Évaluation de la qualité des données dans les systèmes d'intégration de données

Les besoins d'accéder, de façon uniforme, à des sources de données multiples, sont chaque jour plus forts, particulièrement, dans les systèmes décisionnels qui ont besoin d'une analyse compréhensive des données. Avec le développement des Systèmes d'Intégration de Données (SID), la qualité de l'information est devenue une propriété de premier niveau de plus en plus exigée par les utilisateurs.

Cette thèse porte sur la qualité des données dans les SID. Nous nous intéressons, plus précisément, aux problèmes de l'évaluation de la qualité des données délivrées aux utilisateurs en réponse à leurs requêtes et de la satisfaction des exigences des utilisateurs en terme de qualité. Nous analysons également l'utilisation de mesures de qualité pour l'amélioration de la conception du SID et de la qualité des données. Notre approche consiste à étudier un facteur de qualité à la fois, en analysant sa relation avec le SID, en proposant des techniques pour son évaluation et en proposant des actions pour son amélioration. Parmi les facteurs de qualité qui ont été proposés, cette thèse analyse deux facteurs de qualité : *la fraîcheur* et *l'exactitude* des données.

Nous analysons les différentes définitions et mesures qui ont été proposées pour la fraîcheur et l'exactitude des données et nous faisons émerger les propriétés du SID qui ont un impact important sur leur évaluation. Nous résumons l'analyse de chaque facteur par le biais d'une taxonomie, qui sert à comparer les travaux existants et à faire ressortir les problèmes ouverts.

Nous proposons un canevas qui modélise les différents éléments liés à l'évaluation de la qualité tels que les sources de données, les requêtes utilisateur, les processus d'intégration du SID, les propriétés du SID, les mesures de qualité et les algorithmes d'évaluation de la qualité. En particulier, nous modélisons les processus d'intégration du SID comme des processus de workflow, dans lesquels les activités réalisent les tâches qui extraient, intègrent et envoient des données aux utilisateurs. Notre support de raisonnement pour l'évaluation de la qualité est un graphe acyclique dirigé, appelé graphe de qualité, qui a la même structure du SID et contient, comme étiquettes, les propriétés du SID qui sont relevants pour l'évaluation de la qualité. Nous développons des algorithmes d'évaluation qui prennent en entrée les valeurs de qualité des données sources et les propriétés du SID, et, combinent ces valeurs pour qualifier les données délivrées par le SID. Ils se basent sur la représentation en forme de graphe et combinent les valeurs des propriétés en traversant le graphe. Les algorithmes d'évaluation peuvent être spécialisés pour tenir compte des propriétés qui influent la qualité dans une application concrète. L'idée derrière le canevas est de définir un contexte flexible qui permet la spécialisation des algorithmes d'évaluation à des scénarios d'application spécifiques.

Les valeurs de qualité obtenues pendant l'évaluation sont comparées à celles attendues par les utilisateurs. Des actions d'amélioration peuvent se réaliser si les exigences de qualité ne sont pas satisfaites. Nous suggérons des actions d'amélioration élémentaires qui peuvent être composées pour améliorer la qualité dans un SID concret. Notre approche pour améliorer la fraîcheur des données consiste à l'analyse du SID à différents niveaux d'abstraction, de façon à identifier ses points critiques et cibler l'application d'actions d'amélioration sur ces points-là. Notre approche pour améliorer l'exactitude des données consiste à partitionner les résultats des requêtes en portions (certains attributs, certaines tuples) ayant une exactitude homogène. Cela permet aux applications utilisateur de visualiser seulement les données les plus exactes, de filtrer les données ne satisfaisant pas les exigences d'exactitude ou de visualiser les données par tranche selon leur exactitude. Comparée aux approches existantes de sélection de sources, notre proposition permet de sélectionner les portions les plus exactes au lieu de filtrer des sources entières.

Les contributions principales de cette thèse sont : (1) une analyse détaillée des facteurs de qualité fraîcheur et exactitude ; (2) la proposition de techniques et algorithmes pour l'évaluation et l'amélioration de la fraîcheur et l'exactitude des données ; et (3) un prototype d'évaluation de la qualité utilisable dans la conception de SID.