Didelės apimties duomenų analitika ir vizualizaviams

Prof. dr. Igoris Belovas, prof. habil. dr. Gintautas Dzemyda, prof. dr. Audronė Jakaitienė,
prof. dr. Julius Žilinskas, doc. dr. Tadas Žvirblis

Didieji duomenys – tai didžiuliai sudėtingi duomenų rinkiniai, kuriuos yra iššūkis saugoti, analizuoti ir vizualizuoti tiriamiems procesams ar rezultatams gauti. Didžiųjų duomenų analizė arba žinių atradimas iš duomenų - tai operacijos, skirtos įžvalgoms ir žinioms iš didelių duomenų rinkinių gauti. Įmonių duomenys yra didelių duomenų rinkinių pavyzdys, kurį sudaro įvairių verslo funkcijų duomenys, pavyzdžiui, gamybos, atsargų, pardavimų, finansų ir pan. Didelių duomenų rinkinių analizės atlikimo poveikis gali paskatinti įmones vykdyti duomenimis grindžiamą veiklą ir priimti sprendimus. Kitas dažnas didelių duomenų pavyzdys - daiktų interneto (IoT) paradigma. Iš įvairių tipų jutiklių gaunami duomenys sudaro didelius duomenų telkinius. Tokių duomenų pagrindu kuriami įvairūs procesų monitoringo sprendimai medicinoje, pramonėje, energetikoje ir kitose srityse. Didelės apimties duomenų analitika leidžia gauti naudingos informacijos iš duomenų bazių ar duomenų srautų, kurie yra didžiuliai apimties, greičio ir įvairovės prasme. Surinkus (didelius) duomenis, būtina juos analizuoti, kad būtų galima išgauti juose slypinčią informaciją. Šiuo atveju labai svarbu naudoti didžiųjų duomenų analizės priemones. Kuriami didelės apimties duomenų analitikos metodai, reaguojant į poreikį analizuoti didelius greitai surinktų sudėtingų duomenų kiekius. Dėl to duomenų gavimas ir apdorojimas vyksta dideliu tempu, kurio neįmanoma pasiekti klasikiniais skaičiavimo metodais. Duomenų vizualizavimas – tai galimybė duomenis pateikti žmogui suprantama forma, padedančia geriau juos suvokti. Vizualią informaciją žmogus pajėgus suvokti daug greičiau negu skaitinę, ji palengvina naujų žinių atradimą. Daugiamačių duomenų vizualizavimas ir dimensiškumo mažinimas yra neatskiriama duomenų tyrybos ir mašininio mokymosi dalis. Vizualizavimo uždavinys dažnai turi savyje optimizavimo uždavinį, kurio tikslo funkcija yra daugiaekstremė ir kurio kintamųjų skaičius yra labai didelis. Kyla iššūkiai greičiau ir tiksliau išspręsti tą uždavinį, ypač kai vizualizuojamų duomenų kiekiai yra labai dideli. Čia teks pasitelkti ir specialius optimizavimo metodus, ir našiuosius skaičiavimus.

Big data analytics and visualisation

Big data are large, complex datasets that are challenging to store, analyse and visualise to produce the processes or results under investigation. Big data analytics or knowledge discovery from data are operations designed to extract insights and knowledge from large data sets. Enterprise data is an example of a big data set, which consists of data from different business functions such as production, inventory, sales, finance, etc. The impact of carrying out analytics on big data sets can drive data-driven activities and decision-making in businesses. Another common example of big data is the Internet of Things (IoT) paradigm. The data from different types of sensors form big data clusters. Such data is the basis for a wide range of process monitoring solutions in medicine, industry, energy and other fields Big data mining makes it possible to extract useful information from databases or data streams that are massive in terms of volume, velocity and variety. Once (big) data has been collected, it needs to be analysed to extract the information it contains. The use of big data analytical

tools is essential here. Big data mining was developed in response to the need to analyse large volumes of complex data collected quickly. As a result, data acquisition and processing are carried out at high speed, which cannot be achieved by classical computational methods. Arising challenges for big data mining are: completeness, accuracy, and currency of discovered insights/patterns; quality of data to be mined; issues concerning big data storing/processing; modification of mining algorithms and techniques to deal with abundant, heterogeneous, and streaming data; dealing with evolving changes; dealing with dynamics/velocity of big data; flexibility of mining algorithms and techniques; representing and processing big data as events and event sequences. Visualisation is the ability to present data in a human-readable form that helps to understand it better. Visual information can be grasped much more quickly than textual information and facilitates the discovery of new knowledge. Visualisation and dimensionality reduction of multidimensional data is an integral part of data mining and machine learning. A visualisation problem often contains an optimisation problem with a multiextremal function and a very large number of variables. There are challenges in solving that problem faster and more accurately, especially when the amount of data to be visualised is very large. This will require the use of both special optimisation techniques and high-performance and parallel computing. New visualisation techniques will also be developed.