maintenance des vues pour l'autre joueur. L'augmentation du coût de traitement des requêtes d'une vue entraîne une réduction du coût de maintenance et vice versa. Ils définissent par ailleurs une fonction de gain : Pour le premier joueur, la fonction de gain est définie comme suit : la différence entre le coût de maintenance de la vue et le coût de traitement des requêtes d'une vue, et pour le second joueur au contraire,

la fonction de gain est définie comme la différence entre les coûts du traitement des requêtes et le coût d'affichage de la maintenance. Les auteurs [13 - 16] se sont basés sur une approche gourmande TSGV (theory strategy game views). Le principal avantage de TSGV est de trouver des vues appropriées pour la matérialisation en une seule analyse des nœuds du MVPP. En effet, l'évaluation des nœuds se fait comme suit : depuis la racine les nœuds précédents sont matérialisés avec les nœuds antérieurs, afin de supprimer les nœuds redondants et d'examiner plus facilement les nouveaux nœuds. Par conséquent, la résolution du problème dans un temps acceptable n'a pas été abordée. Toutefois, la minimisation du coût total de la matérialisation est considérée comme le but principal du problème de sélection de vue matérialisée. Comme un entrepôt de données peut être très important dans la pratique, la sélection de l'ensemble des vues appropriées dans l'ensemble des vues candidates peut être aussi une recherche très longue dans un énorme espace de recherche. Par conséquent, de nombreux algorithmes de sélection de vue matérialisés récents vont être exploités pour résoudre ce problème.

2-2. La théorie des possibilités (TP)

La théorie des possibilités vise à capturer l'incertitude sur la base des informations disponibles, elle est toujours limitée pour fournir une analyse extrême de cette information [17]. Par conséquent, il déploie deux mesures de confiance [8] : l'une qui couvre la possibilité d'événements et l'autre qui couvre la certitude des événements. En conséquence, face à l'incertitude concernant les prédicats dans P, les propriétés relatives à la complexité informatique ne sont plus garanties. Ainsi le problème de maintenance de l'équilibre entre la modélisation de l'incertitude avec une précision suffisante et une complexité de calcul suffisamment faible reste à résoudre [14]. L'analyse montre que l'évaluation des prédicats est un problème difficile [8]. Outre les informations sur le niveau de qualité, est induite en considérant l'incertitude impliquée lors de la mesure [10]. L'une des propriétés essentielle est que l'application d'une mesure sur une grande collection de données peut être représentée efficacement dans le sens où (i) la représentation a une faible complexité de stockage et (ii) elle peut être mise à jour de manière incrémentielle lorsque de nouvelles données sont observés [12]. La quantité de données qui atteint un certain niveau de qualité peut être résumé comme une distribution de possibilité sur l'ensemble des nombres naturels [9]. Par ailleurs, la quantification numérique de la qualité des données peut être fondée sur la théorie de la mesure [10]. La vision moderne de la qualité des données a été installée par [7, 11]. Dans leurs travaux respectifs, ils ont observé que la «qualité» est une propriété complexe de données qui ne peuvent être mesurées directement. Au lieu de cela, on devrait considérer les différentes facettes (c'est-à-dire les dimensions) de la qualité qui sont pertinentes pour une application spécifique et développer des procédures de mesure pour ces dimensions en conséquence. Cette compréhension a conduit à plusieurs approches pour évaluer les dimensions de la qualité des données [8 - 12]. Dans [3], un cadre général a été proposé dans lequel la qualité des données est essentiellement mesurée sur une échelle ordinale, mais l'incertitude sur le niveau de qualité est explicitement prise en compte. Ce cadre permet de combiner des approches basées sur des règles (par exemple, pour des raisons de cohérence et d'exhaustivité) avec des approches basées sur l'incertitude (par exemple, pour les devises). En général, cela signifie que la qualité des données est modélisée comme une distribution des incertitudes sur un ensemble de niveaux de qualité totalement ordonnés. Le problème de l'agrégation des dimensions de la qualité reste à résoudre.

2-3. Les méthodes de traitement des requêtes (TM)

L'analyse des données massives joue un rôle important dans les entreprises axées sur les données [3]. Cependant, de nombreuses applications peuvent tolérer un certain degré d'imprécision, en particulier pour les requêtes exploratoires, où les premiers résultats approximatifs, suffisamment précis, ont souvent une valeur beaucoup plus grande pour les utilisateurs [13]. La théorie mathématique [14] de base qui sous-tend ces recherches peut être divisée en deux grandes catégories : l'agrégation en ligne (OLA) basée sur le

théorème central limite (CLT) (1); et les méthodes bootstrap (2). Les deux méthodes permettent à l'utilisateur d'observer la progression d'une requête donnée en affichant des réponses approximatives affinées, puis d'arrêter l'exécution de la requête, une fois que le résultat atteint la précision souhaitée (l'intervalle de confiance est réduit en fonction de la taille croissante de l'échantillon). Notez que l'OLA basée sur CLT peut calculer les résultats approximatifs si la taille de l'échantillon est "suffisamment grande" avec une bonne qualité, et subit un échec d'estimation dans le cas contraire. Un modèle probabiliste est dérivé pour prédire l'échec de l'estimation possible de l'OLA basé sur le CLT, pour évaluer la taille de l'échantillon attendu et prendre en charge des requêtes d'agrégation arbitraires car les estimateurs d'échantillonnage de formes fermées ne sont pas nécessaires. De plus, les deux méthodes souffrent d'éventuelles erreurs d'estimation. Pour l'OLA basée sur CLT, on suppose que les échantillons doivent être non biaisés. Enfin, la méthode d'analyse complète [13], montre qu'il y a une forte probabilité d'échec de l'estimation pour les méthodes OLA et bootstrap basées sur CLT, nous adoptons la méthode pour renvoyer des résultats précis aux utilisateurs. Les auteurs [17] notent que la solution pour l'opérateur algébrique - n'est pas seulement liée au taux d'erreur prédéfini par l'utilisateur, mais est également liée aux résultats approximatifs des sous-requêtes Les résultats [10] indiquent que la configuration des paramètres d'estimation de chaque sous-requête n'a rien à voir avec les résultats de sous-requête approximatifs, mais uniquement avec le taux d'erreur prédéfini. L''OLA basée sur la CLT [13] vise à donner une réponse plus rapide aux requêtes d'agrégation à grande échelle avec une estimation statistiquement valide du résultat final. L'idée de base d'OLA est de calculer un résultat approximatif par rapport aux échantillons aléatoires non biaisés basés sur le CLT et d'affiner les résultats. Malheureusement, il n'existe pas de technologie capable de prendre en charge un plus grand nombre de catégories de requêtes dans un temps acceptable, en tenant compte de l'échec de l'estimation, la performance d'exécution et la fiabilité.

2-4. Comparaison des approches existantes

Pour des tailles de données spécifiques, l'échantillonnage aléatoire uniforme peut ne pas fonctionner correctement. Les techniques d'évaluation avec le même taux d'échantillonnage pour différentes distributions ou opérations peuvent entraîner une précision différente des réponses aux requêtes. Les différentes techniques sont illustrées dans la ***Figure 1***. 100 Go de données (plus de 500 graphiques et 250 nœuds) est illustrés pour chaque type de technique pour évaluer la performance des échantillonnages aléatoires et échantillonnage stratifié en comparant la moyenne des échantillons et des données complètes. Les métriques utilisées sont la fraicheur et la précision [14]. Le résultat est obtenu à partir de plusieurs exécutions.

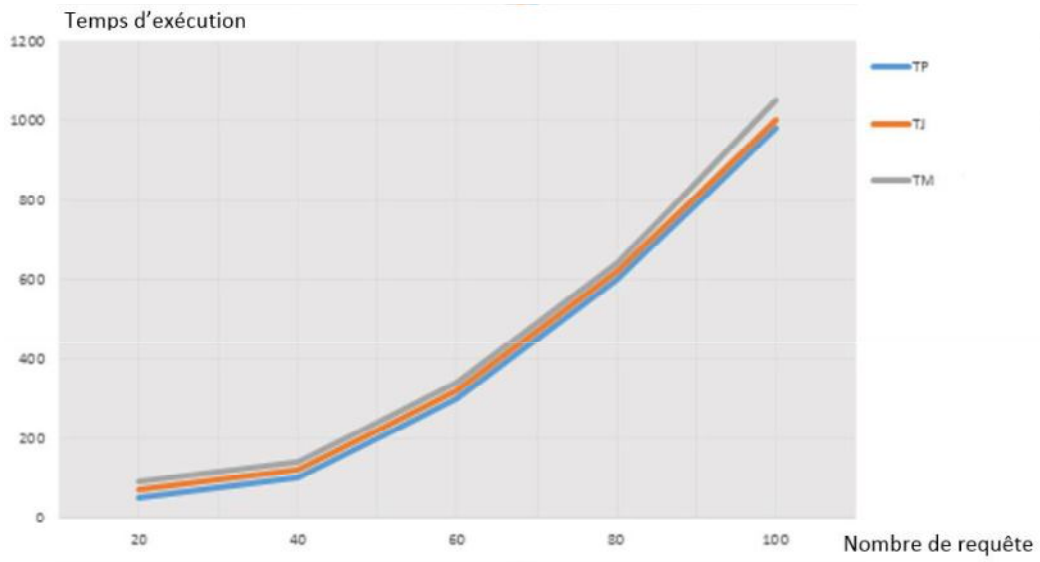

Figure 1 : *Comparaison des trois techniques d'optimisation*

3. Méthodologie

Le cadre d'optimisation de la qualité formalise les différents éléments impliqués dans l'évaluation de la qualité des données. Parmi ces éléments, il y a des sources de données, les cibles de données, le profil utilisateur, les processus ETC, les caractéristiques de l'entrepôt, des mesures de qualité et des algorithmes. Nous commençons à définir le cadre et ses composantes.

3-1. Cadre d'optimisation de la qualité noté MO

Le cadre est défini sous la forme d'une application linéaire qui prend en compte nos paramètres d'optimisation et l'intégrale de Choquet :

$$Q_f : Q_{int} \rightarrow Q_{out} \qquad Q_f = \sum_{i=1}^{3} HC_i \tag{1}$$

avec, Q_f : Fonction de qualité ; Q_{int} : Qualité en entrée ; Q_{out} : Qualité en sortie ; C_i : Facteurs de qualité (exactitude, fraicheur et complétude) ; H : l'opérateur d'agrégation

Pour l'optimisation de notre cadre nous définissons les notions sur lesquelles nous sommes basés pour faire les simulations.

3-2. Erreur liée et confiance

Un sous-groupe g satisfait à l'erreur liée $\epsilon_0 > 0$ et à la confiance $\varepsilon \in [0,1]$ si et seulement si parmi les sources indépendantes M fournies par le sous-groupe, au moins l'estimation $M.\theta$ satisfait $\varepsilon_0 \geq \epsilon_i$ où ε_i est l'erreur de la i-ème source. Lorsque la confiance est faible, la réponse est presque sans valeur car elle risque une forte probabilité. Ainsi, pour rendre le résultat plus robuste, la confiance devrait être limitée par une constante et proche à 1. De cette façon, nous pouvons utiliser l'erreur liée à remplacer le tuple d'erreur de confiance, si les consommateurs ne spécifient pas la confiance. Nous pouvons utiliser la variance v pour exprimer le tuple de confiance et d'erreur comme suit :

$$v = \varepsilon^2(1-\theta) \tag{2}$$

Étant donné que la confiance par défaut est une constante, nous pouvons utiliser la variance pour remplacer la limite d'erreur.

3-3. Graphique de qualité

Un graphique de qualité est un couple $G = (V, E)$ Où : - V est l'ensemble des nœuds. V_s, V_t et V_a sont les ensembles des noeuds de la source, de la cible et de l'activité; avec$V = V_s \cup V_t \cup V_a$. Chaque noeud source ou cible correspond à une source ou cible du cadre.

3-4. Chemin de fraîcheur

Etant donné un trajet de données dans un graphique de qualité [A_0, A_1, ... A_n], en commençant à un A_0 nœud source, le chemin de fraîcheur est la valeur réelle de la fraîcheur se propageant le long du chemin (sans tenir compte d'autres nœuds du graphe): il correspond à la somme des données de la source de fraîcheur réelle du nœud source, des coûts de traitement des nœuds dans le chemin et des retards interprocessus entre les nœuds :

Chemin de fraicheur ($[A_0,\ldots A_p]$) $=$ Souce de fraicheur réelle (A_0) $+ \Sigma_{x=0..p}$ cout des processus (A_x) $+ \Sigma_{x=1..p}$ délai inter processus (A_{x-1},A_x)

Le chemin critique pour une activité donnée est le chemin de données (à partir d'un nœud source à l'activité) qui détermine la fraîcheur de l'activité, à savoir la valeur réelle de la fraîcheur pour l'activité est égale à la trajectoire de fraîcheur. En d'autres termes, si l'on ignore les autres noeuds et nous calculons la fraîcheur des données en utilisant uniquement le chemin critique, on obtient la même valeur de fraîcheur.

3-5. Sélectivité de précision

Compte tenu d'une requête de réécriture QR sur un ensemble de zones $\{R_1 \ldots R_k\}$, la sélectivité de QR, notée *sel*(QR), est la proportion de tuples du produit cartésien des zones $R_1 \ldots R_k$ qui vérifient les prédicats de réécriture. Si la réécriture de la requête n'a pas de sélection (ni jointure) de prédicats, puis tous les tuples sont retournées et donc la sélectivité est 1. Le nombre de tuples de la réécriture est estimé en multipliant le nombre de lignes de zones d'entrée par la sélectivité de la réécriture.

3-6. Distribution de Dirichlet pour la complétude

Soient $\alpha_1, \alpha_2, \ldots, \alpha_r$ tous strictement positifs. Une distribution de Dirichlet d'exposants $\alpha_1, \alpha_2, \ldots, \alpha_r$ (également appelés les hyper paramètres pour ne pas les confondre avec les paramètres θ_i) possède une densité de probabilité de la forme suivante :

$$Dir(\theta|\alpha_1, \alpha_2, \ldots, \alpha_r) = \frac{\gamma(\sum_{i=1}^{r} \alpha_i)}{\prod_{i=1}^{r} \gamma(\alpha_i)} \prod_{i=1}^{r} \theta^{\alpha_i - 1} \quad (3)$$

où, γ est la fonction Gamma, qui satisfait $\gamma(x+1) = x\gamma(x)$ et $\gamma(1) = 1$. Pour l'apprentissage des paramètres de réseaux bayésiens, nous allons faire les hypothèses suivantes

3-7. Personnalisation

Un profil utilisateur est un ensemble de prédicats misent à jour traduisant l'intérêt de l'utilisateur. Pour un utilisateur U le profil est définie par le couple $P_{rof} = (pred_U, \theta)$ où $pred_U$ est une disjonction de prédicats qui spécifie la condition des attentes de U et θ est un nombre réel qui indique le degré d'intérêt de l'utilisateur U sur les données générées par $pred_U$. $\theta \geq 1$. Ce degré d'intérêt est défini par les facteurs de qualité défini.

3-8. Paramètres de simulation

L'ensemble des données des entrepôts de ventes pour le cas adventureworks 2016 [7] est utilisé tout au long de l'expérience. Il y a plus de 2 millions d'enregistrements et 500 champs. Les algorithmes d'évaluation sont exécutés sur différentes tables dont la taille des données est comprise entre 20 et 100 Go. Le nombre de requêtes effectué est compris entre 20 et 100. L'erreur de précision est comprise entre 0 et 1. Le cout total est calculé comme étant la somme des couts de maintenance et de processus. Le taux de confiance (complétude) est compris entre 10 % et 100 %. Les métriques utilisées sont ceux définit. La fonction d'évaluation doit être en diminution exponentielle pour compenser la perte de précision.

3-9. Ligne de base

La fonction d'évaluation est comparée à la théorie de jeux (TJ), la théorie de possibilité (TP) et la théorie de matérialisation (TM) (voir la section 3). Il s'agit d'exécuter des requêtes en ajoutant un bruit à la réponse afin de renvoyer les réponses en tenant compte du profil utilisateur. Cette comparaison est évaluée pour des erreurs limites, un faible taux de complétude et un cout total élevé. Nous avons mené des expériences qualitatives pour illustrer la performance de notre cadre.

4. Résultats et discussion

4-1. Variation des facteurs de qualité

Dans cette simulation, nous évaluons nos facteurs de qualité sur des ensembles de données (contenant un nombre de graphique et de nœud). Comme le temps d'exécution n'est pas constant pour une même requête, alors le cout total est croissant. La précision de la réponse aux requêtes est donnée par le bruit et les préférences de l'utilisateur. Donc la perte de précision fournie des erreurs spécifiques ***(Figure 2)***. Cette ***Figure*** illustre également que le temps d'exécution de la requête avec différents limites d'erreurs. Le cadre indique que le temps décroit peu importe la façon dont l'erreur est liée aux changements. Puisque le jeu des données diminue de façon exponentielle avec la croissance des limites d'erreur.

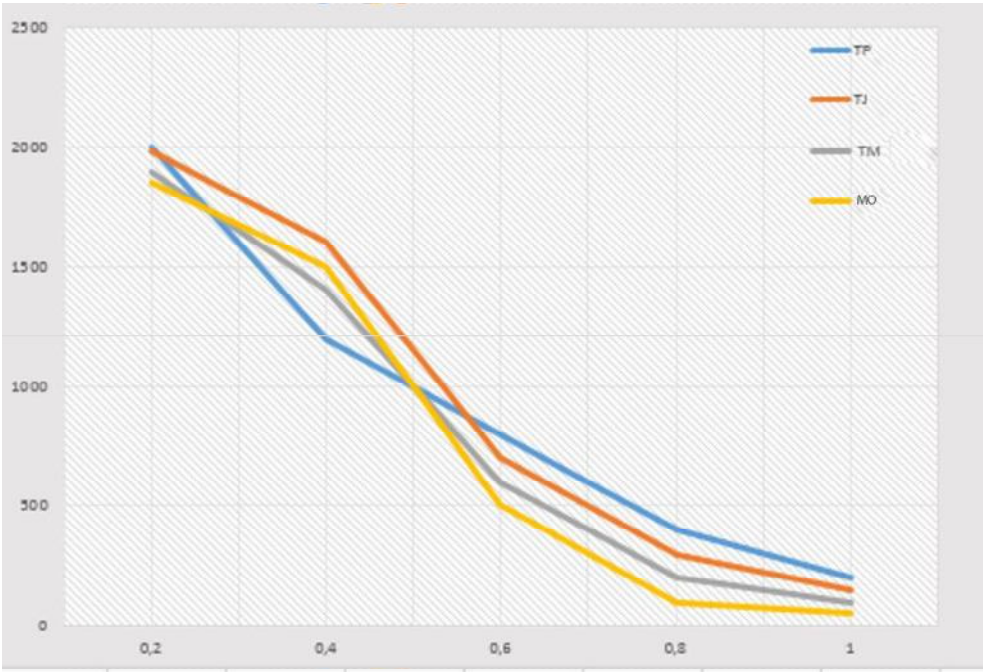

Figure 2 : *Observation de la perte de précision sur les différentes méthodes*

Nous observons que le cadre prend en compte les couts supplémentaires. Le résultat des requêtes sur le jeu de données échantillonné peut ne jamais atteindre la limite d'erreur car l'erreur est toujours liée. Ainsi le taux de complétude devient faible ***(Figure 3)***, ce qui est indépendant de l'écart maximum ou minimum observé.

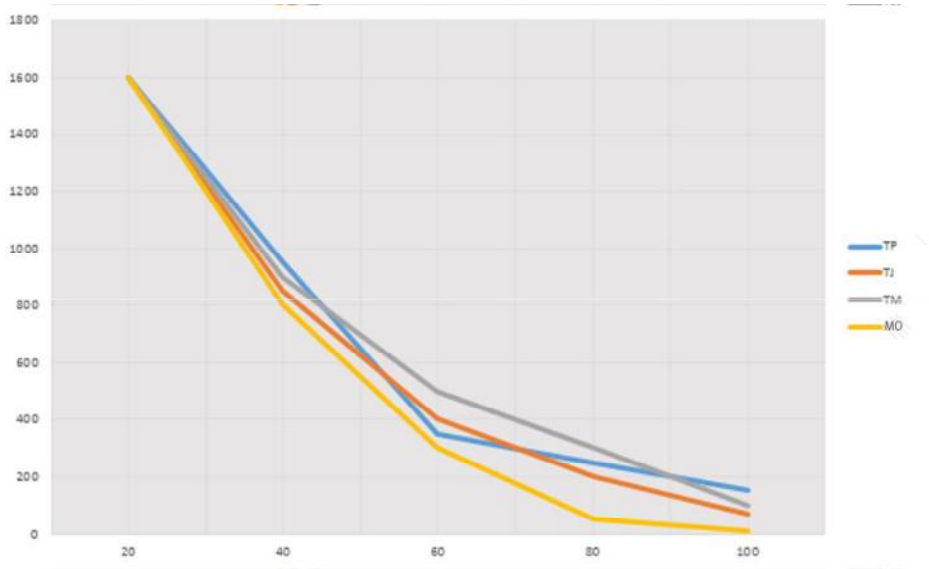

Figure 3 : *Observation du taux d'incomplétude sur les différentes méthodes*

4-2. Test de limitations

Les ensembles de données volumineux (nombre croissant de graphique avec différentes topologies) pris en charge simultanément par le cadre, permet de déterminer les limites du cadre. Toutefois le cadre indique que le temps décroit peu importe la façon dont l'erreur est liée aux chargements ***(Figure 4)***.

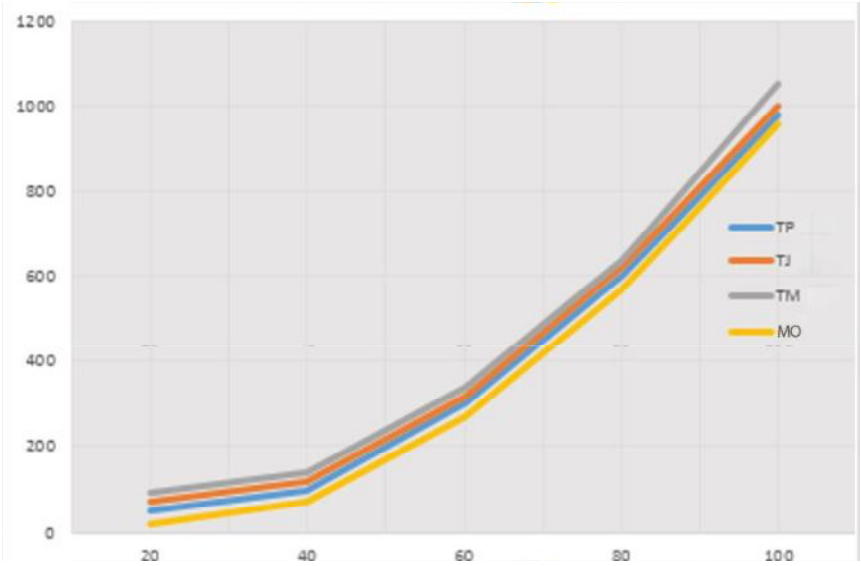

Figure 4 : *Temps d'exécution (ms) sur les différentes méthodes en fonction du nombre de requêtes*

5. Conclusion

Dans cet article, nous avons présenté nos expérimentations avec les facteurs de qualité. Nous avons décrit le prototype, qui gère le cadre d'optimisation de la qualité proposée. Les modèles du cadre, à savoir, les sources de données, les cibles de données, des graphiques de qualité, des propriétés, des mesures, le profil utilisateur et des algorithmes d'optimisation de la qualité. Nous avons utilisé le cadre d'optimisation dans un scénario d'entrepôt de données afin de valider notre approche. Plus précisément, nous avons comparé notre cadre à

trois théories : (i) une théorie des jeux, (ii) une théorie des possibilités et (iii) une théorie de matérialisation de vues. Nous avons montré comment les approches peuvent être modélisées sous forme de graphiques de qualité et comment l'optimisation des algorithmes a été instanciée pour eux. Cette expérimentation permet de valider l'approche dans des applications réelles.

Références

[1] - MIREILLE CHIATSÈ ANASSIN, BOKO AKA and MARCELLIN KONAN BROU, 'Data Extraction in the Warehouse : A Quality-Basesd Approach', *European Journal of Scientific Research*, Vol. 150, N°1 (August 2018) 14 - 20

[2] - I. ALBERT, Visualizing sequencing data quality. In: *The Biostar Handbook: A Beginner's Guide to Bioinformatics*. Biostar, 1st edition. https://www.biostarhandbook.com/, (2016)

[3] - VERONIKA PERALTA, Data Quality Evaluation in Data Integration Systems. PHD thesis, Université de Versailles Saint-Quentin-en-Yvelines (France) and Universidad de la Republica (Uruguay)

[4] - YUXIANG WANG, YIXING XIA, QIMING FANG and XIAOLIANG XU, 'AQP++: a hybrid approximate query processing framework forgeneralized aggregation queries', *Journal of Computational Science*

[5] - L. BELLATRECHE, A. CUZZOCREA, S. BENKRID, Query optimization over parallel relational data warehouses in distributed environments by simultaneous fragmentation and allocation. In: Proc. ICA3PP, (2010) 124 - 135 p.

[6] - S. AGARWAL, B. MOZAFARI, A. PANDA, H. MILNER, S. MADDEN, I. STOICA, BlinkDB: queries with bounded errors and bounded response times on very large data. In : Proc. Eurosys, (2013) 29 - 42 p.

[7] - W. LU, D. MCFARLANE, V. GIANNIKAS, Q. ZHANG, An algorithm for dynamic order-picking in warehouse operations. *European Journal of Operational Research*, 248 (1) (2016) 107 - 122

[8] - E. BOTTANI, R. MONTANARI, M. RINALDI & G. VIGNALI, Intelligent Algorithms for Warehouse Management. In C. Kahraman, & S. Çevik Onar, Intelligent Techniques in Engineering Management : Theory and Applications, Switzerland : *Springer International Publishing*, (2015) 645 - 667 p.

[9] - HOSSEIN AZGOMI, MOHAMMAD KARIM SOHRABI, A game theory based framework for materialized view selection in data warehouse. *Engineering Applications of Artificial Intelligence*, 71 (2018) 125 - 137

[10] - J. B. CLEMPNER, A. S. POZNYAK, Modeling the multi-traffic signal-control synchronization: A Markov chains game theory approach. *Eng. Appl. Artif. Intell.*, 43 (2015) 147 - 156

[11] - M. ARAB, M. K. SOHRABI, Proposing a new clustering method to detect phishing websites. Turk. J. Electr. Eng. Comput. Sci. http://dx.doi.org/10.3906/elk-1612-279, (2017)

[12] - S. BABU, U. MOHAN, An integrated approach to evaluating sustainability in supply chains using evolutionary game theory. Comput. Oper. Res., (2017)

[13] - M. FICCO, C. ESPOSITO, F. PALMIERI, A. CASTIGLIONE, A coral-reefs and gametheory-based approach for optimizing elastic cloud resource allocation. Future Gener. Comput. Syst., (2016)

[14] - A. BRONSELAER∗, J. NIELANDT and G. DETRÉ, "An incremental approach for data quality measurement with insufficient information", ELSEVIER, *International Journal of Approximate Reasoning*, 96 (2018) 95 - 111

[15] - A. BRONSELAER, R. DE MOL, G. DE TRÉ, A measure-theoretic foundation for data quality, IEEE Trans. Fuzzy Syst., https://doi .org /10 .1109 /TFUZZ .2017.2686807, Published Online, (2017)

[16] - H. SCHAEBEN (Eds.), Freiberg, Germany, 1273 - 1278. ISBN 978-3-00-050337-5. Lisowski, P.; Krawczyk, A.; Porzycka-Strzelczyk, S. Possibilities of 3D data storage in spatial databases. ZN Pol. Sl. Studia Informatica. Gliwice, 35 (116) (2014)

[17] - F. CHEN, H. WANG, Y. XIE & C. QI, An ACO-based online routing method for multiple order pickers with congestion consideration in warehouse. *Journal of Intelligent Manufacturing*, 27 (2) (2016) 389 - 408. DOI:10.1007/s10845-014-0871-1