Közgazdasági és
Gazdaságinformatikai
Doktori Iskola

TÉZISGYŰJTEMÉNY

Molnár Géza Gábor

Szemantikus web technológiák alkalmazási lehetőségei az „exploratory” OLAP-ban

című Ph.D. értekezéshez

TÉMAVEZETŐ:

Dr. Kő Andrea, Ph.D.
egyetemi tanár

BUDAPEST, 2022

Számítástudományi tanszék

TÉZISGYŰJTEMÉNY

Molnár Géza Gábor

Szemantikus web technológiák alkalmazási lehetőségei az „exploratory” OLAP-ban

című Ph.D. értekezéshez

TÉMAVEZETŐ:

Dr. Kő Andrea, Ph.D.
egyetemi tanár

Tartalomjegyzék

1 KUTATÁSI ELŐZMÉNYEK ÉS A TÉMA INDOKLÁSA

Az utóbbi évtizedben a strukturálatlan adatok (pl. szövegek) mennyisége sokkal nagyobb mértékben nőtt, mint a strukturált adatoké. Ez nem kis mértékben a közösségi médiának és az okos eszközök elterjedésének köszönhető. Így fordulhat elő, hogy manapság már jóval több a nem strukturált adat a strukturáltnál (Phoebe Wong, 2019). A becslések szerint 2022 végére részesedésük az összes adatból 93%, illetve 7% körül lesz.

A nagy mennyiségű adat tárolása egyelőre nem okoz gondot, mivel a tároló eszközök ára az utóbbi időben jelentősen csökkent, kapacitásuk viszont közben növekedett. Egy GB adat tárolási költsége így a töredékére csökkent. A jelenleg alkalmazott tárolási technológiák ugyan néhány év múlva várhatóan elérik teljesítőképességük határát, de már most megjelentek olyan jövőbe mutató fejlesztések (pl. kvantum tárolók), amelyek képesek lesznek lépést tartani a megnövekedett tárolási kapacitás igényekkel (Klein, 2017).

A strukturálatlan adatok feldolgozásával viszont más a helyzet. Ezeket korábban manuálisan dolgozták fel, vagy egyszerűen figyelmen kívül hagyták. A növekedő részarányuknak köszönhetően viszont egyre többször merül fel ezek elemzésének igénye. Erre tipikus példa, amikor egy cég szeretné az ügyfeleinek visszajelzéseit vagy egy partneréről megjelenő szöveges információkat összesíteni, kiértékelni. A nagy adatmennyiség miatt sokszor reménytelen, hogy hasonló feladatokat manuálisan végezzünk el. Ezért manapság különösen fontos, hogy olyan számítógépes eljárásokat dolgozzunk ki, amely lehetővé teszi a strukturálatlan – többnyire szöveges - adatok feldolgozását és elemzését.

A strukturálatlan adatok feldolgozása, mint probléma sok részterületet érint, többek között a tudásreprezentáció, a szemantikus technológiák, az adat-, web-, és szövegbányászatot, a mesterséges intelligencia részterületeit ezen belül a tanuló algoritmusok témáját is. A struktúra alkotás fontos eszközei a szemantikus technológiák, ezek közül az ontológiák, amelyek egy adott terület fogalmi leírását adják meg (Gruber, 1993). Segítségükkel egy adott területre jellemző fogalmak és a közöttük lévő kapcsolatok definiálhatók.

1.1 A kutatás célja

Kutatási témám a szemantikus technológiák (elsősorban ontológiák) adattárházakban való felhasználásával kapcsolatos. Elsődleges célom olyan megoldások vizsgálata, amelyek lehetővé teszik a strukturált és nem strukturált adatok együttes elemzését az adattárház/üzleti intelligencia rendszerekben. Az ilyen rendszereket a szakirodalomban exploratory OLAP rendszereknek is nevezik.

Az exploratory OLAP rendszerek területe meglehetősen új, a legelső modellek a szakirodalomban alig több, mint tíz éve jelentek meg. Fontos megjegyezni, hogy a legtöbb ilyen modell megmaradt koncepcionális szinten. Néhány esetben készült ugyan prototípus, viszont azok általános alkalmazhatósága még nem igazolt. Emiatt céljaim közé tartozik még legalább egy prototípus elkészítése, és értékelése. A működő prototípusok segítségével következtetéseket lehet levonni az adott exploratory OLAP modell megvalósíthatóságára és alkalmazhatóságára is. A prototípus elkészítésénél a szemantikus web technológiák közül elsősorban az ontológiákat szeretném használni. Mivel ezek az általam választott területen (ticketing rendszer) nem állnak rendelkezésre, ezért további cél a szükséges szakterületi ontológia kifejlesztése.

1.2 Kihívások

Az exploratory OLAP modell, illetve prototípus kidolgozásánál több nehézséggel is meg kell küzdeni. Ezek egy része minden megoldást érint, más részük az ontológiákhoz köthető.

- Első probléma, hogy az adott területet érintő szakterületi ontológiák általában nem állnak rendelkezésre. Ezek – lehetőleg automatikus – előállítása a nem strukturált forrásrendszerekből nem triviális feladat.
- A második nehézség az OLAP kocka tervezésének egy fontos lépésénél, az aggregációk (összegzett adatok) tervezésénél adódik. Ennek megoldására egyelőre csak előzetes kutatási eredmények vannak.
- A harmadik probléma az ontológiák változásával kapcsolatos (evolúció, verziókezelés). Jelenleg ennek kezelése is még gyerekcipőben jár.
- A következő kihívás az ETL folyamatot érinti. A strukturálatlan adatok integrálása az adattárház ETL folyamatai közé nem egyszerű, ugyanis számos olyan elemet tartalmazhat, amelyet a strukturált adatokat kezelő folyamatok nem ismernek. Például nem relációs operátorok, gépi tanulást alkalmazó számítások, összetett adattípusok stb.
- A szemantikai réteg integrálása az ontológiák segítségével performancia problémákat vethet fel, amelyeket figyelembe kell venni az elkészült terv optimalizálásánál is.
- Az exploratory OLAP rendszernek képesnek kell lennie dinamikusan integrálni az adatforrásokat Ez olyan magas számítási költségeket eredményezhet, amelyek akár a kivitelezhetőséget is veszélyeztethetik.
- Az exploratory OLAP automatikus hozzáférést jelent nem csak a sémához, hanem bizonyos mértékben az adatokhoz is. Ez maga után vonja azt, hogy példány szinten is képesnek kell lenni az adatok értelmezésére és az érvelésre.

- Végül az utolsó nehéz kérdés az, hogy az alkalmazott szemantikus web technológia mennyiben alkalmas a vezetői döntés támogatásra. Itt a nemstrukturált adatok integrálása és az egyes adatforrások függetlensége okozza a fő problémát (Abelló, 2015).

1.3 Kutatási kérdések

Kutatásom során a következő kérdések megválaszolását tűztem ki célul:

- Hogyan tehető szemantikussá a hagyományos OLAP és az adattárház, vagyis hogyan lehet megtervezni a szemantikus réteg integrációját az adattárházba?
- Hogyan lehet a megtervezett modellt a gyakorlatban megvalósítani egy prototípus segítségével?
- Mi lehet egy hatékony validálási módszer az „exploratory OLAP" modell esetében?

2 A FELHASZNÁLT MÓDSZEREK

Ebben a fejezetben a kutatás során alkalmazandó módszertan kerül ismertetésre. Ide tartozik az információs rendszerek tervezése és a szoftverfejlesztés során használt design science, az adatgyűjtés és adatelemzés, az ontológia fejlesztési és értékelési módszertan, valamint a prototípus készítése során alkalmazott fejlesztési módszertan.

2.1 Design science

A design science („tervezéstudomány") olyan összegző és elemző technikák, nézőpontok halmaza, amelyek IT-kutatások során hatékonyan alkalmazhatók (Vaishnavi, Kuechler, & Petter, 2004). A cél az adott probléma megértése az információs rendszerek nézőpontjából. Ez rendszerint két alapvető tevékenység segítségével történik:

- Új ismeretek szerzése egy innovatív alkotás (artifaktum) létrehozásán keresztül
- Az artifaktum használata során kapott visszajelzések elemzése

Kutatásom esetén az artifaktum az exploratory OLAP prototípus, illetve a ticketing ontológia. Az artifaktum tervezése és vizsgálata mindig egy adott összefüggésben, környezetben (kontextus) történik. A kontextus esetünkben a ticketing rendszer.

Egy design science projekt mindig iteratív tevékenység, amely két fő tevékenysége a tervezés és a vizsgálat. A tervezési tevékenység három részre osztható. Ezek a következők: a probléma vizsgálata, a probléma kezelésének megtervezése és a kezelés validálása. Ezen három résztevékenység ismétlését tervezési ciklusnak (design cycle) nevezzük.

A tervezési ciklus egy nagyobb ciklus része, amelynek során a tervezési ciklus eredményét (a validált probléma kezelést) átadják a felhasználóknak, akik ezt használják, illetve értékelik. Ezt a nagyobb ciklust fejlesztési ciklusnak (engineering cycle) nevezzük (Wieringa, 2014).

A ciklus a következő lépésekből áll:

- A probléma vizsgálata. Milyen jelenségeken kell javítani? (Pl: hogyan lehet növelni a szöveges adatok kezelésének hatékonyságát az adattárházakban)
- A probléma kezelő eljárás tervezése. Egy vagy több artifaktum tervezése a probléma kezelésére. (Pl: exploratory OLAP prototípus tervezés)
- A probléma kezelő eljárás validálása. Ezek a tervek megoldják a problémát?
- A probléma kezelő eljárás implementálása. A probléma kezelése az egyik artifaktummal. (Pl: egy exploratory OLAP prototípus megvalósítása)

- A megvalósított eljárás értékelése. Milyen sikeres a probléma kezelése? (Pl: az elkészült prototípus mennyire felel meg a követelményeknek?) Ez a lépés sok esetben egy új iteráció kezdetét jelenti.

Ezt a ciklust addig kell ismételni, amíg el nem jutunk a kívánt eredményig.

2.2 Adatgyűjtés és elemzés

Az első adatgyűjtési metódus a szisztematikus szakirodalmi áttekintés kiegészítése az adott területre (ticketing rendszerek) vonatkozóan. Ez segít a kutatási kihívások megfogalmazásának pontosításában, az exploratory OLAP fogalmi keretrendszerének kialakításában, valamint a prototípus létrehozásában is. Az adatgyűjtés során azonosítom a nem strukturált szakterületi adatforrásokat és létrehozom a kezdeti ontológiát.

A következő módszer az összegyűjtött adatok vizsgálata. Ide tartozik az adatminőség elemzése, amelynek során elsősorban a következőket lehet ellenőrizni:

- Hiányzó értékek (pl: nem kitöltött mezők)
- Kiugró értékek (pl: az átlagosnál jóval hosszabb vagy esetleg csak két karakterből álló megjegyzés mező)
- Duplikációk
- Nem megfelelő értékek (pl: szótárban nem szereplő szavak, rövidítések stb.).

Az adatminőség és az adatok vizsgálata alapján szükséges lehet az adatok tisztítása abból a célból, hogy a rendszerbe ne kerülhessenek be nem megfelelő minőségű adatok.

2.3 Ontológia fejlesztés

Bár napjainkban számos ontológia fejlesztési módszertan ismert, ezek közül kevés a széles körűen elfogadott és kellően érett megoldás (Iqbal, Murad, & Mustapha, 2013). Kutatásom során ezek közül – figyelembe véve előnyeiket és hátrányaikat, illetve a konkrét feladathoz való alkalmasságukat - választottam ki az alkalmazandó ontológia fejlesztési módszertant.

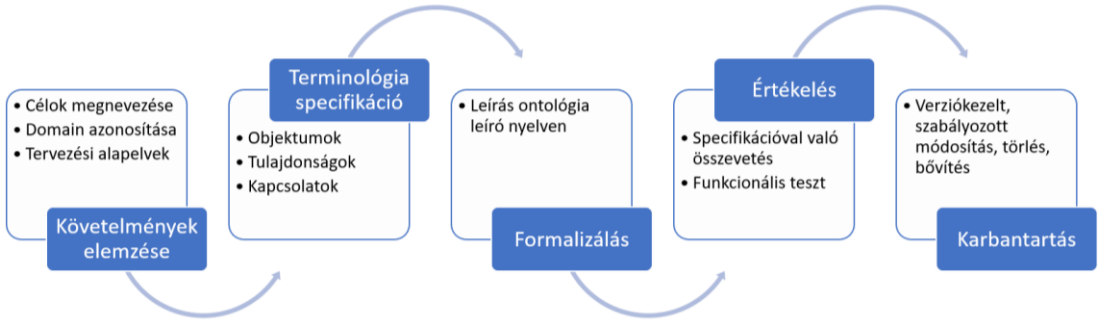
Elsősorban a következő módszertanokat vetttem számításba (Gomez, Corcho, & Fernandez-Lopez, 2002):

- TOVE (TOronto Virtual Enterprise): egy projekt keretében készült. A célja a vállalati működés modellezése és a vállalati integráció leírása volt. Az eredmény felfogható egy második generációs szakértői rendszernek is. A projektből kialakult ontológia módszertant később más célokra (pl: ellátási láncok) is felhasználták.

- CommonKADS: gyakran használt módszertan tudásalapú rendszerek létrehozására. Más módszertanokkal ellentétben ez a megközelítés szorosan kapcsolódik az objektumorientált programozásban használt UML jelölési módszertanhoz.
- SENSUS: a WordNet (az angol nyelv lexikai adatbázisa) kibővítésével készült ontológiára épülő módszertan.
- On-To-Knowledge: egy EU-s projekt keretében létrejött módszertan. A cél nagyszámú, heterogén strukturálatlan vagy félig strukturált dokumentumra épülő tudásbázis létrehozása volt. A dokumentumok nagyvállallatok intranetjeiről és a világhálóról származtak (Iqbal, Murad, & Mustapha, 2013).

A vontakozó szakirodalom áttekintése után a ticketing ontológia fejlesztésénél az On-To-Knowledge módszertan leegyszerűsített változatát választottam, mivel ennél a fogalmak felfedezése van a fókuszban. A felfedezett fogalmak segítenek majd később a multidimenzionális azonosítók, a dimenziók, tények és mértékek előállításához.

A módszertan lépéseit a következő ábra mutatja. Az egyes lépések között szükség esetén visszalépésre is van lehetőség.



1. ábra: Az ontológia fejlesztés lépései az On-To-Knowledge módszertan alapján (saját szerkesztés)

2.4 Ontológia értékelés

Az ontológia értékelésre egy kritérium alapú értékelési módszert választottam. Az értékeléshez, azaz a konkrét kritériumok teljesülésének megvizsgálásához itt a következő kérdéseket kell feltenni (Pâslaru-Bontaş, 2007):

- Tartalmi szempontból releváns-e a kapott ontológia modell az adott területen (esetünkben a ticketing rendszer)?
- Milyen minőségű a kapott ontológia mögötti fogalmi modell terület független nézőpontból, azaz milyen hatékony a tudás ábrázolás?

- Technikai szempontból mennyire felel meg a kapott ontológia figyelembe véve a legújabb technológiákat és eszközöket?
- Mennyire használható a kapott ontológia egy adott gyakorlati feladat esetén?
- Felhasználható-e az ontológia más alkalmazásokban?

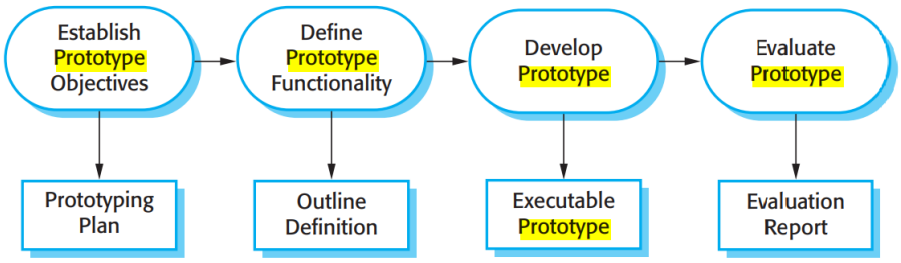
Ezen kritériumokat ellenőriztem a ticketing ontológia második verziójának elkészítése után.

2.5 Rendszerfejlesztés

A prototípus kifejlesztése felfogható egy rendszerfejlesztési, azon belül szoftverfejlesztési feladatnak is, emiatt célszerű ezt egy megfelelően kiválasztott szoftverfejlesztési módszertan szerint végezni (Bánné Varga, 2012). Ezekben a módszertanokban az a közös, hogy a fejlesztés során ugyanazokat a tevékenységeket (részfolyamatok) hajtják végre, csak más módon és megközelítésben. A szoftverfejlesztés részfolyamatai a következők (Sommerville, 2011):

- Specifikáció. Ennek során az ügyfelek és a szoftver fejlesztők pontosan definiálják az előállítandó szoftver feladatait és a működésére vonatkozó korlátozásokat.
- Szoftverfejlesztés. Ez a részfolyamat a szoftver tervezését és elkészítését jelenti.
- Validálás. Annak ellenőrzése, hogy a szoftver a felhasználó igényeinek megfelelően készült-e el.
- Evolúció. A szoftver utólagos módosítása a változó felhasználói vagy piaci igényeknek megfelelően.

Számos szoftver fejlesztési módszertan létezik, a prototípus elkészítéséhez ezek közül az un. rapid prototípus módszert választotttam (Luqi, 2002). Ez a prototípus gyors, iteratív előállítását helyezi előtérbe, ezáltal kézzelfogható eredményeket szolgáltat már a szoftverfejlesztés korai fázisában. A módszernek megfelelően, Prasad (2010), illetve Abello (2015) ötleteit felhasználva hoztam létre prorotípusokat. Egy adott prototípus létrehozásának lépéseit a következő ábra mutatja:



2. Ábra: A prototípus elkészítésének lépései

3 AZ ÉRTEKEZÉS EREDMÉNYEI

Ebben a fejezetben a kutatás eredményeit ismertetem. Ezek közül kiemelném a ticketing ontológia elkészítését, valamint az exploratory OLAP prototípusokat, amelyekre nem találtam példát a szakirodalomban. Az általam készített ticketing ontológia változat nem fedi le az egész ticketing területet, elsősorban a hibabejelentésekre (incidensek) fókuszál.

Az exploratory OLAP prototípusok közül is nagyon kevés működő változat van a gyakorlatban. A kutatás során általam létrehozott prototípusok felhasználnak meglévő modelleket (Prasad, Abello), de azokat átalakítva, és a szükséges mértékig leegyszerűsítve.

3.1 Exploratory OLAP modellek vizsgálata

A szakirodalomban három releváns modellt találtam, ezek Prasad (2010), Abello (2015), illetve Ibragimov (2015) nevéhez köthetők.

Prasad (2010) modelljében a nem strukturált adatokat tartalmazó adatforrásokból kinyert szövegre először szövegelemzési eljárásokat alkalmaz, majd az elemzések eredményét az adatbázis/adattárház megfelelően kialakított tábláiban helyezi el. A ténytábla és a dimenziótáblák csillag elrendezésűek, ezért alkalmasak OLAP kockák, illetve riportok létrehozására.

Abello és társai (2015) exploratory OLAP koncepciója szemantikus web technológiákat használ. A modell legfontosabb része az ontológia-alapú tudásreprezentáció. Az ontológiák lehetővé teszik a tények azonosítását. Mindegyik tény esetén azonosíthatók továbbá a lehetséges dimenzionális fogalmak, amelyek egy jól elrendezett hierarchikus formában vannak tárolva funkcionális függőségek segítségével.

Ibragimov (2015) adatforrásként Linked Open Data (LOD) adatokat használt, amelyek RDF formátumban vannak tárolva. Az általa javasolt rendszer négy modulból áll:

- Global Conceptual Schema – az adatkockáról tárol információkat
- Semantic Query Processor – MDX lekérdezésekből állít elő SPARQL lekérdezéseket
- Distributed Query Processor – lekérdezi a végpontokat, összegyűjti az adatokat
- Source Discovery Schema Builder – a felhasználókkal tartja a kapcsolatot a séma létrejötte alatt.

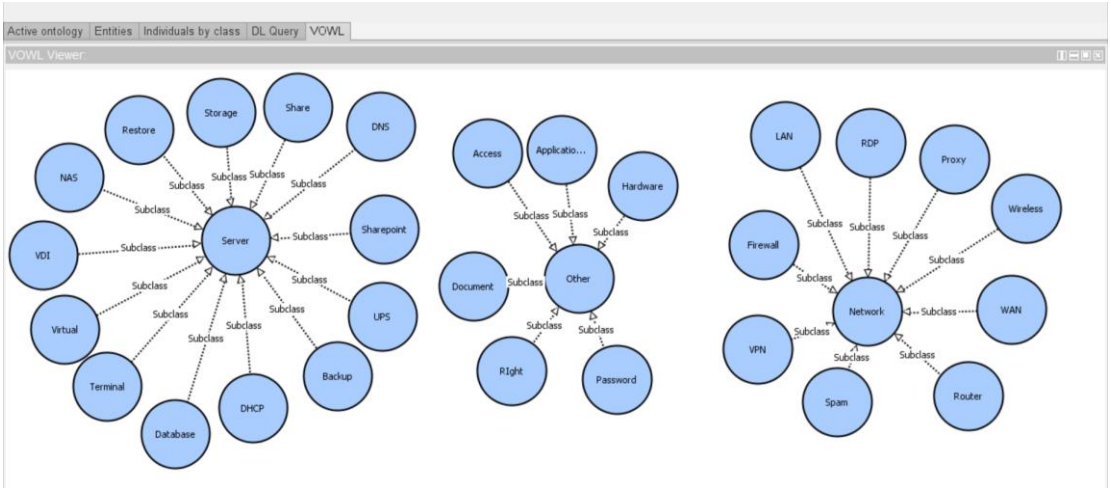
A három exploratory OLAP modell összehasonlítását a következő táblázat mutatja:

Szempont	Abello	Ibragimov	Prasad
Adatforrások (nem strukturált)	**tetszőleges**	**Linked Open Data**	**tetszőleges**
Eszközök, technológiák	**ontológiák**	**RDF, MDX, SPARQL**	**szövegelemzés, XML**
Legnagyobb kihívás	**ontológia –-> MD**	**MDX → SPARQL**	**szövegelemzés**
Létezik-e prototípus terv	**nem**	**igen**	**igen**
Kimenet helye	**kocka**	**kocka**	**adattárház**
Kimenet formája	**OLAP séma**	**OLAP séma**	**Csillag séma**

1. táblázat: Az exploratory OLAP modellek összehasonlítása

3.2 Ticketing ontológia

A ticketing ontológiákból két verziót is készítettem. Az első verzió egy kétszintű taxonómia, amely elősegíti a szemantikus keresést az incidensekben. Első körben az incidensek témájának azonosítása, kategorizálása a leglényegesebb feladat. Ezt a feladatot a LDA (Latent Dirichlet Allocation) modell segítségével oldom meg (Revert, 2018). Ez egy olyan felügyelet nélküli gépi tanulási modell, amely szöveges dokumentumokhoz témákat rendel. Ezen kívül a modell azt is megmutatja, hogy a témák az egyes dokumentumokban milyen mértékben (százalékban) vannak jelen. A kapott eredményt a következő ábra szemlélteti:



3. ábra: A ticketing ontológiák kezdeti rendszere (részlet)

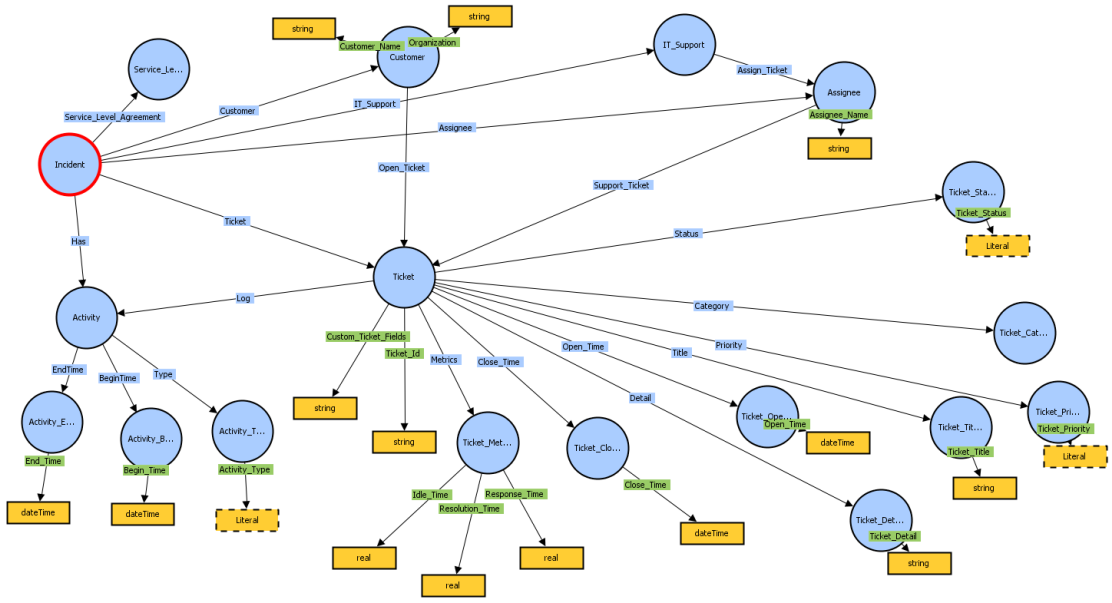
Az ontológia második verziójában (erre a második prototípus miatt volt szükség) a követelmények bővültek. A legfontosabb új követelmény, hogy az ontológia ne csak a hibajegyek kategorizálására legyen alkalmas. Lehetőség szerint minél jobban írja le a ticketing rendszer működését, azon belül is a hibajelentések (incidensek) kezelését az alábbi pontok alapján:

- Az ügyfél szöveges módon jelezi a problémát a helpdesknek, amivel megnyit egy hibajegyet.
- A hibajegy szövege egyrészt egy rövid megfogalmazásból (tárgy), másrészt egy részletesebb leírásból áll. A hibajegy emellett egyéb jellemzőkkel bír. Ilyen például a ticket létrehozásának ideje, vagy a hiba sürgőssége (prioritás).
- A ticket egy hozzáértő személyhez (assignee, analyst) lesz rendelve, aki megoldja a problémát. A megoldás során végzett tevékenységek rögzítésre (logolás) kerülnek.
- A megoldás során a ticket bizonyos jellemzői (pl. a státusza) többször is megváltozhatnak.
- Nem cél a ticket teljes életútjának leírása, itt csak a ticket kezdeti és végső állapota lényeges.
- Legyen lehetőség néhány mérőszám (mérték, KPI) ábrázolására is, pl. mennyi egy hibajegy átlagos megoldási ideje stb.

A ticketing ontológia második verziójában fontos, hogy az osztály-alosztály kapcsolaton kívül más kapcsolatok, illetve az osztály tulajdonságok is megjelenjenek. Konkrét példányok létrehozása az exploratory OLAP prototípus elkészítéséhez nem szükséges.

A fogalmak meghatározása során alkalmazott módszer („reverse engineering”) lényege az adatforrásokból kinyerhető fogalmak és az adott szakterület (ticketing incidensek) szótárának integrálása (Skoutas & Simitsis, 2007). Az így kapott ontológia un. felhasználó-centrikus szakterületi ontológia, azaz az adott szakterületről csak a vizsgált adatforrások szempontjából szükséges ismereteket tartalmazza. A szöveges adatforrásokból származó fogalmak (kulcsszavak) listája az előző prototípusnál már elkészült, így rendelkezésre állt. A szakterületi fogalomgyűjteményt az Internetről letöltött tematikus források segítségével állítottam össze. az ontológia szempontjából releváns fogalmak kiválasztásához egy Python-script segítségével szemantikus összehasonlítást végeztem a szöveges adatforrásból kinyert kulcsszavak és a szakterületi fogalomgyűjtemény szavai között.

Ezután a kapott eredményből leszűrtem a szakterületi fogalomgyűjtemény azon szavait, amelyekhez található magas (elsősorban 0.75 feletti) hasonlósági indexű kulcsszó az adatforrásokból. A kapott szavakból manuálisan választottam ki, illetve határoztam meg az ontológiában szereplő fogalmakat (osztályokat), és a tulajdonságokat. A kapcsolatokat a követelményekből kiindulva hoztam létre. Az eredményt a következő ábra mutatja:



4. ábra: A ticketing ontológia második verziója

3.3 Exploratory OLAP első prototípus

A prototípus elsődleges célja, hogy a ticketing rendszerből származó szöveges adatok feldolgozásával előálljon a Prasad (2010) által javasolt csillagséma.

A prototípussal szemben további elvárás, hogy valósítsa meg a következő funkciókat:

- Szövegfeldolgozási műveletek (kulcsszavak keresése, és címkézése)
- A kulcsszavak tanulási képessége
- OLAP-séma elkészítése a csillagsémát felhasználva
- Dashboard készítése a kapott eredmény szemléltetésére.

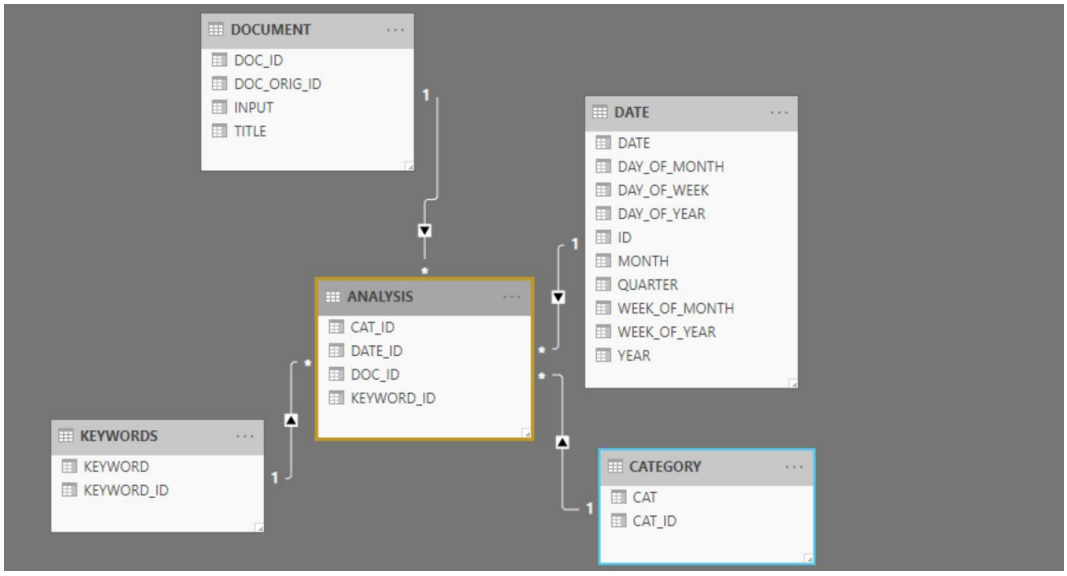
Egyelőre nem cél a félig vagy teljesen automatizált működés. A prototípus létrehozása a Prasad-féle megközelítésre épülő, de attól eltérő fogalmi modell alapján négy lépésben történt:

- Adatok előkészítése
- Szövegelemzések elvégzése
- Csillag séma létrehozása
- OLAP-kocka létrehozása, tesztelése.

A kivitelezés során a következő szoftvereket használtam:

- Service Desk (ticketing) program a forrásadatok exportálására
- Python 3.7 (Jupyter, Anaconda): a szövegfeldolgozó rutinok elkészítésére
- MS SQL Server 2017 Express, SQL Server Management Studio és SQL Server Import és Export varázsló a csillag séma létrehozására
- Power BI Desktop az OLAP-kocka elkészítésére, tesztelésére.

A prototípus a forrásadatokból a következő sémát állította elő:

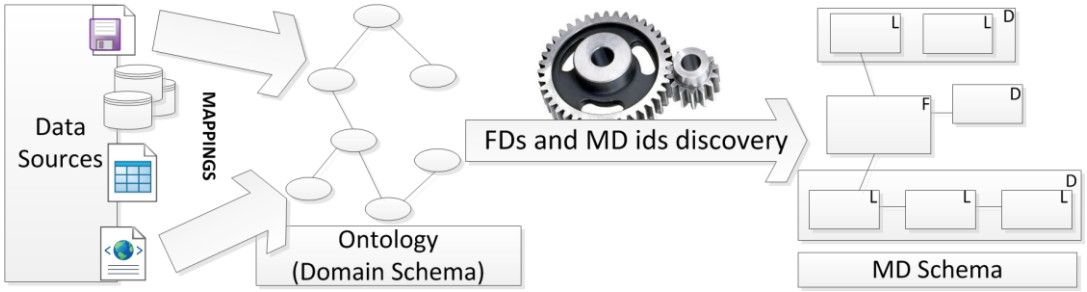


5. ábra: Az első exploratory OLAP prototípus sémája

A prototípus megvalósítása viszonylag egyszerű, viszont alkalmazhatósága az előre definiált végső séma miatt korlátozott.

3.4 Exploratory OLAP második prototípus

Az exploratory OLAP második típusának kifejlesztése a 6. ábra alapján történt.



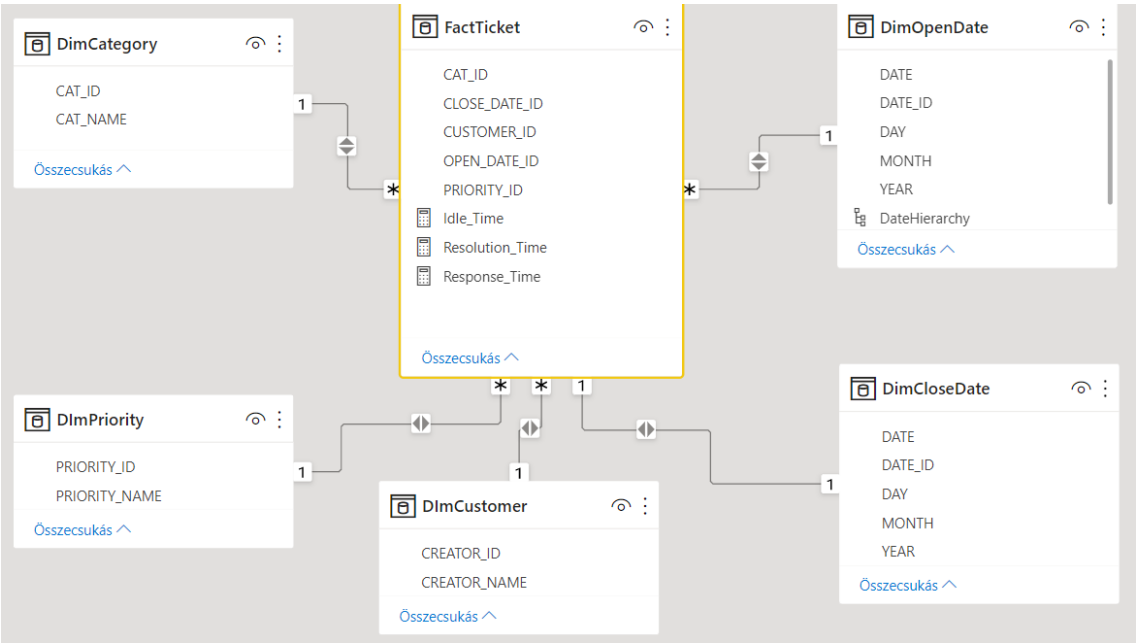
6. ábra: Ontológiák a domain modellezésben

A fejlesztés első lépése az adatok előkészítése volt. A probléma szempontjából releváns forrásadatok a következők:

- A hibajegyek adatai (Azonosító, megnyitás ideje, prioritás stb.)
- Hibajegy kategória adatok
- A hibajegyekhez köthető tevékenységek adatai (Tevékenység kezdetének dátuma, tevékenység típusa stb.)
- A tevékenységeket elvégző informatikusok adatai
- A hibajegyet létrehozó ügyfél adatai
- SLA adatok

Az eredetileg Excel/CSV formátumú adatokat egy relációs sémában helyeztem el. A séma tartalmazza a szöveges adatokat is (hibajegyek tárgya és részletes leírása).

Az adatforrásokból (data sources) jövő adatok egy referencia ontológiára (ld. 3.2) vannak leképezve. Az ontológiára épülő leképezések segítségével azonosíthatók a funkcionális függőségek (FD, Functional Dependency) és a multidimenzionális azonosítók (MD ids). Ezek után a tények és a dimenziók azonosításával előáll az MD séma. Ez utóbbi a következő ábra szemlélteti:



7. ábra: a 2. exploratory OLAP prototípus sémája

Az elkészített prototípus megfeleltethető egy „nulladik” dimenzionális modellnek. Azaz, egy olyan alapot adhat, amelyre építve az OLAP-séma lényegesen könnyebben létrehozható, mintha a folyamatot teljesen az alapoktól kezdenénk.

3.5 A kutatási eredmények összegzése

A következő táblázat a kutatási célokat és az elért eredményeket szemlélteti.

Cél	Eredmény
Exploratory OLAP modellek vizsgálata.	A modelleket összehasonlítottam, majd vizsgáltam a megvalósíthatóságukat és alkalmazási lehetőségeiket. Prasad (2010), illetve Abello (2015) modellje – szükséges módosításokkal és egyszerűsítésekkel – megvalósíthatónak bizonyult. Ibragimov modellje kutatásban preferáltaktól eltérő eszközöket használ, emiatt annak megvalósíthatóságát nem vizsgáltam. A gyakorlatban Abello (2015) modellje a legjobban alkalmazható a három exploratory OLAP modell közül. Lehetővé teszi egy kezdeti dimenzionális modell létrehozását a nyers adatokból kiindulva.
Exploratory OLAP prototípusok elkészítése	Két prototípust is készítettem Prasad (2010), illetve Abello (2015) ötletét felhasználva. Mindegyik egy ticketing rendszer adatait használja fel.
Ticketing szakterületi ontológia elkészítése	Az ontológiából két verzió is készült. Az első egy kétszintű taxonómia, amely az első exploratory OLAP prototípushoz jól alkalmazható. A második prototípushoz szükség volt az ontológia továbbfejlesztésére újabb fogalmakkal, tulajdonságokkal és kapcsolatokkal.

2. táblázat: A kutatási célok és eredmények összefoglalása

3.6 A kutatás jelentősége

Kutatásom segítséget nyújthat:

- Ontológia fejlesztőknek, akiknek egy új szakterületi vagy alkalmazás-specifikus ontológiát kell létrehozniuk teljesen az alapoktól kezdve.
- Adatbázis (azon belül ETL – Extract, Transform, Load) fejlesztőknek, akiknek a feladata a szöveges adatforrások integrálása az adattárházba.
- Adatelemzőknek, akik nem strukturált adatforrásokból szeretnének használható információt kinyerni
- Menedzsereknek, akik együtt szeretnének elemezni strukturált és nem strukturált adatokat.

Az exploratory OLAP rendszerek területe még új és kiforratlan, ezért kutatása külön kihívást jelent. A téma jelentőségét igazolja, hogy egyre több tanulmány és cikk jelenik meg az exploratory OLAP-hoz köthetően. Magyarországon – tudomásom szerint – ez az első kutatás ebben a témában.

4 FŐBB HIVATKOZÁSOK

Abelló, A. (2015). In *Using Semantic Web Technologies for Exploratory OLAP: A Survey.* IEEE. Forrás: https://core.ac.uk/download/pdf/41822169.pdf

Abello, A., & Romero, O. (2010). A framework for multidimensional design of data warehouses from ontologies. Data Knowl. Eng. Vol 69, no. 11.

Bánné Varga, G., 2012. In: Az adattárház készítés technológiája. hely nélk.:Typotext, pp. 19-20, 23, 3.fejezet.

Biemann, C. (2005). Ontology Learning from Text: A Survey of Methods. LDV Forum. Forrás: jlcl.org

Cuzzocrea, A., Bellatreche, L., & Song, I.-Y. (2015). In *Data Warehousing and OLAP over Big Data: Current Challenges and Future Research Directions.* International Journal of Business Process Integration and Management.

Gomez, A., Corcho, O., & Fernandez-Lopez, M. (2002). Methodologies, tools and languages for building ontologies. Data & Knowledge Engineering 46 (2003).

Gruber, T. R., 1993. A Translation Approach to Portable Ontology Specifications. In: hely nélk.:Knowledge acquisition, Vol. 5 No. 2 , pp. 199-200.

Ibragimov, D., Hose, K., Pedersen, T. B., & Zimány, E. (2015). In *Towards Exploratory OLAP Over Linked Open Data – A Case Study.* International Workshop on Business Intelligence for the Real-Time Enterprise. Forrás: https://link.springer.com/chapter/10.1007/978-3-662-46839-5_8

Iqbal, R., Murad, A., & Mustapha, A. (2013). An Analysis of Ontology Engineering Methodologies: A Literature Review. Faculty of Computer Science and Information Technology, Universiti Putra Malaysia.

Kő, A., & Gillani, S. (2019). A Research Review and Taxonomy Development for Decision Support and Business Analytics Using Semantic Text Mining. *International Journal of Information Technology & Decision Making*.

Klein, A., 2017. Hard Drive Cost Per Gigabyte. [Online]
Available at: https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/#:~:text=From%20January%202015%20to%20January,the%20cost%20of%20providing%20storage.

Liang, X. (2018. február). Forrás: https://towardsdatascience.com/textrank-for-keyword-extraction-by-python-c0bae21bcec0

Liu, H., & Wang, P. (2014). Assessing Text Semantic Similarity Using Ontology. *Journal of Software, 9*(2), 490-496.

Luqi, F. K. a., 2002. An Introduction to Rapid System Prototyping. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, 28(9), pp. 817-820.

Nebot, V., Berlanga, R., & Pérez, J. M. (2009). *Multidumensional Integrated Ontologies: A Framework for Designing Semantic Data Warehouses.* doi:10.1007/978-3-642-03098-7_1

Neumayr, B., Anderlik, S., & Schrefl, M. (2012). Towards ontologybased OLAP: Datalog-based reasoning over multidimensional. Proc. 15th Int. Workshop Data Warehousing OLAP.

Pâslaru-Bontaş, E., 2007. A Contextual Approach to Ontology Reuse: Methodology, Methods and Tools for the Semantic Web. In: hely nélk.:REFUBIUM - FREIE UNIVERSITÄT BERLIN, p. 6. fejezet.

Prasad, K. S. (2010). In *Text Analytics to Data Warehousing.* Kalli Srinivasa Nageswara Prasad: Text AInternational Journal on Computer Science and Engineering,. Forrás: https://www.researchgate.net/publication/49941856_Text_Analytics_to_Data_Warehousing

Phoebe Wong, R. B., 2019. Everything a Data Scientist Should Know About Data Management. [Online]
Available at: https://www.kdnuggets.com/2019/10/data-scientist-data-management.html

Revert, F., 2018. An overview of topics extraction in Python with LDA. [Online]
Available at: https://towardsdatascience.com/the-complete-guide-for-topics-extraction-in-python-a6aaa6cedbbc

Romero, O., & Abello, A. (2012). Ontology driven search of compound IDs. Knowl. Inform. Syst., vol. 32.

Skoutas, D., & Simitsis, A. (2007). Ontology-based Conceptual Design of ETL Processes for both Structured and Semi-structured Data. *International Journal on Semantic Web and Information Systems, 3*(4), 1-24.

Sommerville, I., 2011. Software Engineering Ninth Edition. In: hely nélk.:Pearson, pp. 26, 47-52, 73-78.

Vaishnavi, V., Kuechler, B., & Petter, S. (2004). DESIGN SCIENCE RESEARCH IN INFORMATION SYSTEMS. Eds.

Wieringa, R. J. (2014). In *Design Science Methodology for Information Systems and Software Engineering.* Springer.

5 PUBLIKÁCIÓK JEGYZÉKE

2022. május

FOLYÓIRATCIKK

Géza Molnár [2021]: Ticketing Data Warehouse System Development: Challenges and Experiences
In SEFBIS JOURNAL NO.14/2021 pp. 14-24

Géza Molnár [2022]: Decision-Making Through a Self-Service Business Intelligence Solution
In JOURNAL OF APPLIED MULTIMEDIA

KONFERENCIA KÖZLEMÉNY

Géza Molnár [2020]: An Implementation of Exploratory OLAP System Based on Prasad's Approach
In: AIS 2020 PROCEEDINGS, ISBN 978-963-449-209-2, pp. 89-92

KONFERENCIA ABSZTRAKT

Molnár Géza [2019]: Szemantikus web technológiák alkalmazási lehetőségei az exploratory OLAP-ban
In OGIK 2019

Molnár Géza [2019]: Út az adatoktól a riportokig, avagy vezetői információs rendszer kifejlesztése
Microsoft szoftverek használatával – esettanulmány
In MAFIOK 2019, pp. 43

Molnár Géza [2018]: Vezetői információs rendszer kifejlesztése egy informatikai kft számára – esettanulmány
In OGIK 2018, pp. 56

EGYÉB (KÖNYVRÉSZLETEK, KÖNYVFEJEZETEK)