

ÚJ HASONLÓSÁGMÉRTÉK ALAPÚ MEGOLDÁSOK GYAKORI ELEMHALMAZOK FELTÁRÁSÁRA, ELEMZÉSÉRE ÉS VIZUÁLIS MEGJELÉNÍTÉSÉRE

Szerzők: **CSERESNYÉS Adrienn**, VI. évfolyam; **BARTOS Anikó**, I. évfolyam

Témavezető: **Dr. ABONYI János** egyetemi tanár

Intézmény: Pannon Egyetem, Mérnöki Kar, Folyamatmérnöki Intézeti Tanszék, Veszprém

A mai világban egyre népszerűbb adatbányászat segítségével az előzetesen nem ismert, hasznos információkat lehet kiszűrni. A kutatás során olyan jellegű adatok feltárása a cél, amely segítségével a lehető legtágabb értelemben vett eseménysorozatok és az ezekhez kapcsolódó komplex folyamatok hatékonyan jellemezhetőek.

A vásárlói kosár elemzés teljesen újszerű továbbfejlesztéseként nem csupán a termékek és a fogyasztók szimultán elemzése a cél, hanem az adatbányászat eredményének értelmezését megkönnyítendő a generált gyakori elemhalmazok szemléletes ábrázolása is. A javasolt megoldás elvi háttéréként az egyik legismertebb és legszélesebb körben alkalmazott algoritmus, az Apriori szolgál. A dolgozat ennek a módszernek a továbbfejlesztésére tesz kísérletet.

A gyakori elem párok kinyerésére az elemek leszámlálása helyett egy kevésbé ismert Bitable-reprezentációt segítségül hívva alapvető mátrix és vektorműveletekre támaszkodó algoritmus készíthető. A TDK-dolgozat egyik fontos eredménye ennek a Bitable-algoritmusnak a továbbfejlesztése, mely során megfelelő mátrix műveletekkel becslés adható a tranzakciók információtartalmára. Ezek alapján az algoritmus keresési tere jelentősen csökkenthető, azaz a gyakori elemhalmaz-keresés hatékonysága növelhető.

A Bitable, vagyis a vásárlói tranzakciókat tartalmazó mátrix megjelenítése a VAT-algoritmus segítségével történt, Tanimoto hasonlóságmértéket alkalmazva. A generált megoldások a biclustering eszközök fejlesztésének irányába mutatnak. A dolgozat eredményeként tehát új algoritmusok kerültek kidolgozásra a gyakori elemhalmazok feltárására és vizualizációjára, melyek implementálása MATLAB fejlesztői környezetben történt, és benchmarking, valamint folyamatmérnöki problémákon generált eredmények értékelésével volt tesztelve.

Kulcsszavak: **adatbányászat, Bitable, vásárlói kosár, mátrix**

Jelen publikáció a TÁMOP-4.2.2/B-10/1-2010-0025 azonosítójú projekt támogatásával valósult meg.

NOVEL CORRESPONDENCE-BASED APPROACH FOR THE DISCLOSURE, ANALYSIS AND VISUAL REPRESENTATION OF COMMON SETS OF ELEMENTS

Authors: **Adrienn CSERESNYÉS**, sixth-year student; **Anikó BARTOS**, first-year student;

Supervisor: **Dr. János ABONYI**, professor

Institution: University of Pannonia, Faculty of Engineering, Department of Process Engineering, Veszprém

Data mining has become a popular tool for the disclosure of valuable information not known *a-priori*. The aim of the research of this field is to reveal data for the efficient description of event sequences and the corresponding processes.

The present work introduces a novel development to the shopping-cart analysis, which provides an expressive graphical representation for the common element sets in addition to the simultaneous analysis of consumers and products. The proposed approach is an extension of the most widely applied Apriori algorithm, which applies Bittable, i.e., the transaction matrix representation and basic matrix and vector operations to unravel the common sets of elements instead of their enumeration. One of the key results of the proposed approach is a matrix- operation-based approximation method for the information content of the transactions, which can be applied to reduce the search space of the Bittable algorithm, thus enhancing its efficiency. The representation of the Bittable was performed using the VAT algorithm with the Tanimoto similarity coefficient. The approach has been implemented in the MATLAB environment and thoroughly tested with both benchmark examples and the evaluation of results from process engineering related problems.

The proposed approach incorporates biclustering tools, which can be a direction of further research in the future.

Keywords: **data mining, bittable, consumer basket, matrices**

This publication was supported by TÁMOP-4.2.2/B-10/1-2010-0025 project.