Budapesti Corvinus Egyetem

Gazdaságinformatika Doktori Iskola

TÉZISGYŰJTEMÉNY

Saira Andleeb Gillani

A szövegbányászattól a szervezeti tudás létrehozásáig: Integrált szövegbányászati megoldás a szakterületi ontológiák gazdagítására és karbantartására

című Ph.D. értekezéséhez

Témavezető: Kő Andrea, PhD

Budapest, 2015

Információrendszerek tanszék

TÉZISGYŰJTEMÉNY

Saira Andleeb Gillani

A szövegbányászattól a szervezeti tudás létrehozásáig: Integrált szövegbányászati megoldás a szakterületi ontológiák gazdagítására és karbantartására

című Ph.D. értekezéséhez

Témavezető: Kő Andrea, PhD

Budapest, 2015

Tartalomjegyzék

I. A KUTATÁS HÁTTERE ÉS ÁTTEKINTÉSE

A szervezeteknek meg kell küzdeniük a szabályozási, társadalmi és gazdasági környezet összetett és folytonosan változó jellegéből fakadó kihívásokkal. Mindez növekvő igényt jelent a szervezeti tudás menedzsmentje számára olyan kérdésekben, mint például hogy hogyan biztosítható az alkalmazottak számára szükséges munkakör specifikus tudás a megfelelő időben és a megfelelő formában. Az alkalmazottaknak folyamatosan frissíteniük kell tudásukat, fejleszteniük kell kompetenciáikat, hogy alkalmazkodni tudjanak a gyorsan változó, megújuló igényekhez. Tudásmenedzsment szempontból a tudástáraknak kulcsszerepe van. Elsődlegesen ezek tartalmazzák a szervezetek intellektuális tőkéjét, amely egyrészt explicit tudás; miközben az alkalmazottak birtokolják azt a tacit tudást is, amely nehezen kivonható és kodifikálható. Az üzleti folyamatoknak kulcsszerep jut a szervezeti tudás menedzselése szempontjából, hiszen hordoznak mind explicit, mind tacit tudáselemeket. A szervezetek szempontjából meghatározó kérdések egyike az, hogy hogyan kezelhető ez a rejtett tudás a szervezeti tudás, különösképpen az alkalmazottak tudásának javítása érdekében a legmegfelelőbb tanulási és képzési anyagok biztosítása által; illetve hogyan biztosíthatjuk, hogy az üzleti folyamatok során alkalmazott tudás megegyezzen a tudástárakban és az alkalmazottak fejében lévővel.

PhD értekezésem központi témája az információ és tudás folyamatokból való kinyerési eljárásai és ezen szervezeti információ/tudás gazdagítása elsődlegesen szövegbányászati módszer és megoldás által; amely egyben fel tudja tárni a folyamatokba rejtett információt és tudást; dinamikusan kezelve a folyamatok adatait. Egy üzleti folyamat résztevékenységekre bontható és a tevékenységek mindegyike számos jellemzővel bír; ilyen a tevékenységek leírása, a kapcsolódó felelősségi körök, a végrehajtásra vonatkozó információk (kiváltó tényezők és események) és a felhasználandó források. Az értekezés átfogó kutatási kérdése, a folyamatmodellekből történő tudáskinyerés vizsgálata és a kinyert tudáselemek leképezése szakterületi ontológiára.

A folyamatmodellek tevékenységének „leírás” attribútuma beágyazott explicit és tacit tudáselemeket is tartalmaz. A kutatásom célja a rejtett tudáselemek azonosítása és feltárása, illetve azok kontextusba (szakterületi ontológia) helyezése. Kutatásom éppen ezért erősen támaszkodik a szövegbányászatra. Szövegbányászat alatt olyan módszer vagy megközelítés értendő, amely új, azaz addig ismeretlen információ kinyerésére irányul bármilyen típusú szövegből. A szövegbányászat használatával a leírás

kinyerhető egy adott tevékenységből, és ezután felhasználható egy olyan fogalomlista létrehozására, amely a kontextus függést megadó folyamat bemeneteként alkalmazható. Disszertációm átfogó célja egy olyan szövegbányászati megoldás megtervezése és kifejlesztése, amellyel kinyerhető a folyamatokba ágyazott tudás, majd leképezhető szakterületi ontológiára. A folyamatba rejtett információ szakterületi ontológia bővítésében (ontológia populáció, vagy angolul ontology population) való felhasználásához a naplófájlokból elérhető tevékenység leírások szövegbányászati feldolgozása szükséges. Ahogyan az ebből a rövid áttekintésből is kitűnik, a folyamatmodellezés jellege inkább leíró, míg számos ebből a modellezésből adódó kérdés kontextusfüggő válaszokat igényel. Kutatásomban a kontextus a Studio ontológia segítségével biztosított (Vas, 2007).

I.1 Az értekezés felépítése

Ebben a részfejezetben a disszertáció felépítését ismertetem. Az első fejezet a kutatási téma indoklásával, a problémakör megfogalmazásával, a kutatási kérdésekkel és a kutatás módszertanával foglalkozik. Áttekintést nyújtok a kutatás hátteréről, előfeltételezéseiről, az alkalmazott módszertanról. Ez a fejezet tárgyalja a kutatás problémakörét és a kapcsolódó kutatási kérdéseket is. A második fejezetben mutatom be a szakirodalmi áttekintést, a kutatás elméleti hátterét. Részletesen foglalkozom a szövegbányászattal, az információkinyeréssel, az ontológiatanulással, a szemantikus hasonlósági metrikákkal, az ontológiák karbantartási kérdéseivel, az ontológiák továbbfejlesztésére alkalmazott fogalomkiterjesztéssel (concept extraction to enrich ontologies), de kitérek a folyamatmenedzsment és a szemantikus folyamatmenedzsment területeire is. A harmadik fejezetben mutatom be azt az általam javasolt, tervezett és kifejlesztett ProMine keretrendszert, megoldást, amelyet kutatási kérdések megválaszolásában is alkalmazok. Részletesen tárgyalom a ProMine felépítését és funkcionalitását. A kutatási kérdések megválaszolása és a kapcsolódó kutatási állítások bizonyítása a negyedik, ötödik és hatodik fejezetekben található meg. Három esettanulmányon keresztül adok választ a kutatási kérdésekre és bizonyítom a kapcsolódó állításokat. Mindegyik fejezet más-más szakterülethez tartozik és más kutatási kérdéshez kapcsolódik. A hetedik fejezet zárja a disszertációt; ez a fejezet tartalmazza a konklúziót és a jövőbeli lehetséges kutatási irányokat is.

I.2 A kutatás céljai és főbb kérdései

Kutatásom kontextusát a szervezeti tudás menedzsmentje adja. Ennek a tudásnak egyrészről létezik egy explicit formája a tudástárakban, ontológiákban, másrészről egy tacit formája az alkalmazottak fejében. Sajátos tudásreprezentációs forma a folyamat, amely beágyazott formában tartalmazza a tudást. Az általam javasolt és kifejlesztett ProMine szövegbányászat-alapú tudásmenedzsment megoldás segít összekapcsolni egymással a tudásreprezentáció ezen formáit. A szövegbányászati megoldásom egyik fő hozzáadott értéke abban rejlik, hogy automatizálja a teljes folyamatot, nevezetesen az üzleti folyamatokból való tudáskinyerést, illetve hogy lehetővé teszi a szakterületi ontológia gazdagítását. Ha bármilyen változás történik a folyamatban vagy a tevékenységhez szükséges tudásban, a ProMine megoldás segítségével az új vagy módosított tudás eljuttatható az alkalmazottakhoz.

A disszertációm egyik célja a szervezeti folyamatokból és vizsgált szakterülethez kapcsolódó elektronikus formában elérhető dokumentumokból való tudáskinyerésre használt különböző módszerek és technikák azonosítása, kiértékelése.

Kutatásomban kiemelt szerep jut az üzleti folyamatokból való fogalomkinyerésnek (concept extraction) és az ontológiafejlesztésnek. A szövegbányászati megoldások azonosítják a tevékenységek és munkakör specifikus tudás, valamint az egyéb készségek, kompetenciák közötti kapcsolatot. A munkakörhöz kapcsolódó kompetenciák rendszere párosítható a szervezeti struktúrával és üzleti folyamatokkal. Az asszociációs szabályokon keresztül azonosított fogalmak közötti kapcsolatok képezik az ontológia továbbfejlesztésének bázisát. Így a szövegbányászati keretrendszerből kapott kimenetek felhasználhatóak szakterületi ontológia továbbfejlesztésére. Az így megszerzett tudás lehetővé teszi, hogy az alkalmazottakhoz eljuthasson a munkakörükhöz szükséges specifikus tudás. A ProMine megoldás újszerűsége azon alapul, hogy az üzleti folyamatok tevékenységeinek elemzése szövegbányászati technikák segítségével egyrészről lehetővé teszi a tudás kinyerését belőlük, másrészről ezen üzleti folyamatok a szervezeti tudásbázishoz kapcsolhatóak, ahol a folyamatstruktúra a tudásszerkezet felépítésére szolgál. Ebben a megközelítésben az ontológia maga a tudásbázis, amely egy bizonyos terület fogalmi leírására szolgál. Az elsődleges újítás az új szakterületi fogalmak kinyerésére és gazdagítására szolgáló új algoritmusokban és a statikus és dinamikus folyamatok tudásának integrálásában rejlik.

A kutatási kihívásoknak megfelelően a következő kutatási kérdéseket határoztam meg

és válaszoltam meg:

Első kutatási kérdés: Hogyan használhatjuk a szövegbányászatot a folyamatokból való tudáskinyerésre egy már létező ontológia fejlesztése vagy populációja érdekében?

Ez a kérdés a kutatás központi és egyben átfogó kérdése, a további kérdések alkérdéseit képezik ennek a problémafelvetésnek. A legfőbb kihívás ebben a kérdésben egy olyan szövegbányászati megoldás megtervezése és implementálása, amely összeköti a folyamatmodellezést és a kontextust megadó szakterületi ontológiát. A szövegbányászat által kínált lehetőségek vizsgálata és a kutatási kérdés megválaszolása érdekében áttanulmányoztam és elemeztem szövegbányászattal kapcsolatos szakirodalmat.

Második kutatási kérdés:

Milyen módszertan és fogalomkinyerési módszerek léteznek a folyamatokból kinyert tudás gazdagítására? Egy szövegbányászaton alapuló megoldás használható-e és hogyan ontológiatanulási célokra a gépi tanulással összefüggésben?

Ez az első kérdéssel összefüggő alkérdés a fő kutatási problémával foglalkozik, nevezetesen a folyamatokból való tudáskinyerési módszerek tárgyalásával. A kérdés megválaszolását azzal kezdem, hogy megvizsgálom, hogyan artikulálható az üzleti folyamatokban rejtett tudás a különböző szövegbányászati technikák használata révén. Elemzem a szakirodalomban már leírt szövegbányászati megoldásokat és módszereket valamint kitérek a kapcsolódó részterületek, mint a folyamatmenedzsment, folyamatbányászat (process mining), szövegbányászat, szemantikus technológiák és ontológiák bemutatására is. Disszertációm központi témája a szervezeti tudás szövegbányászati megoldásokkal történő kinyerése.

Számos információkinyerési megoldást megvizsgáltam, elemeztem. A már létező, egyéb megoldásokat megvizsgálva megállapítható, hogy ezek nem, vagy nehézkesen használhatóak a már meglévő, üzleti folyamatokból származó tudás gazdagítására szakterületi ontológiatanulási célokkal összefüggően. Ezen eszközök nehézkes alkalmazhatósága még inkább egyértelmű abban az esetben, amikor a szakterületi ontológia által nyújtott kiegészítő kontextus függő tudást az üzleti folyamatok gazdagítására használnánk.

A fentiek miatt olyan módszerek és megoldások megtervezése és implementálása szükséges, amelyek automatizálják az üzleti folyamatokból való információkinyerést, és

amelyek olyan módon fejlesztik tovább ezt a kinyert információt, hogy az hozzájárul a már létező szakterületi ontológiák gazdagításához. Ez a kutatási kérdés a tudáskinyerés automatizálásának lehetőségeit is érinti. Különböző szövegbányászati megközelítési módszereket dolgoztam ki a fogalomkinyerési probléma kezelésére. A sikeres technikák között találhatóak szabályalapú (Mykowiecka, Marciniak, & Kupść, 2009; Xu et al., 2010) és statisztikai módszerek is (Bunescu et al., 2005; Wu & Weld, 2007). Különböző ontológiatanulási megközelítések és eszközök, mint a Text2Onto (Cimiano & Völker, 2005), az OntoLT (Buitelaar, Olejnik, & Sintek, 2003), az OntoBuilder (Gal, Modica, & Jamil, 2004), a DODDLE-OWL (Morita, Fukuta, Izumi, & Yamaguchi, 2006) és az OntoGen (Fortuna, Grobelnik, & Mladenic, 2007), mind képesek félig automatikusan végrehajtani a fogalomkinyerést ontológiai célokra. Azonban a fent említett megoldások által szolgáltatott fogalomlisták minősége nem kielégítő, mert nem illeszkednek megfelelően a vizsgált szakterülethez, és leginkább a hagyományosan használt rangsorolási metrikákat (például TF-IDF) használják. Éppen ezért van szükség a legújabb szövegbányászati megközelítések használatára a fogalomkinyerésben.

Harmadik kutatási kérdés: Egy módosított szemantikus hasonlósági mérték jelentősen javítani fogja a szakterületi ontológia fejlesztésének és gazdagításának hatékonyságát és az ontológia minőségét.

Jelen disszertáció fő célja az üzleti folyamatok strukturálatlan (szöveges) adataiból való fogalomkinyerés. A szakirodalomban megtalálható a fogalomkinyerés néhány általános keretrendszere, ezek azonban kevésbé rugalmasak és alkalmazhatóak, mivel nagy részük az ún. „shallow NLP” technikákra támaszkodnak. A hangsúly ezekben az esetekben nem kontextusfüggő információkinyerésen van. Ezek a megoldások nem veszik kellőképpen figyelembe a szemantikus jellemzők kezelését, és az ilyen sémákból származó fogalmak kevéssé köthetők a vizsgált szakterülethez. Felhasználói szemszögből tekintve a kontextusba helyezett fogalmak sokkal nagyobb jelentőséggel bírnak döntéshozatal során. Vannak olyan szakértők, akik a szemantikus fogalomkinyerést ajánlják (Ercan & Cicekli, 2007; Gelfand, Wulfekuler, & Punch, 1998; Hassanpour, O'Connor, & Das, 2013; Kok & Domingos, 2008). A szemantikus hasonlóságot használó technológia lehetővé teszi egy adott fogalomhoz olyan további fogalmak azonosítását, amelyek még nincsenek jelen a tudásbázisokban és relevánsak a vizsgált szakterülethez. A szemantikus hasonlóság mértéke és a fogalompárok vélhetően javíthatják az ilyen rendszerek teljesítményét.

Mindazonáltal a szakirodalomban megtalálható szemantikai alapú módszerek nem adnak megoldást egy fontos problémára, nevezetesen nem támogatják a felhasználókat az üzleti folyamathoz kapcsolódó specifikus fogalmak azonosításában. Ez a probléma még bonyolultabbá válik, ha az üzleti folyamathoz nem áll rendelkezésre elegendő háttér információ, leírás (mint az általam feldolgozott folyamatok esetében is). Ebből következően a szemantikus fogalomkinyerés még mindig nyitott kérdést képez az ontológiafejlesztésben, és szükség van az NLP és szövegbányászati technikák komplexebb alkalmazására is. Egy olyan új szemantikus hasonlóságmértéket dolgoztam ki, amely segít a fogalomkinyerésben, és amellyel áthidalhatóak a korábbi szemantikus hasonlóságmértékek problémái.

Negyedik kutatási kérdés: Létező ontológiák felső kategóriáinak (osztályainak) használata javítani fogja-e a szövegbányászati megoldás eredményességét az ontológiagazdagítási folyamatban, vagy sem?

A manuális ontológia populáció (ontology population) és ontológiagazdagítás komplex és időigényes feladat, amely szakmai hozzáértést, többszöri szakértői egyeztetést és erőfeszítést igényel. Jelen disszertációban a célom az, hogy egy fél-automatikus megoldást ajánljak e folyamatok (ontológia populáció (ontology population) és ontológiagazdagítás) elvégzésére. Különböző információkinyerésen alapuló megközelítések már eddig is használatban voltak ontológia populáció céljából. Az ontológiatanulás, gazdagítás és karbantartás egy folyamatosan zajló és összetett folyamat, amely számos kihívást hordoz magában (Shamsfard & Abdollahzadeh Barforoush, 2003; Wong, Liu, & Bennamoun, 2012; Zouaq, Gasevic, & Hatala, 2011). Kulcsszerepet játszik az ontológiamenedzsmentben, és olyan problémákat kezel, mint a tartalmakban lévő rejtett mintázatok leképezhetősége ontológiákká.

Az ontológia populáció egy érdekes aspektusa, amely nincs megfelelően tárgyalva a szakirodalomban, a redundancia kezelése. Ha egy ontológia nem ellenőrzött módon bővített, (nem vizsgált, hogy a példány létezik-e már az ontológiában), akkor redundáns példányok keletkezhetnek. Éppen ezért, a konzisztencia fenntartása vagy a redundancia felszámolása az ontológia populáció fő problémáit jelentik. Az általam javasolt megoldás mindezen felvetésekre választ ad lexikális források felhasználásának segítségével.

II. A FELHASZNÁLT MÓDSZEREK

Disszertációm kutatási módszertana, a kutatás kivitelezése követi egyrészről a Gazdaságinformatika Doktori Iskola követelményeinek megfelelő hagyományos, kvalitatív és kvantitatív kutatási módszereket, másrészről azokat a módszertani megközelítéseket is, amelyek használhatók a számítástudomány területén végzendő kutatásokban. Egy a megoldandó kutatási feladatra vonatkozó kutatásmódszertani folyamatot határoztam meg, mérhető és ütemezett feladatokkal. Ezen feladatok megvalósítása érdekében probléma felvetéseket és kutatási kérdéseket kellett megfogalmaznom, a klasszikus hipotézis meghatározás helyett.

A Budapesti Corvinus Egyetem Gazdaságinformatika Doktori Iskolája az egyetemen belül a társadalomtanulmányi doktori iskolákhoz tartozik, tudományterületileg az informatika tudományok tudományágba került besorolásra, ebből adódóan a kutatási módszerek némiképp ”hibrid” módon való alkalmazása elfogadhatónak tekinthető.

II.1 A kutatási módszertan

Az fentiekben részletezett kutatási probléma megoldásához olyan kutatásmódszertant határoztam meg, amely egyaránt épít az informatikában megszokott megközelítésekre és a kvalitatív módszerekre is. A kutatásban használt módszertani megközelítésem kombinálja a kvantitatív és kvalitatív kutatási módszereket egymással. A disszertációban alkalmazott főbb kutatási szakaszok:

- **Problémameghatározás:** Ez a fázis magában foglalja a szakirodalom áttekintését a már létező tudáskinyerési technikák és az ontológia gazdagítási megoldások elemzése céljából. Ebben a szakaszban arra is rámutatok, hogy nincs olyan általános keretrendszer a szakirodalom alapján, amely alkalmas lenne a szervezeti folyamatokból való automatikus tudáskinyerésre, majd a kinyert információ gazdagítása után annak ontológia gazdagítási célokra való felhasználására.
- **A szakirodalom feldolgozása:** Ebben a szakaszban a szakirodalom alapos tanulmányozására és a releváns megoldatlan problémák azonosítására is sor kerül. A kutatási szakasz célja a tudáskinyerési módszerekkel, ontológia gazdagítással, fogalom kategorizálással és ontológiatanulás mintáival kapcsolatos szakirodalom részletes elemzése. A disszertációban leírt kutatás szorosan kapcsolódik egy jelenleg zajló informatikai kutatási projekthez is (PROKEX projekt, EUREKA_HU_12-1-2012-

0039), ebből adódóan a projekt követelményeihez illeszkedő technikákat és szakterületeket is figyelembe vettem.

- **Kutatási kérdések megfogalmazása:** A szakirodalmi elemzés után a következő lépés a kutatási kérdések megfogalmazása az általános célkitűzésnek megfelelően. Ez a szakasz meghatározó fontosságú, hiszen ezek a kérdések hajtóerőként működnek a kutatás kezdeteitől egészen a végéig. Négy kutatási kérdés került definiálásra ebben a disszertációban, ahol is az első kérdés képezi a kutatás központi kérdését, míg a másik három ehhez kapcsolódó alkérdések.
- **A kutatás fogalmi keretrendszerének megtervezése:** Ez a keretrendszer segít a kutatási terület kulcsfontosságú problémáinak tisztázásában és feltérképezésében. A keretrendszert egy átfogó ábrával is bemutatom. Itt kerül meghatározásra a szövegbányászati előfeldolgozási technikák összessége a folyamatokból kinyert adatok és a tudáskinyerési technikák számára. Ez a megközelítés már kitér arra is, hogyan használhatóak a szövegbányászati technikák információkinyerésre, hogyan használhatóak különböző automatizált tanulási technikák fogalom gazdagításra a különböző források segítségével, illetve hogy hogyan lehet kiszűrni az irreleváns fogalmakat különböző statisztikai és szemantikus eszközök révén. Ez a keretrendszer hosszas szakértői egyeztetés után alakult ki.
- **Fejlesztés:** A kutatási kérdésekre választ adó ProMine megoldás kifejlesztése az előző tervezési fázisokra épül, lényegében az előző szakaszban kialakított keretrendszer implementálásáról van szó. Ebben a szakaszban egy a keretrendszert leíró, működőképes prototípust fejlesztettem ki JAVA nyelven, (ez egyben a PROKEX projekt egy moduljaként működött). A prototípus fejlesztése során iteratív megközelítést alkalmaztam, a fejlesztés tapasztalatai alapján újabb finomítási, javítási fázis következett. Ezt a javasolt prototípust, a ProMine-t, alkalmazom tudáskinyerésre.
- **Kiértékelés:** Ebben a szakaszban egy olyan kiértékelési módszer került kidolgozásra, amellyel az általam kifejlesztett megoldás minősége és hatékonysága vizsgálható (March & Smith, 1995). Az értékelési szakasz célja a fejlesztési folyamat vizsgálata, kiértékelése. Vizsgáltam, hogy a fejlesztés eredménye megfelel-e az előzetesen várt eredményeknek, illetve hogy milyen kapcsolatban áll a ProMine program terve a valós kimenettel (Balbach, 1999). A kifejlesztett ProMine megoldás értékelésre került a szakterület lefedése és a pontosság szempontjából is. A megoldás teljesítményének mérhetőségére kvalitatív és kvantitatív értékelési módszereket is

használtam. Kvalitatív értékelési módszerként három esetet választottam három különböző területről: 1) élelmiszerlánc-biztonság; 2) biztosítás; 3) IT audit. Ezen területek kiválasztásának oka az összekapcsolódó feladatok komplexitásában és a mindennapi teljesítésük során felmerülő problémákban keresendő. Az általam kifejlesztett ProMine megoldás teljesítményének vizsgálata ellenőrzött teszteken keresztül zajlott az adott szakterületek szakértőinek segítségével, majd ezek az eredmények felhasználásra kerültek egy létező szakterületi ontológiában (STUDIO ontológia).

- **Konklúzió:** Ez a kutatás utolsó fázisa, amelyben leírom a kutatás újszerűségét. A szakirodalom különböző nevekkel illeti ezt a szakaszt, mint például konklúzió, eredmények vagy kommunikáció. A konklúziók bizonyítékkal megfelelően alátámasztottak kell, hogy legyenek, ez esetemben az előző szakasz tesztjein keresztül biztosított. A javasolt megoldás korlátaira és jövőbeli lehetséges kutatási irányokra is rávilágít ez a szakasz.

A fent említett kutatási módszerrel fejlesztettem ki a ProMine prototípust. A ProMine végleges verziójának létrehozása érdekében az úgy nevezett „design science" megközelítés alkalmaztam (March & Smith, 1995). Eszerint a megközelítés szerint két alapvető tevékenység, fejlesztés és kiértékelés, különböztethető meg. Míg a fejlesztés a megoldás specifikus célra való konstrukciójának folyamata, addig az értékelés az a folyamat, amelyben megállapításra kerül a megoldás teljesítménye. Ebben a kutatásban ezen tevékenységek végrehajtása egy ismétlődő jellegű inkrementális kutatástervezési folyamatban történik, amely meghatározott számú iterációt tartalmaz:

- Első iteráció – A központi keretrendszer-fejlesztés végrehajtása, amely magában foglalja kulcskifejezések kinyerését és a WordNet használatával történő fogalomgazdagítást. A fogalmak szűrését statisztikai mértékekkel és információ-nyereséggel hajtottam végre. A megoldás értékelése valós adatok felhasználásával történt, míg a kinyert fogalmak értékelése az értékelési mutatók segítségével zajlott.
- Második iteráció – A fogalomkinyerés eredményeinek javítása érdekében a keretrendszer kiterjesztése, illetve a WordNet-hez a Wiktionary hozzáadása történt meg. A fogalomszűréshez egy szemantikus mérték statisztikai eszközzel való kombinációját hoztam létre. Ez a megközelítés javított az előzetes eredményeken, mindemellett kapcsolatot teremtetett a kinyert fogalmak és üzleti folyamatok között.

 Harmadik iteráció – Ebben az iterációban javítottam a fogalomkinyerési technikán az összetett szóválasztási módszer módosításával, így ennek az iterációnak a fő fejlesztése az a fogalomkinyerési modul, amely egy alapontológia (seed ontology) használatával állítódik elő. Ez a kategorizálási modul támogatást biztosít az ontológia gazdagításában.

 Negyedik iteráció – Ebben az iterációban történt a keretrendszer validálása a kinyerési módszer más területeken való alkalmazásával és értékelésével. A keretrendszer és a ProMine eszköz általános érvényűsége a három különböző terület értékelési mértékeinek összehasonlításán keresztül demonstrálható.

A fent bemutatott iteratív eljáráson keresztül fejlesztettem ki és tesztelem a szövegbányászati prototípust. Három esettanulmányon keresztül mutattam be és validáltam a ProMine megoldást (March & Smith, 1995). Értékeltem a tudáskinyerési módszert és eszközt, a pontosság és a tárgyalt szakterület ontológiamodelljének lefedése szempontjából, felhasználva a ProMine segítségével bővített ontológiából nyert mérőszámokat.

III. A PROMINE: A JAVASOLT KERETRENDSZER ÉS MEGOLDÁS

Az általam kifejlesztett ProMine prototípus három alapvető feladatot támogat. A tudáselemek kinyerését a szakterület dokumentum korpuszából és egyéb forrásokból, majd a kinyert tudáselemekből való fogalomszűrést a terület releváns fogalmainak azonosítása érdekében. A harmadik feladatot a szemantikus fogalomkategorizálás (semantic concept categorization) képezi, amely segíti a szakterületi ontológia gazdagítását és populációját. A ProMine fogalomkinyerés első változatát mutatja be Gillani és Kő cikke (Gillani & Kő, 2014). A ProMine prototípus segítségével bizonyítható a javasolt keretrendszer hatásossága a szemantikus fogalomkinyerés, szűrés és kategorizálás esetében. A ProMine folyamata a következő fázisokból áll:

III.1 Adatkinyerés

A ProMine bemeneti fájlja egy szervezeti folyamat kimeneti fájlja. Ez a bemeneti fájl XML formátumú. A keretrendszer első lépésében (adatkinyerési szakasz) a releváns információ automatikusan kivonásra kerül ebből a bemeneti fájlból a ProMine által. Az adott szöveg bemeneti fájlokból való kivonása után ez a szöveg a folyamat tevékenységeinek megfelelően szöveges fájlokba mentődik el.

III.2 Az adatok előfeldolgozása

A szöveg kinyerése után veszi kezdetét a legkritikusabb szakasz, a kinyert szöveg tisztítása. Az előfeldolgozáshoz tartoznak azok a transzformációk, amelyek strukturálatlan szöveget feldolgozásra alkalmas formátumra alakítják. Ez az előfeldolgozási modul biztosítja az adat előkészítését az ezt követő tevékenységekre, amelyeket a későbbiekben tárgyalok. A szöveg előfeldolgozása szerves része a természetes nyelvű feldolgozásnak (natural language processing, NLP). Magában foglal olyan különböző NLP és szövegbányászati technikákat, mint a tokenizálás, stopszószűrés, „part-of-speech (POS) tagging”, a szótövezés és a fogalmak gyakoriságának számlálása. Ezen technikák alkalmazásával a bemeneti szöveg átalakul fogalom vektorrá, és minden fogalom súlya az adott fogalom bemeneti fájlban jellemző gyakoriságán alapul.

III.3 Fogalomgazdagítás

Az előző szakasz során egyedi kulcsszókészlet került meghatározásra a folyamat minden tevékenységéhez. Ez a fázis két lépésre osztható, elsőként a különböző lexikai forrásokból kinyert szinonimák azonosítása, majd második lépésként összetett kifejezések (szópárok) alkotása a szakterületi korpusz segítségével történik meg.

III.4 Szemantikus hasonlóság mértéken alapuló fogalomszűrés

Az előző szakaszokban az adott területhez nem kapcsolódó fogalmak eltávolításra kerültek egy adott kulcsfogalom szinonim fogalmainak készletéből (Wordnet & Wiktionary). Ennek ellenére az eredményként kapott szólista több száz kifejezést tartalmazott. Ez a szólista magas dimenziójú és ritka vektorként is értelmezhető. Az ajánlott ProMine keretrendszerben a szólistát leszűkítem a releváns fogalmak kiválasztásával, egy fogalomszűrési módszer alkalmazásával. A legtöbb ontológiatanulási eszköz esetében hagyományosan olyan statisztikai mértékeket használnak szűrési folyamatokra, mint például a TF-IDF, RTF, valamint entrópia és valószínűségi mutatók (Cimiano & Völker, 2005). Fontos lexikális kifejezések azonosításához a ProMine egy olyan innovatív megközelítést használ, amely statisztikai és szemantikai eszközöket kombinál. Egy új hibrid szemantikus hasonlósági mérték használatát javaslom az adott szervezeti folyamat szempontjából releváns ontológiai minták azonosítására. Ez a modul két szakaszból épül fel; elsőként minden egyes fogalomjelöltre vonatkozóan a domain korpusz felhasználásával kiszámolásra kerül a belőlük származó információ-nyereség. A második szakaszban a szemantikailag jobban reprezentatív fogalomjelöltek megtalálása érdekében egy hibrid szemantikus hasonlósági mérték használatát javasolom, amely mérték különböző információforrásokra épít, mint például lexikai szemantikus hálózatok (WordNet) és a domain korpusz.

III.5 Fogalmak szemantikus kategorizálása

ProMine architektúra lényegében két fő feladat ellátására szolgál; az egyik a fogalomkinyerés, míg a másik a fogalmak szemantikus kategorizálása. A harmadik szakasz végére már rendelkeztem a domain specifikus fogalmak egy javított listájával minden egyes kulcskifejezésre (a kulcskifejezések a második szakaszban kerültek

kiválasztásra). Ezek a fogalmak szemantikailag hasonlóak a kulcskifejezéshez. A következő lépésben az ontológia gazdagítási elveinek megfelelően kategorizálom ezeket a fogalmakat. Emiatt a fogalmak kapcsolatának feltérképezése is szükséges e szavak és a már létező (mag)ontológiák között. A fogalmak szemantikus kategorizálásának egy újfajta módszerét ajánlom a létező ontológia gazdagítása érdekében. Ez a módszer képes új domain specifikus fogalmak osztályozására a magontológia már létező struktúrájának megfelelően. A fogalomkategorizálásra ez a módszer az ontológia már létező fogalomkategóriáinak struktúráját (az osztályok taxonómiája) használja, és teszi mindezt külső tudásforrások segítségével, mint például a Wiktionary. Az ajánlott megközelítés az ontológia egy adott részének felhasználásával talál szemantikus hasonlóságot a kinyert fogalmak között.

IV. A KUTATÁS TUDOMÁNYOS EREDMÉNYEI

A disszertációmban a szövegbányászat üzleti folyamatokhoz kapcsolódó egy új paradigmáját mutattam be. A ProMine szövegbányászati keretrendszer és megoldás megtervezésével és kifejlesztésével igazoltam kutatási állításaimat. E szövegbányászati megoldáson keresztül azonosítom és kinyerem az üzleti folyamatokban rejtőző tudáselemeket, majd kontextusba (domain ontológia) helyezem őket a tudássá alakításukhoz.

Három esettanulmányon keresztül válaszoltam meg a kutatási kérdéseket. Az első esettanulmány megmutatta, hogy a ProMine fogalomkinyerési megközelítés képes a domain szakértők és ontológiafejlesztők támogatására a szakterület specifikus ontológiák hatékony felépítésében. Mindazonáltal a végső listában szereplő fogalmak számának csökkentése és relevanciájának növelése érdekében néhány egyéb módszert is alkalmazok az információ-nyereség mellett. Az információ-nyereség kritériumának gyengeségei statisztikai jellegéből adódnak (Novaković, Štrbac, & Bulatović, 2011), ahogyan azt a dolgozatban részletesen kifejtem. Ezért a jobb eredmények elérésének érdekében szemantikai mértékeket használok. Így egy új hibrid hasonlóságmértéket (Cloud Kernel becslés) ajánlok a fogalmak szűrésének folyamatához.

A második esettanulmány a javasolt hasonlósági mértékhez kapcsolódik. A fogalomkinyerésen túl a ProMine foglalkozik a releváns fogalmak rangsorolásával és szűrésével is az új hibrid hasonlóságmérték segítségével. A ProMine tudáskinyerési megoldás újszerűsége az alábbiakban foglalható össze: 1) kismértékű, a szervezeti folyamatokba beágyazott leírásból képes fogalmakat kinyerni, és külső források felhasználásával gazdagítja ezt a tudást, valamint nagyszámú új fogalmat nyer ki automatikusan, emberi beavatkozás nélkül. 2) Fogalomszűrési megközelítése komplex szintaktikai és szemantikai elemzést hajt végre a fontos, releváns fogalmak kiszűrésére. A fent javasolt új hibrid hasonlóságmérték alkalmazható a mesterséges intelligencia egyéb területein, a pszichológiában és a kognitív tudományokban is. Ez az esettanulmány azt is bemutatja, hogy a ProMine teljesítménye szakértői értékelés szerint is meggyőző, mivel az alkalmazásával kapott eredmények azt mutatják, hogy számos új releváns fogalom került azonosításra és kinyerésre és később az ontológia populáció során alkalmazásra.

A harmadik esettanulmányban a ProMine ontológiagazdagítási képességét teszteltem

egy ontológia alapú e-learning környezetben, amelyet informatikai auditorok tréningjein, a CISA előkészítő kurzusain használnak. Egy ideális magas pontossággal (accuracy) és felidézési (recall) értékkel rendelkező rendszer nagyszámú eredményt szolgáltat, ahol az összes eredmény helyesen címkézett. Adott számú tesztelési ciklus után a ProMine rendszer kellő pontosságot mutatott a kapcsolódó szakterületi ontológia gazdagítása esetében. Ha mind pontossággal és felidézési értékekkel is foglalkozunk, azt tapasztaljuk, hogy a vizsgált esetben mindkettő fokozatosan emelkedik. Mindkét mutató magas értékei azt mutatják, hogy a javasolt kategorizálási módszer pontos értékeket ad vissza (magas pontosság), illetve hogy a pozitív eredmények többségét képes visszaadni (magas felidézési érték). A teszteset, mind a pontossági, felidézési értékeket, valamint a szakértői értékelést alapul véve igazolta, hogy a ProMine megoldás meggyőzően működik az új fogalmak kategorizálása során.

A disszertáció főbb hozzáadott értéke és tudományos eredményei a következőek:

- A legfontosabb eredmény egy általános szövegbányászati megoldás és keretrendszert megtervezése és kifejlesztése, amely összekapcsol két különböző megközelítést; nevezetesen a folyamatmodellezést, amely procedurális jellegű, és a kontextust vagy ontológiát, amely deklaratív természetű.
- A kutatás egyik fő eredménye a fogalomkinyerési megoldás és annak gazdagítása olyan források segítségével, mint a WordNet, Wiktionary és szakterületi korpusz. A legújabb szövegbányászati és NLP technikákat is felhasználom az információkinyerési megoldásomban.
- További eredmény egy a releváns tudás azonosítására szolgáló módszer kifejlesztése kulcsfogalmak és összetett szavak azonosításán keresztül.
- A fogalomszűrés számára egy új módszert javasoltam, amely statisztikai és szemantikus mértékek kombinációját használja. Ezen a folyamaton keresztül bármely szakterülethez szemantikailag releváns fogalomlista állítható elő. Ez az ajánlott módszer a hibrid hasonlóságmérték.
- További fontos eredmény a fogalomkategorizálási módszer az ontológia populáció, gazdagítás számára.
- Szövegbányászati technikák újszerű integrációját valósítottam meg a szervezeti folyamatokban lévő tudáselemek azonosításához és kinyeréséhez, és ezen tudáselemeknek a szakterületi ontológiában való elhelyezéséhez.

A ProMine megoldás és módszer többféle módon továbbfejleszthető, mind a

fogalomkinyerési eljárás, mind a szűrési eljárás kínál erre lehetőséget. A fogalmak közötti relációk kiterjesztése, valamint további szakterületi ontológiák alkalmazása a tesztelésre lehet egy újabb fejlesztési irány.

V. FELHASZNÁLT IRODALOM

Balbach, E. D. (1999). Using case studies to do program evaluation. *Sacramento, CA: California Department of Health Services*.

Buitelaar, P., Olejnik, D., & Sintek, M. (2003). *OntoLT: A protege plug-in for ontology extractionfrom text.* Paper presented at the Proceedings of the International Semantic Web Conference (ISWC).

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., & Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial intelligence in medicine, 33*(2), 139-155.

Cimiano, P., & Völker, J. (2005). Text2Onto *Natural language processing and information systems* (pp. 227-238): Springer.

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Information Processing & Management, 43*(6), 1705-1714.

Fortuna, B., Grobelnik, M., & Mladenic, D. (2007). *OntoGen: semi-automatic ontology editor*: Springer.

Gal, A., Modica, G., & Jamil, H. (2004). *Ontobuilder: Fully automatic extraction and consolidation of ontologies from web sources.* Paper presented at the Data Engineering, 2004. Proceedings. 20th International Conference on.

Gelfand, B., Wulfekuler, M., & Punch, W. (1998). *Automated concept extraction from plain text.* Paper presented at the AAAI 1998 Workshop on Text Categorization.

Gillani, S. A., & Kő, A. (2014). Process-based knowledge extraction in a public authority: A text mining approach *Electronic Government and the Information Systems Perspective* (pp. 91-103): Springer.

Hassanpour, S., O'Connor, M. J., & Das, A. K. (2013). A semantic-based method for extracting concept definitions from scientific publications: evaluation in the autism phenotype domain. *J. Biomedical Semantics, 4*, 14.

Kok, S., & Domingos, P. (2008). Extracting semantic networks from text via relational clustering *Machine Learning and Knowledge Discovery in Databases* (pp. 624-639): Springer.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems, 15*(4), 251-266.

Morita, T., Fukuta, N., Izumi, N., & Yamaguchi, T. (2006). DODDLE-OWL: a domain ontology construction tool with OWL *The Semantic Web–ASWC 2006* (pp. 537-551): Springer.

Mykowiecka, A., Marciniak, M., & Kupść, A. (2009). Rule-based information extraction from patients' clinical data. *Journal of biomedical informatics, 42*(5), 923-936.

Novaković, J., Štrbac, P., & Bulatović, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research ISSN: 0354-0243 EISSN: 2334-6043, 21*(1).

Shamsfard, M., & Abdollahzadeh Barforoush, A. (2003). The state of the art in ontology learning: a framework for comparison. *The Knowledge Engineering Review, 18*(04), 293-316.

Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR), 44*(4), 20.

Wu, F., & Weld, D. S. (2007). *Autonomously semantifying wikipedia.* Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association, 17*(1), 19-24.

Zouaq, A., Gasevic, D., & Hatala, M. (2011). Towards open ontology learning and filtering. *Information Systems, 36*(7), 1064-1081.

VI. A TÉMAKÖRREL KAPCSOLATOS SAJÁT PUBLIKÁCIÓK

2015 **Saira Gillani**, Andrea Kő

Incremental Ontology Population and Enrichment through Semantic-based Text Mining: An Application for IT Audit Domain.

International Journal on Semantic Web and Information Systems (IJSWIS). (Under Second Review).

ISSN: 1552-6283, DOI:10.4018/IJSWIS

2014 **Saira Gillani**, Andrea Kő

Process-based Knowledge Extraction in a Public Authority: A Text Mining Approach

2014 3rd International Conference (EGOVIS 2014) on Electronic Government and the Information Systems Perspective, Springer International Publishing.

ISBN: 978-3-319-10178-1, pp. 91-103, DOI: 10.1007/978-3-319-10178-1

2015 Noureen Zafar, Saif Ur Rehman, **Saira Gillani** and S. Asghar

Segmentation of Crops and Weeds using Supervised Learning Technique.

Improving Knowledge Discovery through the Integration of Data Mining Techniques. Book Chapter, Advances in Data Mining and Database Management (ADMDM) Book Series, IGI Global.

ISBN: 9781466685130, DOI:10.4018/978-1-4666-8513-0

2015 Mohsin Iqbal, Saif Ur Rehman, **Saira Gillani** and S. Asghar

An Empirical Evaluation of Feature Selection Methods.

Improving Knowledge Discovery through the Integration of Data Mining Techniques., Book Chapter, Advances in Data Mining and Database Management (ADMDM) Book Series, IGI Global.

ISBN: 9781466685130, DOI:10.4018/978-1-4666-8513-0

2013 Muhammad Naeem, **Saira Gillani** and Sheneela Naz,

MfWMA: A Novel Web Mining Architecture for Expert Discovery

International Journal of Advanced Science and Technology (IJAST).

ISSN: 2005-4238, Vol. 52, March, 2013.

2013 **Saira Gillani**, Peer Azmat Shah, Amir Qayyum, Halabi Hasbullah

MAC Layer Challenges and Proposed Protocols for Vehicular Ad-hoc Networks

In Vehicular Ad-hoc Networks for Smart Cities, Springer International Publishing.
ISBN: 978-981-287-158-9, 10.1007/978-981-287-158-9_1, pp. 3-13.

2013 **Saira Gillani,** Farrukh Shahzad, Amir Qayyum and Rashid Mehmood

A Survey on Security in Vehicular Ad hoc Networks

In LNCS series (Springer) Nets4Cars-Nets4Trains
ISBN: 978-3-642-37974-1, DOI:10.1007/978-3-642-37974-1_5

2013 **Saira Gillani,** Muhammad Naeem, Raja Habibullah and Amir Qayyum

Semantic Schema Matching Using DBpedia

International Journal of Intelligent Systems and Applications (IJISA)
ISSN: 2074-9058, Vol. 5, No. 4

2012 **Saira Gillani**, Shahnela Naz, Muhammad Naeem, Tanvir Afzal, Amir Qayyum

ErraGMap: Visualization Tool

In 8th International Conference on Digital Content Technology and its Applications IDCTA2012, Jeju Island, Korea.

2008 **Saira Gillani,** Imran Khan**,** Shahid Qureshi, Amir Qayyum

Vehicular Ad hoc Network (VANET), Enabling Secure and Efficient Transportation System

In Technical Journal, University of Engineering and Technology, Taxila, Pakistan. Dec.
ISSN: 1813–1786, Volume 13.