Válasz opponensi bírálatra

Doktori (PhD) értekezés címe: "Frequent pattern and process mining-based process operation analysis" (Folyamatirányítási rendszerek gyakori mintázat keresés és folyamatbányászat alapú elemzése)

Opponens: Dr. Fogarassyné dr. Vathy Ágnes, tanszékvezető egyetemi docens,
Pannon Egyetem Rendszer- és Számítástudományi Tanszék

Mindenekelőtt köszönöm Tanszékvezető Asszony szakértő és alapos opponensi véleményét, segítő értékelését és kérdéseit. Gondolatébresztő megjegyzései lehetőséget adtak az értekezésben bemutatott eredmények alapos elemzésére, értékelésére és továbbgondolására. Köszönöm bírálómnak a dolgozatomban előforduló elírásokkal és pontatlan szóhasználattal kapcsolatos észrevételeit. Ezek mindegyikével egyetértek. A bírálatban feltett kérdésekre, illetve megjegyzésekre válaszaimat az alábbiakban adom meg. Válaszaim sorrendje a kérdések illetve megjegyzések bírálatbeli előfordulási sorrendjét követi.

Válaszok

1. ***A kidolgozott módszertan egyik kulcseleme az időablak méretének meghatározása, amely meghatározza a trace-ek elkülönítését. A szerző ezt empirikus alapon közelíti meg: az esettanulmányban alkalmazott 220 másodperces küszöbértéket az alarm események kezdési időpontjai közötti időintervallumok mediánja alapján választotta meg. Ez a megközelítés azonban több kérdést is felvet. Implicit módon azt feltételezi, hogy az eseményközök fele külön trace-ket határoz meg, ami nem feltétlenül tükrözi a valós eseménykapcsolatok szerkezetét. Véleménye szerint hogyan lehetne ezt a megközelítést finomítani? Mit gondol például egy, az eseményközök gyakorisági eloszlása alapján történő „rés"-detektálási technika alkalmazásáról?***

   A felvetés jogos és releváns. Az esettanulmányban választott trace definiálási módszer az adott rendszer jellege alapján került kiválasztásra. Az időablakos módszer esetén az időablak meghatározása valóban lényeges, ez történhet statisztikai alapon vagy az üzemi tapasztalat alapján is (ahogyan megemlítem a 2-es fejezet idevágó részében). Természetesen más diszkrét események láncolatán alapuló folyamat esetén a trace definíciója eltérhet ettől, ez függ a rendszertől és az elemzési feladattól. Ilyen eltérő módszer lehet a kérdésben felvetett „rés"-detektálási módszer is, de szigorú szabályokkal rendelkező folyamatok esetén esemény-alapú szegmentálás is alkalmazható. Felmerülhet még az események közti szünetek hossz alapú szegmentálása, a különböző csoportok eseménnyé alakítása, majd gyakori szekvenciák keresése, ami feltárhatja az esetleges eseménylánc-eseményrés kapcsolatokat.

2. ***A gyakori szekvenciák kinyerésére alkalmazott GoKimp algoritmus előnye, hogy nem igényel előzetesen megadott paramétereket, mivel a mintakiválasztást tömörítési nyereség alapján végzi. Ez az „önhangoló" működés egyes esetekben előnyt jelenthet, ugyanakkor az automatizmusból fakadóan a felhasználói kontroll csökkenése potenciális kockázatot is hordoz. Kérdésem, hogy milyen típusú logfájl-struktúrák vagy alkalmazási környezetek esetén ajánlott különösen a GoKrimp algoritmus alkalmazása, illetve milyen jellegű hibák vagy torzítások kockázatával kell számolni az algoritmus használatakor?***

   Köszönöm a felvetést, a felhasználói kontroll csökkenése valóban jelenthet problémát. Az algoritmus ugyan hatékony eszköz adattömörítéshez, de a nem megfelelő bemenet esetén valóban

torz eredményeket adhat. Ilyen nem megfelelő bemenet lehet egy rosszul struktúrált log file, ahol nincsenek szűrve a „zajok" (irreleváns események) vagy nincsnek a trace-ek jól definiálva. A javasolt módszertanban a „Data-preprocessing" lépésben van lehetőség a beavatkozásra, így a felhasználó által ellenőrzött és feldolgozott adatot adunk a GoKrimp algoritmusnak, ami elegendő kontrollt adhat a kezünkbe. Az előfeldolgozáshoz a disszertáció 3-as, 5-ös és 6-os fejezetében tárgyalt módszerek nyújthatnak megoldást.

3. ***A 3. fejezetben bemutatott eljárások célja az előfeldolgozás és a releváns eseményláncok azonosítása, míg az 5. fejezetben a logfájl gyakori minták mentén történő strukturálása történik meg. Véleménye szerint e két módszertani megközelítés integrálható-e egy közös rendszerbe, és ha igen, milyen szinergiák vagy nehézségek adódhatnak a kombinálásukból?***

   A két megközelítés abszolút kombinálható, a 3-as fejezet gyakorlati példája bizonyította, hogy hiányosan vagy gyengén címkézett adatoknál az összetartozó eseményláncok szétválogatása nélkül bonyolult és nem pontos folyamatmodelleket kapunk. A dolgozatban erre a problémára kívántam rámutatni, illetve javasolni egy megoldást, ami a gyakori eseményszekvenciákon alapul. Természetesen a bemutatott módszertanok az adott rendszer és feladat függvényében fejleszthetőek, személyre szabhatóak. A disszertáció célja egy komplex megközelítés tárgyalása volt, ami az ipari naplófájlok feldolgozását segíti, figyelembe véve azok hiányosságait és sajátosságait.

4. ***A 4-es algoritmusban, ha a TransDist(A) tranzitív távolságokat ad vissza, akkor a következő lépésben miért szükséges ezen értékeket 1-ből kivonni? A kivonást követően ugyanis hasonlósági értéket kapnánk vissza, s nem különbözőség értékeket; holott az MDS távolságmátrixot vár bemenetként.***

   Ennek oka, hogy a TransDist(A) értékét a közös támogató trace-ek számából kalkuláljuk, így a hasonló szekvenciák nagyobb értéket vesznek fel, ezáltal ha ezeket kisebb távolsághoz szeretnénk rendelni, szükséges az 1-ből kivonás. Tehát valójában a TransDist(A) hasonlóság értéket számít (a tranzitív távolság számítás technikáját alkalmazva), a távolság a kivonás után áll elő.

Ismételten köszönöm Vathy Tanszékvezető Asszony szakértő és gondos bírálói munkáját.

Veszprém, 2025. július 3.

Bántay László