Streszczenie

Problematyka niniejszej dysertacji wiąże się z ideą stosowania w systemach zarządzania dużymi bazami danych koncepcji przybliżonego przetwarzania zapytań (ang. *Approximate Query Processing, AQP*). W podejściu tym odchodzi się od wymogu używania i przechowywania oryginalnych zbiorów danych, na rzecz szybszego operowania ich istotnie mniejszymi podsumowaniami i wykonywania na nich zapytań zwracających wyniki przybliżone. Głównym celem było przy tym opracowanie metod ewaluacji i strojenia silników typu AQP bazujących na strukturach granularnych, gdzie podsumowania pełnych danych zastępuje się kolekcjami podsumowań opisujących mniejsze ich fragmenty (tzw. granule, mikro-klastry).

W ramach rozprawy zaproponowano rozszerzone środowisko ewaluacyjne dla silników typu AQP, które uwzględnia pomiar dokładności wyników testowych zapytań, wyrażany poprzez odpowiednio zaprojektowaną miarę podobieństwa między wynikami przybliżonymi i ich dokładnymi odpowiednikami. Pokazano również, jak wykorzystywać metody bazujące na podsumowaniach danych nie tylko na potrzeby klasycznych zapytań SQL, ale również w celu przybliżonego wykonywania zadań bardziej zaawansowanych, związanych w szczególności z zagadnieniami uczenia maszynowego oraz wizualnej analizy dużych zbiorów danych.

Praca składa się z ośmiu rozdziałów. Rozdziały 1 i 2 przedstawiają tezy rozprawy oraz przegląd ogólnoświatowych badań związanych z omawianą tematyką. Rozdziały 3, 4 i 5 opisują nowe podejście do oceny silników AQP, które – oprócz standardowych aspektów skalowania wielkości danych i zasobów obliczeniowych – uwzględnia także dokładność obliczeń, wyrażaną wspomnianą wyżej miarą podobieństwa. Specyfikację miary wsparto empirycznymi badaniami dotyczącymi zarówno oczekiwań co do jej własności, jak i percepcji użytkowników co do bliskości dokładnych i przybliżonych wyników liczbowych. Pokazano ponadto, jak wykorzystać zaproponowane podejście w testach służących udoskonaleniu algorytmów zaimplementowanych w jednym ze stosowanych obecnie komercyjnie silników AQP, przechowującym wyłącznie granularne podsumowania danych. Rozdziały 6 i 7 wzbogacają zaproponowane środowisko ewaluacyjne o aspekty wykraczające poza język SQL w celu pełniejszej oceny, na ile analizy i rozumowania bazujące na danych zapisanych w sposób przybliżony mogą odbiegać od analogicznych procesów bazujących na danych oryginalnych. W charakterze przykładu opracowano tu nowe wersje wybranych metod selekcji cech (ang. *feature selection*) działające na podsumowaniach danych i pokazano, jak badać podobieństwo pomiędzy ich wynikami w wersji klasycznej i aproksymacyjnej. Co więcej, uzupełniono omawiane środowisko o warstwę repozytorium metadanych, dającą bezpośredni dostęp do podsumowań granularnych przechowywanych we wspomnianym silniku. W Rozdziale 8 podsumowano zawartość i główne wyniki rozprawy.

Abstract

This dissertation's topics relate to the utilization of the concept of approximate query processing (AQP) in large database management systems. The idea is to loosen the requirements of using and storing the original data sets and, instead, to faster operate with significantly smaller data summaries, in particular using them to derive approximate SQL query answers. The main goal of this work was to develop new methods for evaluating and tuning the AQP database solutions based on granular structures, whereby summaries of the complete data are replaced by collections of summaries describing smaller data fragments (so-called granules, micro-clusters).

The main contributions include the design of an extended framework for evaluation of the AQP engines, which takes into account the accuracy measurement of test queries, expressed by means of an appropriately defined measure of similarity between approximate outcomes and their exact counterparts. It is also demonstrated how to utilize methods based on data summaries not only to accelerate classical SQL queries but also to support approximate execution of more advanced tasks, related in particular to the areas of machine learning and visual analysis of big data sets.

This work consists of eight parts. Chapters 1 and 2 present the dissertation's theses and the overview of worldwide research corresponding to the considered topics. Chapters 3, 4 and 5 describe a new approach to evaluation of the AQP engines, which – besides standard aspects of scaling data volumes and computational resources – reflects the accuracy of computing, by means of the aforementioned similarity measure. The measure's specification was supported by empirical investigations referring to both, the expectations related to its properties and the users' perception of closeness between exact and approximate numerical results. It was also shown how to deploy the proposed framework in tests aimed at improvement of algorithms implemented in one of the currently commercially used AQP engines, which is based entirely on granular data summaries. Chapters 6 and 7 enrich the proposed evaluation approach with some aspects going beyond SQL language, with the purpose of assessing to what extent the analytical and reasoning processes based on the approximate information can mimic the analogous processes based on the original data sets. As an example, it was presented how to create the new versions of some feature selection techniques performing on data summaries. It was proposed how to investigate the similarity between the outcomes of classical and approximate modes of feature selection. Moreover, the considered AQP evaluation framework was completed with the metadata repository layer that provides the users with a direct access to granular summaries stored in the considered engine. Chapter 8 concludes the contents and the main dissertation's results.