Válasz opponensi véleményre

Véleményező: Dr. Pollner Péter

Cím: Network science based analysis of human factors in systems engineering

Szerző: Gadár László

Témavezető: Dr. Abonyi János

Tisztelt Dr. Pollner Péter!

Nagyon szépen köszönöm, hogy dolgozatom igen részletes átolvasására időt és energiát fordított, és rendkívül érdekes és hasznos kérdéseket tett fel, megjegyzéseket fűzött hozzá.

A felvetődött témákra az alábbiakban reagálok. Bízom benne, hogy válaszaim kielégítőek lesznek.

Tisztelettel:

Gadár László

Az egyes tézispontok megfogalmazását nagyon általánosnak érzem. A vastaggal szedett tézispontok az adott téma feldolgozását lehetővé tévő módszertant nevezik meg. A tézispontok kifejtésében, az „Eredmények" pontban találtam olyan megfogalmazásokat, amely a jelölt munkáját és konkrét megoldásait, levont következtetéseit tartalmazzák.

A dolgozat célkitűzése elsősorban a hálózattudományi módszerek alkalmazása, és probléma orientált továbbfejlesztése volt. Ennek megfelelően a tézispontok megfogalmazásánál a vastagon szedett részek a módszertani újszerűségeket foglalják össze, amelyek egyedi módszertani fejlesztések eredményei, saját munkát tartalmaznak. A módszertanok alkalmazási területei és a kapott újszerű empirikus eredmények, mint a kutatómunka másodlagos hozadékai, külön pontban, az „Eredmények" részben kerültek ismertetésre a tézisek részeként.

Az oktatási programok és munkahelyek kapcsolásakor lehetségesnek tartja-e olyan mintázatok kibányászását, amellyel esetleg egy-egy személy beazonosítható lenne? Milyen módszert javasolna arra, hogy egy adatbázis anonimitását ellenőrizni lehessen?

Az államigazgatási adatbázisok egyedi rekordok szintjén történő összekapcsolásával kapott adathalmaz valóban magában hordoz anonimitási veszélyeket. Az adatbázis összeállítói ezt úgy kezelik, hogy egyes demográfiai változók kombinációjával kialakult csoportok esetén a csoportlétszám ne legyen kevesebb, mint 3 fő. Azonban saját tapasztalat is az, hogy baráti körben levő illetőről a demográfiai adatokon túli

ismert információ esetén a beazonosíthatóság lehetséges. Az anonimitás tartása érdekében az adatbázis összes változójának kombinációját figyelembe vevő (demográfia, munkahelyi-, diákhitel adatok) összes csoportnak legalább 3 fő nagyságúnak kellene lennie. Ez azonban nagy mértékben rontaná az adatbázis gazdagságát.

A 2.3 és 2.4 ábrán bemutatott kereset-különbséget meg tudja-e magyarázni, vagy legalább előzetes jelét látja-e a tanulmányok során befutott életpályákból, vagy az adatbázis egyéb jellemzőiből?

Az adatbázis tartalmazza az összes végzett fizetés adatát, így a mintavételes kutatások hibalehetőségeinek kiküszöbölésével bányászható a férfi-női fizetések különbségei is. A kutató munka során pusztán az adatbázis „képességeinek" demonstrálására került bemutatásra a férfi-női fizetéskülönbség tématerület, azonban ennek részletes vizsgálata már a fő célkitűzésen túli, ezért a véglegesített dolgozatban már nem szereplő eredmény, mivel egy önálló kutató munkát megérdemelne. Az előzetes eredmények azt sejtetik, hogy a férfi-női fizetés különbségek összefüggenek a diploma szakterületével, a foglalkozás szakterületével és szintjével, a foglalkozás bemeneti képzési követelményeivel és a fizetések szabályozottságával.

Hogyan változnak a 2. fejezetben meghatározott klaszterek jellemzői, ahogy az alfa paraméter változtatásával ritkul a hálózat? Vannak-e olyan elemek, amelyek átvándorolnak egyik klaszterből a másikba? Ha igen, hogyan magyarázható ez a vándorlás?

Az alfa paraméter növelésével, növekszik a rendszerben a klaszterek száma, csökken a klaszterekben levő csúcspontok száma, mert a (kibocsátási és fogadási kapacitásokat figyelembe vevő) konfigurációs modellhez képest a legerősebb kapcsolatok maradnak a hálózatban. Az alfa paraméter változtatásával egyes csúcspontok közösségei megváltoznak, amit a dolgozat 2.9 ábrája is megpróbált érzékeltetni. Van olyan divergensnek mondható csúcspont, amely a konfigurációs modellhez képest sok, de kis súlyú kapcsolattal kötődik egy közösséghez, és egy modulba kerül azokkal az alfa paraméter kis értéke esetén. Ugyanakkor egy-két erős kapcsolattal egy másik közösséghez is kapcsolódik. Az alfa paraméter növelésével, a kis súlyú, kis erősségű kapcsolatok megszűnésével, a divergens csúcspont a nagyobb súlyú, korában látens „szomszédaival" fog egy modulba kerülni. Pl. Agrármérnök szak α=0 esetén gazdasági és bölcsészeti szakokkal kerül egy modulba, míg α=2,9 esetén mérnöki szakokkal.

A 3.9 és 3.10 ábrákon egy-egy település külön ki van írva. Mi a szerepe ezeknek a településeknek, miért pont ezek vannak kiírva?

A kiírt településeknek nincs konkrét célja, az ábra azt érzékelteti, hogy az ábrázolásra kerülő pontok egy-egy települést reprezentálnak.

A dolgozatban bemutatott eredmények mennyire robosztusak a kiválasztott csoportkeresési eljárás algoritmusára nézve? Például a 3.13-as ábra eredménye mennyire lenne reprodukálható más klaszterezési eljárásokkal?

A 3. fejezetben bemutatott csoportkeresési eljárás robosztus, többszöri futtatásra is ugyanazokat az eredményeket adta. Más klaszterezési eljárások, mivel más null-modelleket alkalmaznak, biztosan

különböző csoportosítási eredményeket adnak. A cél éppen ez volt, hogy a null-modellek által meghatározott referencia csoportképző hatása kerüljön vizsgálatra. A csoportok összetétele a közgazdászok számára hasznos elemzési információt jelent.

A 4-ik fejezetben bemutatott eredményeknél az egyes cégekre külön-külön készültek elemzések. Ez egy természetes hozzáállás. De ha az anonim adatokat összekeverjük a cégek között, lehetséges-e létrehozni egy „átlagos" céget, ahol az egyedi cégekben megállapított pozíciójú személyek vannak összevonva, és a kapcsolataik átlagolva? Elképzelhetőnek tartja-e, hogy ha sok cégről rendelkezésére állna ilyen jellegű kimutatás, akkor a cégek összeátlagolásával létrehozható lenne-e egy tapasztalati null modell? Hogyan próbálna létrehozni egy ilyen, akár referenciaként használható, átlagos céget?

Egy referencia cég létrehozása, így benchmarkok kialakítása igen hasznos lenne. Az eltérések vizsgálatával értékelhetővé válnának a csúcspontok és maguk a hálózatok is, és az empirikus eredmények kevésbé lógnának a levegőben. A referencia létrehozásánál megvizsgálnám azt, hogy miért, milyen feladat megoldására hoznám létre. Óvatos lennék azzal, hogy alma az almával legyen átlagolva, és valószínűleg a hasonlók, de mégis különbözők megtalálása okozná a legnagyobb problémát. A gondolat érdekes, egy önálló kutatási témát mindenképp megérdemelne.

Feltehetőleg a szövegalkotás folyamán megváltozott a szöveg tördelése, így néhány elválasztójel a sorok közepére került a tézis füzetben. Esztétikai hátrányon kívül azonban nem okoznak ezek problémát, bár egy ilyen rövid, de mégis fontos részre nagyobb gondot lehetett volna fordítani.

A bírálónak teljesen igaza van ebben. A dolgozat és a tézisfüzetek LaTeX szövegtördelővel készültek. A magyar elválasztásokat kezelő program is tett a pdf fordítás során sorközi elválasztó jeleket, azonban a jelölt magyar helyesírási hiányosságai is szerepet játszanak a hibákban.

A disszertációban egyetlen zavaró megfogalmazást találtam: a 4 ig fejezetben a vásárlói korás formalizmusnál a jelölések azt indukálják, hogy A és B eseményeket jelöl pl. P(A|B), azonban itt A és B dimenziókat jelöl. Pl. a konfidencia definíciójánál az A unió B jelölés eseményekkel megfogalmazva A metszet B lenne. Bár ez egy tankönyvi szintű klasszikus probléma, de a dolgozat jellege miatt várhatóan olyanok is olvassák, akik nem jártasak az adatbányászat szabály-kinyerési zsargonjában, és ezért hibásnak gondolhatják a leírást.

A jelöléstechnika során a gyakori elemhalmaz keresés (frequent pattern mining) szakirodalma szerint jártam el, és ez valóban zavaró lehet más tudományágak művelői számára. Pl. Jiawei Han, Micheline Kamber, Jian Pei: Data mining concepts and techniques, Elsevier, 2011 tankönyve szerint (21. oldal) X unió Y egy olyan tranzakciót jelöl, amelyben X és Y együttesen van jelen.