Exploration de données massives à l'aide d'estimations de cardinalités

Pierre Nerzic, Grégory Smits, Olivier Pivert, Marie-Jeanne Lesot

Exploration de données massives à l'aide d'estimations de cardinalités

Pierre Nerzic[1] Grégory Smits[1] Olivier Pivert[1] Marie-Jeanne Lesot[2]
IRISA - Université de Rennes 1, UMR 6074, Lannion, France, {pierre.nerzic,gregory.smits,olivier.pivert}@irisa.fr
Sorbonne Université CNRS, LIP6, F-75005 Paris, France, marie-jeanne.lesot@lip6.fr

Résumé :
Cet article présente un outil d'exploration interactive de données massives stockées dans un système de gestion de base de données (SGBD), nommé FuzViz. Il repose sur une méthode de construction automatique de résumés linguistiques, qui fournissent une synthèse concise et intelligible du contenu des données. Il offre une vue interactive de ces résumés recalculée dynamiquement selon les demandes de l'utilisateur. Pour assurer une exploration fluide des propriétés décrites par ces résumés, FuzViz s'appuie sur la proposition d'une méthode efficace d'estimations de leurs cardinalités, produites à partir des statistiques sur la distribution des données gérées par le SGBD et consolidées par une approche basée sur un échantillonnage. L'outil propose de plus un mécanisme d'inférence de vocabulaire flou à partir de ces statistiques.

Mots-clés :
Résumés linguistiques, inférence de vocabulaire, estimation de cardinalités, données massives.

Abstract:
This paper describes FuzViz, a tool to explore interactively massive relational data. FuzViz relies on a method building automatically linguistic summaries, that provide concise and intelligible insights in the data content. It offers an interactive view of these summaries, dynamically recomputed on demand. To ensure a fluid exploration of the data, FuzViz exploits the proposition of a highly efficient method for estimating the cardinality of the summary properties, estimated from statistics about the data distribution stored in the relational data base, consolidated by a sampling-based approach. The proposed workflow also involves a vocabulary inference mechanism from these statistics.

Keywords:
Linguistic summarization, vocabulary inference, cardinality estimation, big data.

1 Introduction

Résumer linguistiquement un jeu de données consiste à fournir à un utilisateur un ensemble d'énoncés qui décrivent des propriétés observées dans les données. Ces énoncés sont construits selon un modèle de la forme « Q X *sont* P », appelé *protoforme*, où X désigne les données analysées, P la propriété observée, définie comme une combinaison conjonctive de modalités floues tirées d'un vocabulaire flou, et Q un quantificateur linguistique. Pour des données concernant des vols d'avion, un tel résumé peut par exemple être « *très peu de vols sont (tels que) la distance est courte et le retard à l'arrivée est très long* ». Ces résumés linguistiques fournissent un aperçu à la fois concis et informatif du contenu des données, permettant à des experts métier de traduire des données massives en connaissances utiles.

Cet article décrit un outil complet, appelé FuzViz, qui calcule efficacement de tels résumés pour des données stockées dans un Système de Gestion de Base de Données (SGBD). FuzViz offre de plus des fonctionnalités pour aider les utilisateurs dans la définition du vocabulaire sur lequel reposent les résumés, ainsi qu'une interface interactive de navigation dans les résumés.

Comparée aux approches de la littérature [1, 2], l'originalité de FuzViz est de travailler non pas à partir des données, mais des statistiques maintenues par le SGBD, qui lui permettent de produire les résumés en moins d'une seconde, quelle que soit la quantité des données.

La section 2 rappelle brièvement les notions sur lesquelles repose FuzViz. La section 3 décrit des travaux liés sur l'élicitation de vocabulaire et l'exploration de données, soulignant les caractéristiques de FuzViz. Les différentes étapes de son fonctionnement sont détaillées dans la section 4 puis illustrées sur un jeu de données représentatif dans la section 5.

2 Contexte

Cette section rappelle quelques notions utiles sur les données relationnelles et la définition formelle de vocabulaire.

2.1 Données relationnelles et métadonnées

Le principal rôle d'un SGBD est de stocker des données relationnelles et de fournir des fonctionnalités d'interrogation efficaces, même dans le cas de données massives. Pour cela, le SGBD exploite notamment des statistiques sur la distribution des données : pour chacun des attributs, numérique ou catégoriel, d'une relation, un SGBD enregistre dans une table interne, appelée *métadonnées*, la liste des k valeurs les plus fréquentes ainsi que leurs fréquences relatives ($k = 100$ par défaut dans PostgreSQL). De plus, pour les attributs numériques, le SGBD construit généralement un histogramme de type *equi-depth* [4], contenant par défaut $h = 100$ classes, pour représenter la distribution des valeurs moins fréquentes.

Pour des raisons d'efficacité, les SGBD ne maintiennent que des statistiques monodimensionnelles, même s'il existe des travaux sur des histogrammes multidimensionnels [5].

2.2 Vocabulaire de l'expert

FuzViz exploite par ailleurs un vocabulaire, duquel sont tirés les termes linguistiques utilisés dans les résumés. Formellement, un vocabulaire est défini comme un ensemble $\mathcal{V}$ de n variables linguistiques $V_i = \langle A_i, \{v_{i1}, \ldots, v_{iq_i}\}, \{l_{i1}, \ldots, l_{iq_i}\}\rangle$, $i = 1..n$: A_i désigne l'attribut concerné, q_i le nombre de modalités, v_{is} leurs fonctions d'appartenance respectives et l_{is} les étiquettes linguistiques associées. Ainsi, un attribut A_i décrivant l'heure de départ d'un vol d'avion peut être associé à $q_i = 4$ modalités, ayant les labels, l_{i1} =«*matinée*», l_{i2} =«*après-midi*», l_{i3} =«*soirée*» et l_{i4} =«*nuit*».

Les modalités d'un attribut sont contraintes à former une partition floue forte, c'est-à-dire être telles qu'une valeur du domaine de l'attribut considéré ne peut satisfaire que deux modalités au plus (qui doivent être adjacentes lorsque l'attribut est numérique). De telles partitions augmentent en effet l'interprétabilité des partitions.

3 Travaux liés

Cette section décrit brièvement des travaux existants sur l'élicitation de vocabulaire et l'exploration de données.

3.1 Élicitation du vocabulaire

Les résumés extraits d'une base de données dépendent du vocabulaire considéré, qui fournit une interface entre l'espace de description des données et l'espace symbolique du raisonnement humain. Il est essentiel que l'expert ait une bonne compréhension de la signification de ses modalités. Aussi, certains travaux proposent des interfaces intuitives pour définir et modifier ce vocabulaire [6]. Il est également crucial que les partitions floues permettent une description adéquate de la distribution des données [7], des stratégies coopératives d'élicitation, basées sur des mesures numériques de cette adéquation, ont été proposées [8]. D'autres approches de construction de vocabulaire incluent explicitement des contraintes d'interprétabilité [10] ou de discrimination entre classes à distinguer [11].

Comme décrit dans la section 4.1, le système FuzViz propose une méthode d'élicitation de vocabulaire non supervisée, exploitant les métadonnées fournies par le SGBD.

3.2 Résumés et exploration de données

De nombreux travaux ont été effectués pour rendre le processus de résumé de données efficace ou l'adapter à différents contextes applicatifs, voir par exemple [12]. Un atout important des outils est de permettre à l'utilisateur d'explorer les données et leurs résumés [13]. Comme décrit dans la section 4.3, le système FuzViz proposé offre des fonctionnalités d'exploration riches, affichant des vues interactives des données mises à jour dynamiquement.

Une question cruciale est alors d'être capable de générer rapidement des résumés, en particulier dans le cas où les données sont massives. Dans

ce but, il a été proposé d'exploiter les statistiques gérées par le SGBD pour fournir des estimations fiables et rapides des cardinalités d'ensembles de propriétés candidates [3]. Comme décrit dans la section 4.2, FuzViz propose une stratégie permettant d'améliorer la précision de cette approche, en exploitant une consolidation par échantillonnage.

4 Exploration avec FuzViz

Cette section présente les trois composants de FuzViz permettant de guider les utilisateurs dans l'exploration des données, permettant respectivement l'inférence du vocabulaire, l'estimation des cardinalités des résumés et l'interface de navigation.

4.1 Inférence de vocabulaire

Vue d'ensemble. FuzViz intègre trois stratégies pour aider l'utilisateur à définir un vocabulaire pour chaque attribut numérique, illustrées sur la figure 1. La première, indépendante de la distribution des données, consiste à décomposer le domaine en q modalités de même largeur, q étant fourni par l'utilisateur ($q = 4$ par défaut [10]). La deuxième stratégie décompose le domaine en q modalités représentant approximativement le même nombre de données, à la manière d'un histogramme *equi-depth*. FuzViz propose de plus une troisième stratégie dite « adaptative », détaillée ci-dessous, qui identifie des régions denses par analyse de la distribution des données basée sur un histogramme reconstitué à partir des métadonnées.

Les modalités générées par ces stratégies peuvent ensuite être modifiées facilement, pour définir leurs étiquettes ou modifier leurs bornes.

Approche adaptative proposée. On note $H_A = \{b_1, b_2, \ldots, b_h\}$ un histogramme consolidé décrivant la distribution des n-uplets sur le domaine D d'un attribut A, tel que stocké par le SGBD (cf Sec. 2.1) : H_A regroupe l'histogramme *equi-depth* des métadonnées ainsi que les valeurs les plus fréquentes écrites sous forme de classes. Chaque classe $b_j, j = 1, ..., h$ est associée à sa fréquence relative notée σ_{b_j}. On note $\hat{\sigma}_{H_A}$ la fréquence relative moyenne calculée par $\hat{\sigma}_{H_A} = \frac{1}{h}\sum_{j=1}^{h} \sigma_{b_j}$.

H_A, vu comme une séquence de h classes, est traduit en un mot de h symboles $\triangle$ ou ∇, selon que la fréquence relative est supérieure ou inférieure à la fréquence moyenne :

$$b \longrightarrow \begin{cases} \triangle & \text{si } \sigma_b \geq \hat{\sigma}_{H_A} \\ \nabla & \text{sinon.} \end{cases} \quad (1)$$

Dans l'esprit de l'approche présentée dans [11], la stratégie proposée vise à identifier des séquences contenant une large majorité de $\triangle$ qui définissent le noyau de modalités floues. En partant de chaque $\triangle$, l'algorithme recherche les intervalles les plus larges contenant une proportion de ∇ inférieure à un paramètre δ. Ce paramètre est un seuil empiriquement fixé par défaut à $0,25$. L'algorithme ajoute ensuite des transitions graduelles entre les noyaux adjacents, afin de satisfaire les contraintes de construction d'une partition floue forte.

4.2 Estimation des cardinalités floues

Problématique. FuzViz propose d'estimer la cardinalité relative de propriétés construites conjonctivement à partir du vocabulaire considéré en utilisant seulement les métadonnées du SGBD.

Pour une propriété atomique, définie par sa fonction d'appartenance pour un attribut donné, FuzViz applique la méthode proposée dans [3, 9] : elle agrège, par une intégrale de Choquet, les classes de l'histogramme et les valeurs fréquentes stockées par le SGBD pour l'attribut concerné, pondérées par le degré d'appartenance de la moyenne de chacune. Cette approche présente plusieurs avantages : (i) elle peut être appliquée pour les attributs numériques comme catégoriels, (ii) son temps de calcul est indépendant du nombre de données et de l'ordre de quelques millisecondes, (iii) elle a une très bonne précision par rapport à un parcours intégral des données (cf section 5).

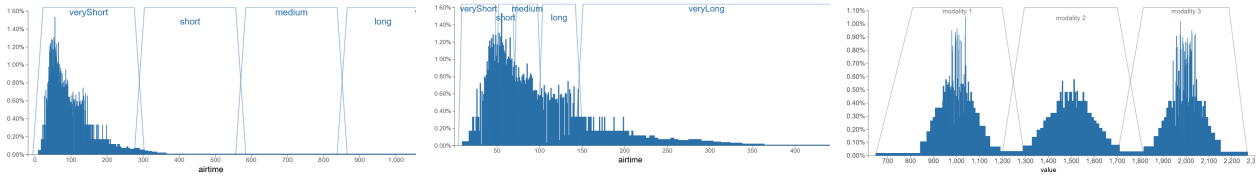

FIGURE 1 – Stratégies d'inférence de vocabulaire : (gauche) Partition equi-width, (milieu) Partition equi-depth, (droite) Partition adaptative.

Les propriétés conjonctives combinent des modalités de plusieurs attributs, en considérant au plus une modalité par attribut. Ainsi, pour une base décrivant des vols d'avion, une conjonction peut être *DureeVol.longue* $\wedge$ *Distance.courte*, pour les modalités $\ll$ *longue* $\gg$ et $\ll$ *courte* $\gg$ des attributs $\ll$ *DureeVol* $\gg$ et $\ll$ *Distance* $\gg$ respectivement. Comme indiqué dans la section 2.1, la plupart des SGBD ne gèrent pas de statistiques concernant des conjonctions d'attributs, à cause des coûts de calcul et de stockage. Plusieurs heuristiques peuvent être envisagées pour estimer la cardinalité de telles conjonctions, leur validité dépend de la question centrale de la dépendance entre les modalités, qui détermine le nombre de n-uplets qui satisfont toutes les modalités.

En effet, si les modalités sont indépendantes et qu'il n'y a pas de lien entre elles, la cardinalité relative d'une conjonction de propriétés $P = m_1 \wedge ... \wedge m_p$ est égale au produit de leurs cardinalités respectives : $\sigma_P = \prod_{m \in P} \sigma_m$, notée σ_{prod}. C'est l'hypothèse choisie par la majorité des SGBD. Elle peut être illustrée par le cas de *Distance.courte* $\wedge$ *Periode.été*.

Si les modalités sont dépendantes, elles sont satisfaites ensemble par les mêmes n-uplets. Par exemple *DureeVol.longue* $\wedge$ *Distance.longue*. Dans ce cas, la cardinalité relative de la conjonction P est $\sigma_P = \min_{m \in P} \sigma_m$, notée σ_{min}.

Si les modalités sont incompatibles, aucun n-uplet ne satisfait toutes les modalités simultanément, par exemple *Distance.longue* $\wedge$ *DureeVol.courte*. Dans ce cas, $\sigma_P = 0$.

Approche proposée. En absence de statistiques appropriées, et aussi parce que l'utilisateur peut modifier les modalités à tout instant, FuzViz implémente une méthode heuristique pour estimer la cardinalité de conjonctions avec un faible temps de calcul et une meilleure précision qu'avec l'hypothèse d'indépendance des attributs. Elle consiste à extraire un petit échantillon des données pour évaluer la pertinence de l'hypothèse d'indépendance, en calculant un *score de dépendance*, noté s_P, qui qualifie la dépendance observée sur cet échantillon sous la forme d'un réel dans l'intervalle $[-1, +1]$.

S'il est positif, ce score indique que les modalités sont plutôt dépendantes. La cardinalité relative de leur conjonction est estimée entre σ_{prod} et σ_{min}. S'il est négatif, les modalités sont plutôt incompatibles ou indépendantes. La cardinalité de leur conjonction est entre 0 et σ_{prod}.

Les scores de dépendances de toutes les conjonctions sont affichés dans FuzViz sous forme d'un tableau qui peut être classé par score croissant, resp. décroissant, faisant apparaître en tête les conjonctions des modalités les moins, resp. les plus dépendantes. Cela permet à l'utilisateur de mieux comprendre les relations entre les modalités dans le jeu de données.

FuzViz propose de calculer la cardinalité de la conjonction σ_P par une interpolation linéaire entre les deux cas extrêmes, en calculant leur moyenne pondérée par le score de dépendance :

$$\sigma_P = \begin{cases} (1 - s_P)\sigma_{prod} + s_P\sigma_{min} & \text{si } s_P \geq 0 \\ (1 + s_P)\sigma_{prod} & \text{sinon.} \end{cases}$$

Ce calcul est extrêmement rapide, une fois connu le score de dépendance. Pour l'estimer, FuzViz construit un échantillon des données dont la taille est calibrée pour que le temps de calcul soit d'environ 15 secondes. L'échantillon doit contenir suffisamment de données représentatives pour que l'estimation du score de dépendance soit fiable. La cardinalité relative de la conjonction σ_P ainsi que les cardinalités relatives des modalités individuelles, σ_m, $m \in P$, sont ensuite calculées, de même que les valeurs σ_{min} et σ_{prod}. Enfin, le score est calculé par l'équation ci-dessous, qui est la réciproque de l'équation précédente :

$$s_P = \begin{cases} (\sigma_P - \sigma_{prod})/(\sigma_{min} - \sigma_{prod}) \\ \quad \text{si } \sigma_P \geq \sigma_{prod} \\ (\sigma_P/\sigma_{prod}) - 1 \quad \text{sinon.} \end{cases}$$

Les scores de dépendance de toutes les modalités des attributs sélectionnés par l'utilisateur sont calculés avec le même échantillon de données, pour diminuer le temps de calcul. En effet, les cardinalités individuelles sont partagées par beaucoup de conjonctions. Ces scores sont mémorisés et ne sont recalculés que si l'utilisateur modifie une modalité ou ajoute de nouveaux termes à la conjonction.

4.3 Affichage et exploration des résumés

Le troisième composant de FuzViz est l'interface de navigation dans les résumés, construits à partir du vocabulaire tel que défini dans la section 4.1 et extraits selon la méthode présentée dans la section précédente. Ces résumés sont un ensemble d'énoncés de la forme « *Q X sont P* » qui décrivent les propriétés P observées dans les données.

Une première visualisation prend la forme de listes, comme illustré sur la figure 2. Il est possible de classer les résumés par cardinalité, quantificateur ou conjonction.

FuzViz propose également une vue interactive pour explorer les propriétés des données, illustrée sur la figure 3. Elle repose sur un diagramme de type *Sunburst*, qui constitue une extension hiérarchique d'un diagramme circulaire représentant les différentes modalités d'un attribut sous la forme de secteurs dont l'amplitude angulaire correspond à la cardinalité relative. Il comprend plusieurs couches qui représentent chacune les attributs employés dans la conjonction choisie par l'utilisateur. Un cadre flottant apparaît à proximité de la souris pour indiquer la cardinalité estimée (ou réelle si elle est disponible) de la modalité désignée. Chaque modalité est cliquable et la vue plonge alors sur le détail de cette modalité. Avec cette visualisation, l'utilisateur peut explorer différentes conjonctions de termes. Lorsque l'utilisateur modifie la définition de l'une des modalités, cette vue est entièrement mise à jour en une fraction de seconde.

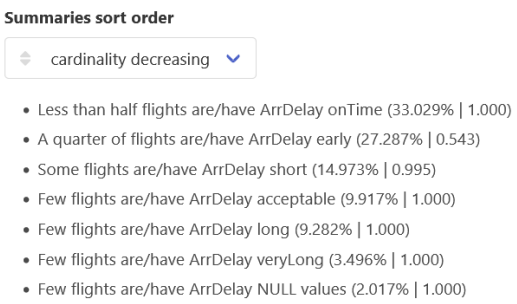

FIGURE 2 – Extrait des résumés générés par FuzViz sur la base des vols d'avion (cf Section 5).

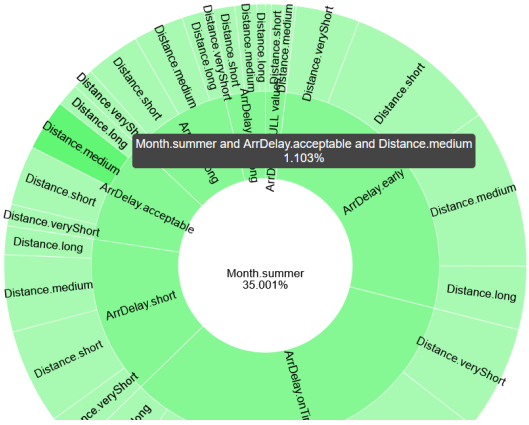

FIGURE 3 – Extrait du diagramme solaire.

5 Résultats expérimentaux

Cette section présente les résultats de FuzViz en termes de précision des estimation, puis de rapidité de calcul, obtenus sur la base *flight database* [1] Celle-ci contient plus de 123 millions d'enregistrements décrivant des vols d'avion aux États-Unis entre 1987 et 2008 selon des attributs tels que *Month*, *DayOfWeek*, *Distance*, *AirTime* pour n'en citer que quelques uns. Dans les scénarii présentés, FuzViz est implémenté sous la forme d'un serveur web portable (Python Flask) avec une interface graphique dynamique (Vue.js, D3.js, WebGL). Il tourne sur un Xeon 2.8Ghz avec 32Go de mémoire RAM. Les données explorées sont stockées sur un serveur PostgreSQL 13 sur la même machine.

5.1 Cardinalités de propriétés atomiques

La figure 4 montre la distribution des taux d'erreur relative, définis par $erreur(\sigma, \sigma_{reelle}) = (\sigma - \sigma_{reelle})/\sigma_{reelle}$, sous forme d'histogrammes. On observe que la plupart des modalités sont correctement estimées, quelques unes le sont très mal. La plus mauvaise estimation est sur la modalité $SecurityDelay.short$: la cardinalité estimée est de $0,0003$ pour une cardinalité exacte de $0,000127$, soit une erreur relative de 136%. Cette erreur considérable est due principalement au très grand nombre de valeurs `NULL` dans la colonne concernée, plus de 72%. La présence de valeurs indéfinies a un fort impact sur la qualité des métadonnées du SGBD car elles réduisent d'autant le nombre de valeurs pertinentes.

5.2 Cardinalités de conjonctions

Trois cardinalités ont été calculées sur toutes les conjonctions possibles de quelques attributs : les cardinalités réelles σ_{reelle}, les cardinalités estimées sous hypothèse d'indépendance σ_{prod} et les cardinalités estimées avec le score de dépendance σ_P. Ensuite, l'erreur relative entre les estimations et les cardinalités réelles est comparée par la différence $|erreur(\sigma_{prod}, \sigma_{reelle})| - |erreur(\sigma_P, \sigma_{reelle})|$.

1. `https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009`.

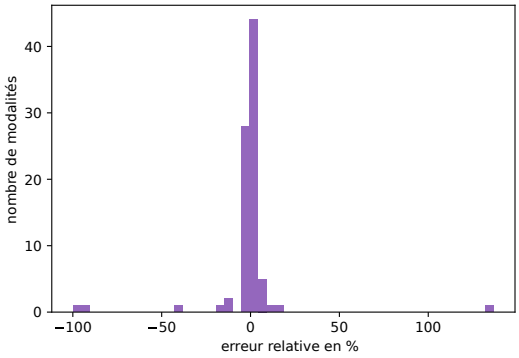

FIGURE 4 – Erreurs d'estimation des modalités individuelles.

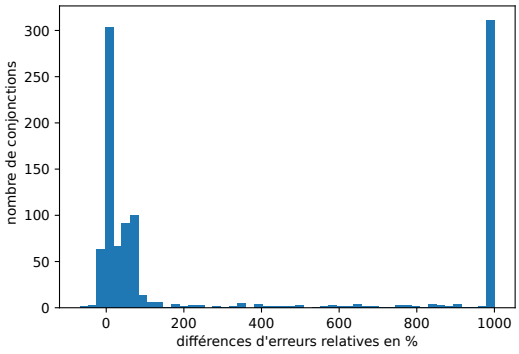

FIGURE 5 – Différences d'erreur d'estimation sur des conjonctions de 4 attributs dépendants.

Attributs dépendants. Le premier cas concerne 1049 conjonctions construites sur 4 attributs, *AirTime*, *ArrDelay*, *DayOfWeek* et *Distance*, qui présentent des dépendances : la durée du vol est corrélée avec la distance et dans une moindre mesure avec le retard à l'arrivée. En ce qui concerne les temps de calcul, les cardinalités réelles ont demandé plus de 7 heures, les cardinalités estimées à l'aide des métadonnées en moins d'une seconde. Enfin, le calcul de tous les scores de dépendance a duré 13 secondes sur un échantillon de 62061 n-uplets.

La figure 5 montre l'histogramme des

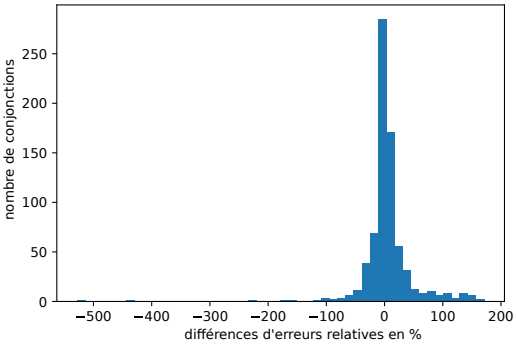

FIGURE 6 – Différences d'erreur d'estimation sur des conjonctions de 4 attributs indépendants.

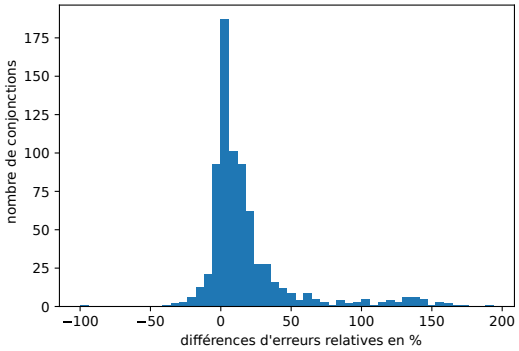

FIGURE 7 – Différences d'erreur d'estimation avec un échantillon plus grand.

différences d'erreurs relatives en pourcentage, bornées à 1000%, entre l'hypothèse d'indépendance et la méthode utilisant le score de dépendance. La figure montre clairement que la plupart des barres se trouvent du côté positif, c'est à dire que la méthode proposée dans cet article donne de bien meilleurs résultats que l'hypothèse d'indépendance des attributs. La méthode proposée donne d'un peu moins bons résultats pour seulement 60 conjonctions sur les 1049.

Attributs indépendants. Le second cas (figure 6) montre une situation moins favorable, lorsque les modalités ne sont pas dépendantes. 749 conjonctions ont été construites sur 4 attributs, *DepTime*, *Distance*, *Month* et *Origin*. La méthode proposée montre des améliorations dans les estimations de cardinalités pour un certain nombre de conjonctions, mais aussi des erreurs plus nombreuses que dans le cas précédent.

L'une des plus mauvaises estimations correspond à la conjonction *DepTime.night* $\wedge$ *Distance.veryShort* $\wedge$ *Month.summer* $\wedge$ *Origin.small* : sa cardinalité réelle est $\sigma_{reelle} = 0,000764$. Elle est estimée à $\sigma_{prod} = 0,000531$ avec 30% d'erreur sous hypothèse d'indépendance, tandis que la méthode proposée l'estime à $\sigma_P = 0,00142$, soit 85% d'erreur relative. Il est à noter que toutes ces cardinalités sont extrêmement faibles. L'erreur plus importante est due à un échantillonnage trop faible. La figure 7 montre les résultats obtenus avec 2 minutes de calcul pour extraire $0,5\%$ des données de la base, environ 6×10^5 n-uplets. Elle montre clairement que la méthode proposée obtient de meilleurs résultats.

Discussion. La méthode proposée s'avère nettement meilleure que l'hypothèse d'indépendance dans le cas d'attributs dépendants. Ce score est évalué assez rapidement, environ 15 secondes sur un échantillon aléatoire des données. Toutefois, il n'y a pas d'amélioration dans le cas de modalités indépendantes, sauf si l'utilisateur accepte de consacrer davantage de temps de calcul à l'extraction d'un échantillon plus grand.

5.3 Efficacité

La figure 8 montre le temps nécessaire pour résumer environ 123 millions de n-uplets en fonction du nombre d'attributs considérés. Étant basé sur des statistiques maintenues par le SGBD, le temps de calcul des résumés est indépendant de la taille de la base de données et est inférieur à 1 seconde. Le délai additionnel nécessaire pour calculer les scores de dépendance est un temps constant d'environ 15 secondes, dépensé une seule fois tant que les modalités ne sont pas redéfinies ou que la conjonction n'est pas modifiée.

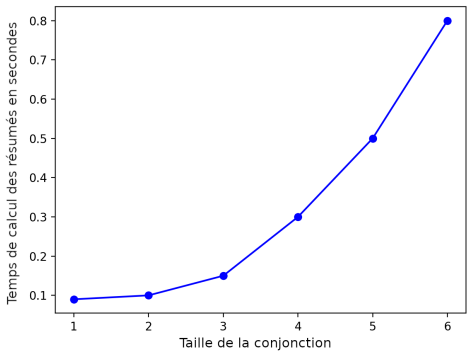

FIGURE 8 – Temps de calcul des résumés selon la taille de la conjonction.

Il est important de noter que, du fait que les cardinalités sont exprimées à l'aide de quantificateurs flous, de petites erreurs d'estimation ont un impact très faible sur les énoncés générés : elle aboutissent souvent à des résumés identiques à ceux qui seraient calculés par un parcours intégral, très lent, des données (cf [3]).

6 Conclusion

L'outil FuzViz présenté dans cet article permet de fournir à un utilisateur une vue concise des propriétés présentes dans les données et ainsi de l'aider à en extraire des connaissances, à partir des résumés linguistiques : il combine une grande efficacité calculatoire à une visualisation interactive et dynamique, permettant de naviguer dans les données. Le calcul d'un résumé linguistique d'un jeu de données contenant des millions de n-uplets peut être obtenu et affiché en moins d'une seconde. L'approche proposée exploite les statistiques gérées par le SGBD, à la fois pour suggérer un vocabulaire correspondant à la distribution des données et pour estimer les cardinalités des combinaisons conjonctives de termes de ce vocabulaire.

Les travaux en cours visent à améliorer encore la précision des estimations de cardinalités de conjonctions de termes, notamment en envisageant des modèles auto-appris, comme par exemple des forêts aléatoires ou des auto-encodeurs, de représentation de la distribution multidimensionnelle des données.

Remerciements :
Ce travail s'inscrit dans le cadre du projet SEA Defender financé par la Direction Générale de l'Armement.

Références

[1] R. A. E. Andrade, R. B. Pérez, A. C. Ortega, J. M. Gómez, and A. R. Valdés, *Soft Computing for Business Intelligence*. Springer, 2014.

[2] G. Smits, O. Pivert, and R. R. Yager, "A soft computing approach to agile business intelligence," in *2016 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, 2016, pp. 1850–1857.

[3] G. Smits, P. Nerzic, O. Pivert, and M.-J. Lesot, "Efficient generation of reliable estimated linguistic summaries," in *2018 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2018.

[4] Y. Ioannidis, "The history of histograms (abridged)," in *Proceedings 2003 VLDB Conf.*. Elsevier, 2003, pp. 19–30.

[5] N. Bruno, S. Chaudhuri, and L. Gravano, "Stholes : A multidimensional workload-aware histogram," in *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, 2001, pp. 211–222.

[6] R. R. Yager, M. Z. Reformat, and N. D. To, "Drawing on the ipad to input fuzzy sets with an application to linguistic data science," *Information Sciences*, vol. 479, pp. 277–291, 2019.

[7] M.-J. Lesot, G. Smits, and O. Pivert, "Adequacy of a user-defined vocabulary to the data structure," in *2013 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2013. 1–8.

[8] G. Smits, O. Pivert, and M.-J. Lesot, "Vocabulary elicitation for informative descriptions of classes," in *2017 Joint 17th World Congress of Int. Fuzzy Systems Association and 9th Int. Conf. on Soft Computing and Intelligent Systems (IFSA-SCIS)*. IEEE, 2017.

[9] G. Smits, P. Nerzic, O. Pivert, and M.-J. Lesot, "FRELS : Fast and Reliable Estimated Linguistic Summaries," in *IEEE International Conference on Fuzzy Systems*. IEEE, 2019.

[10] S. Guillaume and B. Charnomordic, "Generating an interpretable family of fuzzy partitions from data," *IEEE Trans. on fuzzy systems*, vol. 12, no. 3, pp. 324–335, 2004.

[11] C. Marsala, "Fuzzy partition inference over a set of numerical values," in *Proc. of the IEEE Int. Conf. on Fuzzy Systems*. Citeseer, 1995, pp. 1512–1517.

[12] F. E. Boran, D. Akay, and R. R. Yager, "An overview of methods for linguistic summarization with fuzzy sets," *Expert Systems with Applications*, vol. 61, pp. 356–377, 2016.

[13] G. Smits, R. R. Yager, and O. Pivert, "Interactive data exploration on top of linguistic summaries," in *2017 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2017. 1–8.