Egyetemi ökoszisztéma vizsgálata big data környezetben [1]

Hornyák Miklós – Kruzslicz Ferenc
Pécsi Tudományegyetem

A TANULMÁNY CÉLJA

Az egyetemi versenyképesség értelmezésére a Triple Helix modell az elterjedt, amelynek egyik fontos befolyásoló eleme az a környezet, amelyben az egyetem működik (Webster & Etzkowitz 2000, Etzkowitz 2003a, Etzkowitz & Klofsten 2005, EU 2011). Az egyetemi versenyelőny egyik meghatározója a kutatói közösség tudáspotenciálja, melynek mérését jellemzően kvantitatív adatokból előállított változók útján valósítják meg (Etzkowitz 2003b, NCEE 2014). Kutatói kérdéseink megválaszolása során kísérletet tettünk az egyetemi szférát és munkavállalóit speciális vállalatként értelmezve többfajú adatforrások (1) és technológiák (2) azonosítására épülő modell kialakítására (3), mely elvezet egy működő rendszer kiépítéséig.

ALKALMAZOTT MÓDSZERTAN

A Pécsi Tudományegyetemen Tudáspark projektjén belül kialakítandó döntéstámogató rendszer célja az egyetemi tudástőke elemeinek azonosítása kvalitatív adatok felhasználásával, melyek forrásai a kutatói kompetenciák, a folyó kutatási tevékenységek, egyéb kommunikációs tartalmak és a régiós környezet jellemzői.

LEGFONTOSABB EREDMÉNYEK, ÚJDONSÁGOK

A tartalmi változások követésével és elemzésével lehetőség nyílik stratégiai szempontból használható, automatikusan előállított kutatói profilok és kapcsolati hálózatok előállítására. Szövegbányászati módszerekkel kialakított indikátorok felhasználásával az egyetemi versenyképességet jobban jellemző modell alkotható.

GYAKORLATI JAVASLATOK

Munkánk során az egyetemi versenyképességi alapmodellek, a mérésekhez használt indikátorhalmaz és az egyetem kutatói közösség és környezet vizsgálati lehetőségeit és IKT eszközeit mutatjuk be, melynek eredményeként alakítottuk ki és ismertetjük a tervezett rendszer alapját képező modellünket, melynek gyakorlati alkalmazása a későbbiekben várható.

Kulcsszavak: big data, adatbányászat, szövegbányászat, versenyképesség, Triple Helix

[1] A szerzők köszönetet mondanak Bencze Péter PTE TTK I. éves programtervező informatikus hallgatónak a crawler elkészítésében nyújtott segítségért.

BEVEZETŐ

A gazdasági versenyképességnek jellemzően négy szintjét különbözteti meg a szakirodalom (nemzet/régió, ágazat, vállalat, termék). A nemzeti versenyképesség esetében az ország által előállított nemzetközileg piacképes termékek, szolgáltatások és a termelékenység mutathatják (Rapkin & Avery 1995, Lewis 2004) Beszélhetünk továbbá területi (régiós) versenyképességről is, amikor a területegység képességét vizsgálhatjuk relatíve magas jövedelem és relatíve magas foglalkoztatottsági szint létrehozására (Lengyel 2003).

Elterjedt nézet szerint egyedül a *vállalati* szint értelmezhető, ahol a piaci versenyben való sikeres szereplés képessége mutathatja egy vállalat versenyképességét (Porter 1990, Krugman 1994). A vállalati szinthez szorosan kapcsolódó termékek versenyképessége esetén, többek közt az ár, a minőség, a műszaki színvonal mentén vizsgálódhatunk (Chikán 2006, Némethné 2010). Tanulmányunkban az egyetemet speciális terméket előállító vállalatként kezeljük.

A versenyképességet alkotó tényezők más-más súllyal esnek latba az adott ország fejlettségétől függően. Ameddig az alacsony fejlettségű országokban az alapvető infrastruktúra, a hatékonyság vezérelt országokban a hatékonyságemelők játsszák a fő szerepet, Az innováció vezérelte gazdaságokban a versenyképesség erősítése alapvetően az innovatív tevékenységekhez kapcsolódó intézményi, gazdasági környezet támogatásával történhet (Etzkowitz & Klofsten 2005). A fejlődést jellemzően előmozdító elemek az agglomerációval összefüggő *tudásmegosztás*, a spillover hatások, a *kommunikációs* és a kulturális hatások (Varga 2004, Aghion & Jaravel 2015). Ezeket az elemeket támogató környezet kialakítása a hatékonyság vezérelte gazdaságok, melybe Magyarország is tartozik, esetében is fontos. Ezen gazdaságokban a foglalkoztatottak egyre jelentősebb része a kreatív osztálynak is nevezett munkavállalói csoportba tartozik. Az egyetemi munkavállalókra a kreatív osztály tagjaiként tekintünk. Az egyetemek a hatékonyságvezérelt gazdaságokban is megjelenhetnek a tudásrégiók központjaiként, ennek révén az un. intelligens specializáció (smart-specialisation) keretében kialakított ökoszisztéma meghatározó elemévé válhatnak (Lengyel *et al.* 2016). Ezen ökoszisztéma vizsgálatában jelentős szerep hárulhat a rendszer elemei közötti kommunikáció során keletkező adatok elemzésére.

A *kreatív osztály* tagjainak területi elhelyezkedését Richard Florida elemezte. Az általa vizsgált területegységek (nála városok) kreatív potenciáljának összehasonlítására új, úgynevezett „puha" indexeket vezetett be, amely 3T modell néven is ismeretes (Tolerancia – Technológia – Tálentum). Elmélete szerint a prosperáló városokat a nyitott környezet, a másság elfogadása, az élénk és sokszínű éjszakai élet, valamint a kiforrott (IT) infrastruktúra jellemzi. Ezáltal képesek magukhoz vonzani a munkavégzésükhöz modern technológiát használó, magasan képzett kreatív osztály tagjait. Florida szerint a hagyományos „kemény" mérőszámok mellett alkalmazandó *„puha" változók* (pl. szabadidős lehetőségek) jól támogatják az adott térség versenyképességének vizsgálatát (Florida 2002).

Modellünk fontos részét képezik a humán tőke és az intézményi társadalmi tőke elemek, valamint a hírforrások vizsgálata, amelyek a Lengyel-féle piramis modellben is szerepelnek (Lengyel 2003).

Kutatói kérdéseink kialakításában az előzőekben vázolt modellek jelentették az inspirációt. Kísérletet tettünk az egyetemi szférát és munkavállalóit speciális vállalatként értelmezve többfajú adatforrások (1) és technológiák (2) azonosítására épülő modell kialakítására (3), mely elvezet egy működő rendszer kiépítéséig.

A tervezett rendszer egy alkalmazási területe a felsőoktatási szektor területi versenyképességének meghatározása. Az egyetem több szempontból is ideális a versenyképesség vizsgálatokhoz, hiszen a régiót jelentősen befolyásoló vállalatnak tekintjük, melynek jelentős a kreatív osztályba sorolható munkavállalói létszáma, régiós innovációs tevékenysége meghatározó és mind az intézményi, mind a gazdasági környezettel való kapcsolata kiemelkedően szoros. E kapcsolatban mind alaptényezőket, mind külső tényezőket is figyelembe veszünk, melyek elemzésében a szövegbányászat eszköztárát használjuk. A bemutatásra kerülő modell gyakorlati alkalmazása még nem történt meg, az egyes modulok fejlesztése folyamatban van. A kialakítandó rendszer az egyetemen belüli kommunikáció vizsgálatának segítségével azon erőforrások azonosítását támogatja, melyek felhasználásával az egyetem, mint egy tágabb ökoszisztéma tagja, kapcsolódni tud a rendszer többi eleméhez. Az ökoszisztéma elemei közti kommunikáció vizsgálatával a tervezett rendszer támogatást képes nyújtani a rendszerben elfoglalt pozíció javításában.

A VIZSGÁLAT TÁRGYA – AZ EGYETEMI SZEKTOR

Az egyetemi ökoszisztéma kutatásában három jelentős hatású modell terjedt el: a vállalkozói egyetem, a harmadik generációs egyetem, valamint a Triple Helix. Mindhárom modell középpontjában a társadalmi, gazdasági, kulturális változásokhoz való alkalmazkodás áll, azonban a hangsúlyok eltérőek. A vállalkozó egyetem (entrepreneur university) modellje szerint az egyetemnek az erőforrásai szűkössége okán a vállalati működési mód felé kell fordulnia. E modellben a gazdaság egyéb szereplői és az egyetem közti tudástranszfer háttérbe szorul, míg az egyetem keretein belül a vállalkozói szemlélet elterjesztése kiemelten fontossá válik (Hrubos 2001, Polónyi 2005). A harmadik generációs egyetem (3GU) modellje szerint a modern egyetemnek az ipari kapcsolatok (igények) által erőteljesen meghatározó interdiszciplináris tudásmunkások képzését kell végeznie. A modell hangsúlya az ipari kapcsolatokon van, melyek révén biztosítható a szükséges diáklétszám és így az egyetem fő profiljának a tudástermelésnek a folytatása (Wissema 2009).

A tervezett rendszer elméleti kerete a Triple Helix modell, mely a szereplők – akadémiai szektor, a magánszektor és a kormányzat hármasa alkotta rendszer – belső erőforrásaira épülő, kölcsönösen előnyös kapcsolatokra helyezi a hangsúlyt. A modell elméleti háttere szerint a XX. század végére a tudás felértékelődött a hagyományos termelési tényezőkkel szemben, azonban a tudást összpontosító egyetem önmagában nem elégséges a területi versenyképesség növelésében. Ezért a Triple Helix rendszerben gondolkodik: az egyetem által előállított tudás, innovációs potenciál paraméterei a magánszektor versenyképességének egyik pillére, amelyet a gazdasági, kormányzati környezet jelentős mértékben befolyásol. Így a modell e három szektor kölcsönös függőségét is mutatja (Bajmóczy 2005). A magánszektor versenyelőnyre tehet szert a jól működő egyetemi kapcsolataiból, míg az állam a gazdasági növekedés támogatásában érdekelt félként, közreműködik az egyetemeken képződő eredmények ipari hasznosulásában. A magánszférából érkező támogatások és megbízások kiegyenlíthetik az állami költségvetés hiányosságait az egyetemi kutatások támogatásában. Gazdasági szereplők bevonásával az egyetemek kiegészítő forrásokhoz is jutnak (Etzkowitz & Leydesdorf 2000).

Az 1. táblázat foglalja össze a Triple Helix modell alapkategóriáit, melyek az egyes szereplők alapján kerültek felosztásra. Látható, hogy a három szereplő (egyetem, ipar és kormányzat) mind céljaiban, mind a sikeres eredmények azonosításában, mind a tevékenységeket végző ágensekben igen jelentősen eltér egymástól.

1. táblázat. A Triple Helix modell alapkategóriái

	Egyetem	Ipar	Kormányzat
Kultúra	Autonómia	Ipari célok	Erős gazdaság
Tevékenység	Kutatás	Innováció	Támogatás
Cél	Minőség	Profit	Felülvizsgálat
Eredmény	Új tudás	Új termékek	Stimulálás
Ágensek	Kutatók	Cégek	Politikusok
Jövő	Vállalkozások	Ipari kutatások	Adminisztráció

Forrás: Etzkowitz 2000, 321. o. alapján

A modell szereplőinek közös erőforrása a tudástőke, aminek bázisa az egyetemi környezet kutatói, eredménye pedig az ipari környezetben innovációkon keresztül termel profitot olyan támogató környezetben, amelynek célja egy erős gazdaság létrehozása és szabályozása. A Triple Helix modell alapján az egyetemen képződő új tudás hasznosulásának egyik indikátora az ipar és az egyetem közti kommunikáció intenzitásának mértéke és a tudástranszfer minősége, amelyek mind a tudástőke részét képezik. A tudástőke felhalmozás révén létrejövő tudáscentrumok a regionális fejlesztés

központi mozgatóivá válnak, hatásuk kiterjed a régió gazdasági szereplőire, melynek révén a regionális versenyképesség javításában jelentős szerepet töltenek be (Lengyel 2005).

Az új tudás létrehozásának és ipari alkalmazásának alapját képező vállalati tudástőke mérése fontos szerepet tölt be az egyetemi versenyképesség jellemzésében. Ezért a vállalati tudástőke mérhetőségének megteremtése fontos eleme az innovációval foglalkozó kutatási területeknek. Maeques *et al.* (2006) eredményei alapján a vállalati tudástőke (IC) három fő komponense: a humán tőke (HC), a strukturális tőke (SC) és a kapcsolati tőke (RC). A humán tőke kategóriájába soroljuk ez egyéneknek a vállalat szempontjából lényeges összes tacit (hogyan) és explicit (mit) tudását. A strukturális tőke elemei közé tartozik a vállalat szervezeti szintjén megjelenő közösségi tudás, mely a mindennapi működésben, stratégiaalkotásban, vezérelvek kidolgozásában és az üzleti folyamatok menedzselésében jelenik meg. A kapcsolati tőke a vállalat és annak külső környezete (partnerek, szabályozók, versenytársak) közötti interakciókhoz kapcsolódó ismeretek és képességek, melyeknek egyre fontosabb részét képezi az úgynevezett közösségi tőke (SoC). A közösségi tőke az egyének közötti kapcsolati hálózatokban meglévő és onnan kinyerhető tudás forrása.

Ezen belül az egyetemi tudástőke elemeinek mérése ma még kevéssé elterjedt annak összetettsége okán. A University Entrepreneurial Scorecard (NCEE 2014) rendszerében jellemzően a strukturális tőke (SC) és kisebb mértékben a humán tőke (HC) dimenziókba tartozó, egyszerű kérdésekre válaszolva és a megszerzett pontszámok összesítésével a vizsgálati egységet jellemző „entrepreneurship" mérőszám kalkulálható. A módszer hátránya a kitöltő szubjektivitása, az önműködővé tehető folyamatos elemzés teljes hiánya. Megemlítendő az Európai Bizottság és az OECD által a felsőoktatási intézmények számára készített HEInnovative (HEI) önértékelési rendszere is (HEI 2016). A HEI online elérhető rendszere hét fő területre (Vezetés és kormányzás, Szervezeti teljesítmény: Költségvetés, Emberek és Ösztönzők, Vállalkozói szemlélet oktatása, Vállalkozói elökészítés és támogatás, Tudáscsere és együttműködés, Nemzetköziesedés, Hatásvizsgálat) fókuszált kérdésekre adott pontszámok alapján, jól vizualizált formában teszi értelmezhetővé az eredményeket. Az intézmény a stakeholderekkel közösen, különböző időszakokban történő kitöltések összehasonlításával, folyamatában vizsgálhatja a tevékenységeik eredményeként bekövetkező változásokat.

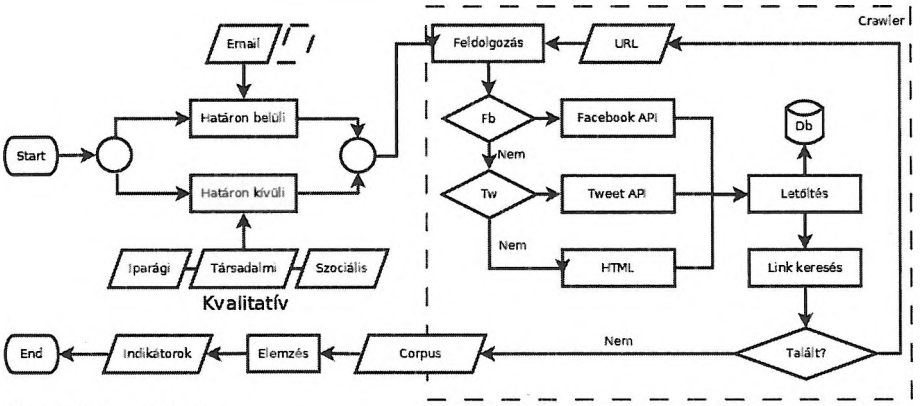
A Tudáspark rendszer az önértékelési módszerekkel szemben az egyetemi kommunikációban már meglévő dokumentumokra épül, és automatizáltsága révén csökkenti a szubjektivitást, amely az információs technológiák fejlettsége révén biztosítható. A tervezett rendszernek a következő fejezetben bemutatásra kerülő modelljében e technológiák felhasználásának kereteit fektetjük le.

A VIZSGÁLAT ESZKÖZE – A TERVEZETT RENDSZER MODELLJE

Az egyetemi tudástőke elemei a hagyományosnak számító a strukturált adatokra épülő *mennyiségi* mutatószámok mellett (publikációk, szabadalmak száma stb.) leginkább *minőségi*, nem strukturált adatok (levelezések, dokumentumok, hang és videó állományok, prezentációk stb.) tartalomelemzésével, illetve a résztvevők kommunikációs csatornáinak (beosztás, szolgálati útvonal, folyamatmenedzsment, üzenetek) és kapcsolati hálózatának elemzésével nyerhetők ki. A feldolgozási módszereik eltérése miatt érdemes ezeket a csatornákat két csoportra osztani. Az egyirányú kommunikáció során nincs interakció, helyette hosszabb ideig elérhető, nagyobb terjedelmű anyagok műsorszórása történik. A kétirányú, párbeszédes kommunikáció rövidebb anyagai már időrendiséget és a kölcsönhatások következtében egymáshoz rendeltséget is tartalmaznak. A nemzetközi szakirodalomban a legutóbbi időszakban kezdtek megjelenni big data környezetre vonatkozó egyetemi, vállalati innovációs folyamatokat elemző, támogató vizsgálatok. Mikova (2016) konferencia anyagok elemzésével technológiai trendek azonosítására tesz kísérletet szövegbányászati eszközök alkalmazásával, Lin (2016) ismerteti egy konceptuális rendszer alapjait, mely a vállalkozások jövőbeli innovációs lehetőségeinek azonosítását tervezi támogatni Tech Mining Engineering adatbányászati technikával. Efimenko *et al.* (2016) R&D területről begyűjtött strukturálatlan adatok feldolgozását végzi innovációs területek feltérképezésére a Map of Science (www.mapofscience.com) segítségével. Végezetül a döntéstámogatásban alkalmazható, automatizált szövegbányászati technikákra épülő koncepcionális keretrendszert mutat be Kayser & Shala (2016), melyben heterogén adatforrások felhasználásával témaazonosítás és trendelemzés segítségével a körültekintőbb döntések meghozására nyílik mód.

A big data típusú adatforrások felhasználásával és ezek szövegbányászati feldolgozásával feltárt külső környezeti tényezők, belső erőforrások és „puha" indikátorok vizsgálata nem elterjedt. Ezért a hazai egyetemi szféra esetében ilyen irányú kutatások újdonságnak számítanak. Magyarországon a képzési és kimeneti követelmények (KKK) elemzését végezték el szövegbányászati eszközök felhasználásával (Kruzslicz 2013). Nemzetközi publikációs és konferencia absztrakt adatokra alapozott, szövegbányászat segítségével azonosított kutatói hálózatok feltérképezésében is történtek próbálkozások hazai kutatók közreműködésével. (Sinozic *et al.* 2015) A tervezett rendszer modelljének egyes részei az előbb felvázolt kutatásokban már megvalósításra kerültek.

1. ábra: Vállalati tudástőke felmérési folyamata



Forrás: Saját szerkesztés

Az 1. ábra mutatja a tervezett feldolgozási folyamat modelljét, melybe mindkét típusú kommunikációs csatorna beépítését elvégeztük. A folyamat első felében a minőségi típusú strukturálatlan szerkezetű adatok azonosítása történne. Ezen adatforrások tekintetében, a kommunikációs csatornák helyzete alapján az egyetem/vállalat határain *kívüli*, illetve azon *belüli* felosztást értelmeztünk, melyek részletesebb bemutatása a következő fejezetekben látható. Az adatforrások feldolgozását biztosító crawler modul kimenete az a szöveges dokumentumhalmaz (szakmai kifejezéssel korpusz), amely a szövegbányászati elemzés alapját adja. A crawler[2] egy olyan számítógépes program, amely a bemenethen megkapott webcímen található oldalakat a programot futtató számítógépre másolja. A crawler modul HTML alapú webforrásokból képes az adatgyűjtésre, melyek elérési címét egy bemeneti lista tartalmazza. A moduláris felépítés okán a későbbiekben tetszőleges HTML adatforrásokkal bővíthető az alrendszer: Researchgate, LinkedIn, Twitter, Facebook, stb.

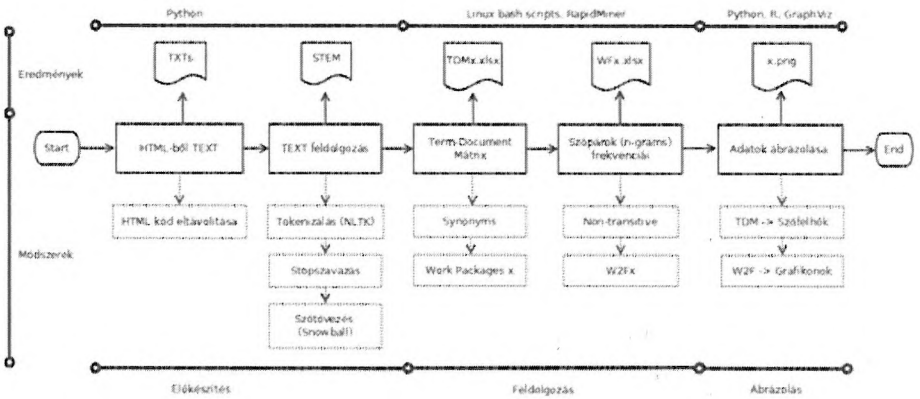
A crawler modul a feldolgozás során – a weboldalon azonosított linkek felhasználásával – további útvonalakat is bejárhat, mely bejárási mélység egy konfigurációs állomány segítségével szabályozható. A bejárás eredményeként a kiinduló központi adatforrásokon lévőhöz képest jelentősen dúsabb adat begyűjtésére nyílna mód. A megoldás hátránya, hogy sok esetben kevéssé releváns adattal dúsul a korpusz. E hátrány kiküszöbölésére Choudhary & Roy (2013) nyomán a fókuszált web crawling technika beépítése nyújt segítséget. A fókuszált crawler által letöltött weboldal egy relevancia pontszám alapján kerül elvetésre vagy eltárolásra. Mivel a pontszámítás szövegbányászati elemzésen alapszik, ezért a célzott dokumentumok szűrésének köszönhetően a korpusz minősége jelentősen növelhető.

A crawler modul által begyűjtött adatok relációs adatbázisban tárolódnak. A tárolási szerkezet több korpusz eltárolását biztosítja, melyekhez tartozó dokumentumok esetében lehetővé teszi a nyelv, szerző, időpont, forrás és típus metaadatok

[2] A Crawler modul a szerzők által fejlesztett Python nyelven készített program.

rögzítését is. Az adatbázisban tárolt weblapokból egy tisztítási folyamat eredményeként létrejövő szöveges állományok biztosítanak lehetőséget a további szövegbányászati vizsgálatokra. A korpusz elemzése Tikk (2007) alapján szövegbányászati technikával történik, melynek egy lehetséges folyamatát a 2. ábra mutatja.

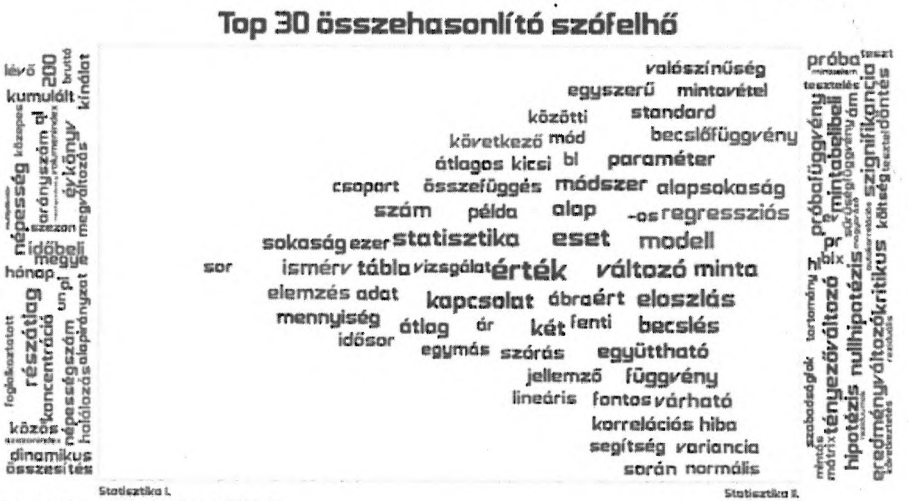
2. ábra: Szövegbányászati elemzés folyamata



Forrás: saját szerkesztés

A szövegbányászati elemzés eredményeként létrejövő mutatók jelentik a folyamat eredményét és végét. E mutatók felhasználásával különböző típusú szófelhő segítségével láttatható az eredmény. A hagyományos szófelhős ábrázolás a korpuszban szereplő szavak gyakorisága alapján történő ábrázolás révén növeli az értelmezés minőségét. További lehetőség összehasonlító (Conway) szófelhők segítségével különböző szempontrendszerek egymáshoz viszonyított eredményeit grafikusan ábrázolni. E két módszer együttes alkalmazásával az eredmények értelmezését nagymértékben megkönnyítő módszer adaptálása történhet, melyre a 3. ábrán látható példa. A tervezett rendszer modelljében bemutatott folyamat alkalmas az állandó adatfrissítésre és elemzésre, így a vizsgálati egységet jellemző indikátorok naprakészen tartására.

3. ábra: A Herman – Rédey (2006) és a Pintér – Rappai (2006) tankönyvek egybevetése



Forrás: Kruzslicz és társai (2016)

A tervezett rendszer modelljének alfolyamatai közül jelen fázisban a crawler modul elkészítése fejeződött be szűrt adatforrásokra (Twitter, Facebook). A crawler által legyűjtött weboldalak eltárolását biztosító relációs adatbázis és az adatbázisban elérhető elemre épülő adattisztító folyamat szintén elkészült. A szövegbányászati elemzés eredményeként előálló mutatók és az adatmegjelenítést biztosító szófelhők Python programjai tesztelési fázisban vannak.

A VIZSGÁLAT ELEMEI – EGYETEM HATÁRAIN KÍVÜL ÉS BELÜL

A következőkben a tervezett rendszer modelljének részét képező elemeket mutatjuk be.

Egyetem határain kívüli iparági, társadalmi és közösségi környezet vizsgálata

Az egyirányú (emissziós) kommunikáció elemzésekor a feladat a szerző által a külvilág számára közölt információk feldolgozása. Publikált tartalmak esetén feltételezhetjük, hogy a célközönség részben ismert, de a valódi felhasználók köre általában nem. Az ilyen kommunikáció általában időben aszinkron jellegű, és a dokumentumok terjedelmesebbek. A vizsgálati egység határain kívüli adatforrásokat az alábbiakban csoportosítottuk:

- Központi közösségi médiafelületek
- Központi hírek, blogok és weboldalak
- Központi egyéb szöveges dokumentumok

A fenti forrásokból elérhető adatok begyűjtését az előző fejezetben részletesen bemutatott crawler modul végzi.

Egyetem határain belül elérhető kutatói profilok és projektek vizsgálata

Az egyetemi tudásvagyon megjelenési formáinak egy része az előzőekhez hasonló egyirányú kommunikációs technológiák felhasználásával érhető el. Az egyetemen belül azonban lehetővé válik a kétirányú, többségében informális csatornák üzeneteinek mérési célú hasznosítása is. Noha ennek jogi és etikai kérdései további vizsgálódások tárgyát képezik, az ilyen adatok hasznosításának lehetősége technológiailag rendelkezésre áll.

Egy kutató stratégiai értéke több komponensből tevődik össze: a *tudása*, a *képessége*, a *motivációja*, *kapcsolati hálója* és az *attitűdjei*. Ezek közül az első három, legkönnyebben fejleszthető elem adja a kompetencia felszínét, míg az utolsó kettő, nehezen fejleszthető tényező adja a magját. Ezek az ismérvek sokszor egymástól elszigetelve vannak tárolva egy intézményen belül, és akár többszörözötten is. Begyűjtésük legtöbbször ad hoc jellegű, manuális úton történik, és gyakran csak adminisztratív célokat szolgálnak. Gyorsan változó információk esetén mindezek naprakészen tartása sokkal költségesebb, mint maga a begyűjtés. Puha információk esetén további problémát jelent az adatok megbízhatósága és tisztasága. A kutatói profilok mintázatainak felismerése, szervezeti célokhoz történő, vagy akár fordított illesztése elengedhetetlen az innovációk hasznosulásához.

Egy kutató kompetenciáját rendszeres adatfelvételekkel feltérképezve csak látható teljesítménye (felismert tudása és képességei) mérhető. A folyó tevékenységek és viselkedések megfigyeléséből ehhez képest naprakész információ szerezhető (akár a motivációk és attitűdökről is). A kutatók kommunikációs csatornáinak (email, blog, Skype, újság, konferencia) feltérképezése után lehetőség nyílik arra, hogy ezek változásainak követésével és elemzésével automatikusan állítsunk elő stratégiai szempontból használható kutatói profilokat és kapcsolati hálózatokat. Intézményfejlesztéskor vagy pályázati előkészítésnél például a múltbeli kimutatásoknál sokkal fontosabb, hogy a kompetenciák jelen fejlődési trendjeit és erőforrásait is megismerhessük.

Egy ilyen profilkészítő alkalmazás lehetséges adatforrásait felsőoktatási intézményekre vonatkoztatva Amini *et al.* (2014) gyűjtötték össze. Vizsgálataik szerint egy-egy kutató profilja az alábbi négy összetevőből áll össze, amihez Soares *et al.* (2013) nyomán egy újabb elemet lehet hozzácsatolni, hogy az alábbi táblázatot kapjuk:

2. táblázat. Kutatói profil forrásai

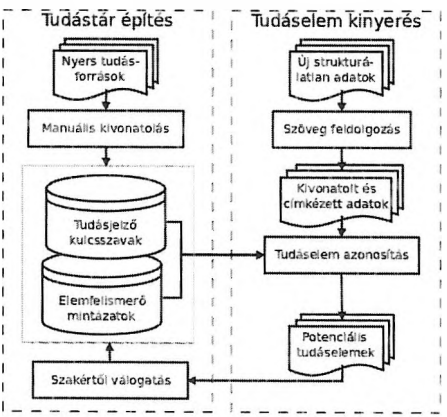
Kutató				
Publikáció	**Oktatás**	**Szakmunka**	**Könyvtárazás**	**Kommunikáció**
Absztraktok Konferencia diák Poszterek Cikkek Könyvrészletek Könyvek Recenziók	Tematikák Előadásvázlatok Tananyagok Számonkérések Szakdolgozatok	Pályázatok Projektek Szakmai blogok Érdeklődési területek Közösségi tartalmak	Kölcsönzések Szervezeti tagságok Kutatási katalógusok Keresési szokások	Belső folyamatok Levelezés Beszámolók Közösségi profilok Hobbi

Forrás: Amini et al. 2014 és Soares et al. 2013 alapján

Ezekben a forrásokban a kompetenciaelemek eltérő mennyiségében és minőségben jelennek meg. Fontos tehát, hogy egy ilyen rendszer kiépítése előtt a lehetséges adatforrásokat ilyen szempontból is feltérképezzük, és priorizáljuk azokat költség, elérhetőség és tartalom alapján. Egy már működő modellbe ugyanis újabb adatforrás bevonása a későbbiekben már csak feldolgozási kapacitási kérdéseket vet fel.

A tartalomelemzés alapú, félig automatikus tudáskinyerés folyamatának Tsui et al. (2014) által javasolt általános sémája két jól elkülöníthető tevékenységkörből áll: a tudástár építésből, melyhez jelentős emberi munkára van szükség, illetve a tudástár elemeinek gazdagításából, új lehetséges elemek kinyerése által. Az ilyen elven működő rendszerek tehát elsősorban az új tudáselemek önműködő kinyerésének fázisát hivatottak támogatni.

4. ábra: Párbeszéd alapú tudástár építésének kétfázisú modellje



Forrás: Aminin et al. (2014)

A modell szempontjából lényeges, hogy az elemzés egységes bemenetét egyszerű szöveges állományok képezik. A különféle egyéb formátumú (txt, doc, pdf, ppt) tartalmak becsatornázásához készítendő interfészek már képesek lehetnek a formázási információk felhasználására is. Mivel a modell elég összetett (szövegelemzés, egyértelműsítés, ontológiaépítés) ezért szükség van a tartalmak előzetes szűrésére és a változások detektálására is. Hogy a modell milyen eredményességgel lenne képes működni egyetemi környezeten kívül, arról az eddigi publikációkból érdemi információ nem áll rendelkezésre. Egy ilyen modellre épülő rendszer elkészítéséhez egy kezdeti, szakértők által összeállított kiindulási kompetenciakör szűkítő lista elengedhetetlen. A siker további kulcstényezője, hogy a feldolgozásra kiválasztott csatornából kinyert információk kizárólag pozitív és minden érintett fél részére előmozdító és ne büntető céllal legyen felhasználva. Hiszen különben a kommunikáció az adott csatornán meg fog szűnni, vagy csak kényszer hatására fog tovább működni. Ekkor azonban lényegében visszakerülünk a manuális adatgyűjtés szintjére. Maga a modell tehát alkalmas mind az egyirányú, mind a kétirányú kommunikációkból származó szöveges adatok feldolgozására. A továbbiakban elsősorban mégis csak a párbeszéd jellegű tartalmakból történő tudáskinyerési lehetőségeit mutatjuk be.

Diskurzus során az információknak nemcsak a szerzői, hanem a címzettjei ismertek és a feldolgozás szempontjából fontosak is. A kérdezz-felelek és vita formájú párbeszédeknél a résztvevők szövegrészekhez hozzárendelése plusz feladat.

A szinkron csatornák elemzésének is több szintje képzelhető el. Alapszinten (*statisztikai szint*) elegendő a kommunikációs csatornák szerkezetét kvantitatív mérőszámokkal (pl. gyakorisági, terjedelmi, olvasottság, impakt faktor) jellemezni. Noha ezek mára már az informatikai rendszerekből,

illetve azok naplóállományaiból könnyen kinyerhetőek lennének, jelenleg a legtöbb helyen mégis újra és újra manuálisan kérik be. A következő (*hálózati szint*) a kapcsolatok számszerűsítése adódik, (ki hova publikál, ki kivel levelez, ki milyen projektben vesz részt stb.) aminek szintén az időbeli változása is érdekes lehet. A harmadik (*tartalmi szinten*) a nyelvi technológia bevonásával sor kerülhet a kompetencia elemek (pl. témakör, szakterület, megoldások) felismerésére (*retrieval*) vagy kinyerésére (*extraction*) is. Ezek a szintek azonban még csak mennyiségi alapon közelítik meg a problémát. További szemantikai elemzésekkel még finomabb, fogalmi szintű információk is kinyerhetők, melyek nélkülözhetetlenek a motivációs és attitűd háttér megismeréséhez.

KAPCSOLATI HÁLÓK

A kutatói profilok képzéséhez szükséges állományok felhasználhatóságának adatvédelmi és jogi feltételei intézményenként is eltérőek lehetnek. Elegendő, ha ezek közül csak a munkahelyi levelezésekbe való munkáltatói betekintés kérdését emeljük ki. Márpedig az érzékeny adatok kinyerésétől való félelem csak akkor lehet jogos, ha ez technológiailag meg is oldható. Ennek igazolására a tudáskinyerő módszereket leginkább nyilvános adatforrásokon lehetséges kipróbálni, mint ahogyan Chang és Poon (2009) is tesztelték az email osztályozási modelljüket. Az intézményi emailforgalomhoz tartalmi, de főleg formai szempontból az interneten elérhető levelezési listák archívumai a leghasonlóbbak. Ezek közül az R-help nevű (https://stat.ethz.ch/mailman/listinfo/r-help) fórum teljes anyagát dolgoztuk fel, hogy hálózatelemzési módszerek segítségével megkeressük az egyes témakörök szakértőit.

A cikk írásakor R-help összesen 31 999 felhasználó 383 777 üzenetét tartalmazta a kezdetektől (1997. április) egészen a vizsgálat megkezdéséig (2015. októberig) bezárólag. Az R-help lista célja az R adatelemző nyelvvel kapcsolatos általános problémák megvitatása és különböző javaslatok felvetése. Más-más céllal és szűkebb körnek további üzenetarchívumok is elérhetőek: R-announce – újdonságok bejelentése, R-devel – a nyelv fejlesztőinek fóruma, R-packages – a kiegészítő csomagok használatainak támogatása, valamint számos egyéb speciális témakört érintő, mélyebb ismereteket igénylő SIG-fórum. Az elsősorban, de nem kizárólag angol nyelvű üzenetek R nyelvű kódrészleteket is tartalmazó párbeszédek szerkezete az elektronikus levelezésben használt formai elemekre épül. Például a korábbi üzenetekből származó idézeteket a sor eleji „nagyobb" (>) karakterek jelölik:

From lists at xxxx Sun Oct 4 13:41:42 2015
From: lists at xxxx
Date: Sun, 4 Oct 2015 12:41:42 +0100
Subject: [R] Calling external file
In-Reply-To:<CAKTtY6RRkiW2CVA-B2A@mail.gmail.com>
References: <CAKTtY6R=6GdrowJozS9KH@mail.gmail.com>
<CAKTtY6RRkiW2CVA-B2A@mail.gmail.com>
Message-ID: <56111076.9030901@dewey.myzen.co.uk>

In line

On 03/10/2015 23:56, xxx wrote:
> Thanks Bill. Simplified content of max.calls.R (with repeated calls
> to maxLik removed) are shown below in the message. No, fn does not
> exist in the environment.

Which explains why R cannot find it.

> I call a routine (say probit.R compiled into a library) to
> use maxLik. Inside this routine,
> 1. In probit.R. likelihood function is defined yet in another nested
> routine;

Azt tapasztaltuk, hogy a tisztán mennyiségi adatok elemzésével is érdekes eredmények kaphatóak. A levelek fejlécéből kinyerhető feladó, címzett, időpont, és hivatkozási adatokból például akár az is megtudható, hogy a munkaidejének ki mekkora részét áldozza mások támogatására. Mivel az emberek munkahelye, emailcíme és néhol még a neve is megváltozott (pl. *Peter Dalgaard BSA* vs. *peter dalgaard*), szükség volt egy egyértelműsítési eljárás kidolgozására. Az alkalmazásnak ez a része egy-egy ember levelezésének időtengelyén haladva gyűjtötte össze a vele kapcsolatos változások listáját. A levelek fejléceinek válasz (reply) hivatkozásai alapján felgöngyölíthetőek az egyes kérdésekkel kapcsolatos megjegyzések és hozzászólások. Az, hogy az R-help valóban támogató lista, az hamar beigazolódott azzal, amikor 5 657 olyan tagot találtunk, akik rendszeresen válaszolnak kérdésekre.

A tagok maradék 82%-a (26 342 ember) csak kérdéseket tett fel, vagy megköszönte a kapott válaszokat. A levélforgalom mennyisége alapján meghatározhatjuk a támogató csapat magját is.

3. táblázat Az R-Help legaktívabb tagjainak listája

Levélforgalom irány szerinti bontása					
Helyezés	**Név**	**Válasz**	**Kérdés**	**Összesen**	**Részarány**
1	Prof Brian D Ripley	12187 db	296 db	12483 db	3,25%
2	David Winsemius	10291 db	26 db	10317 db	2,69%
3	Gabor Grothendieck	8397 db	401 db	8798 db	2,29%
4	Uwe Ligges	6865 db	24 db	6889 db	1,80%
5	Duncan Murdoch	6623 db	59 db	6682 db	1,74%
6	Peter Dalgaard BSA	4699 db	44 db	4743 db	1,24%
7	jim holtman	4480 db	25 db	4505 db	1,17%
8	arun	3695 db	531 db	4226 db	1,10%
9	Berton Gunter	3404 db	130 db	3534 db	0,92%
10	Thomas Lumley	3147 db	31 db	3178 db	0,83%

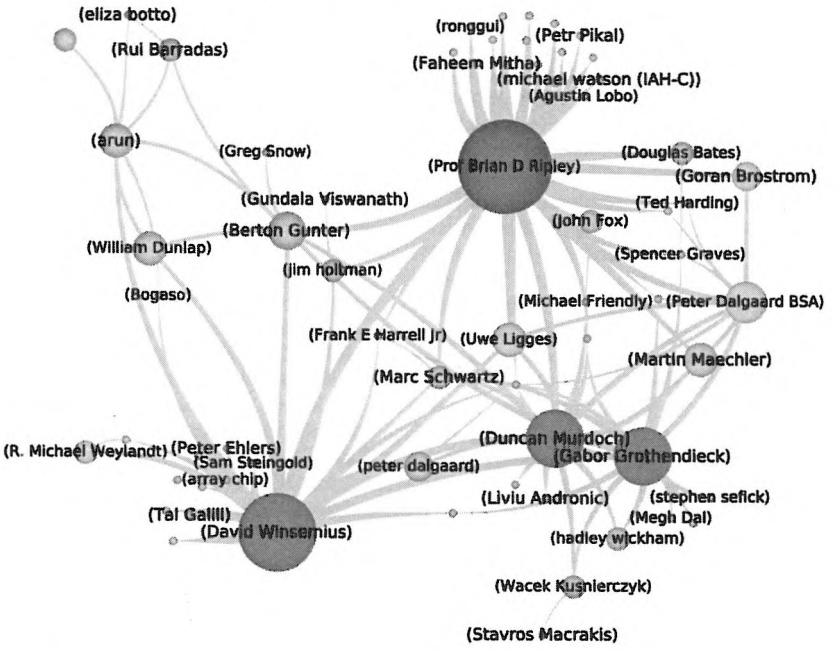
Forrás: Saját számítás

A támogató hálózatok egyik fontos tulajdonsága, hogy a segítségnyújtó csapat tagjai mennyire specializálódtak egy-egy területre, illetve menyire helyettesíthetők. Ez ugyanis jelentősen befolyásolja a válaszadás idejét, és a tagok cseréje esetén meghatározza a szolgáltatás stabilitását. Az üzenetek egymáshoz kapcsolódását kétféleképpen lehet definiálni: szintaktikai és szemantikai alapon. A szintaktikai elemzések során az üzenetek tartalmát nem használjuk fel, csupán a postázásra használt információk (címzett, küldő, hivatkozások, időrend) alapján építjük fel a kapcsolati hálózatot. A szemantikai elemzéseknél az üzenetek tartalmát is felhasználjuk annak azonosítására, hogy az melyik korábbi elemhez tartozik. Az elektronikus levelek esetén ez a módszer hatékonyan alkalmazható, lévén bevett szokás a válaszban a korábbi üzenet egészét, vagy annak lényegi részeit jól elkülöníthető módon idézetként feltüntetni. Az üzenetek szövegének mélyebb értelmezése lehetőséget a jelentésen alapuló kérdés-válasz párok megtalálásában és az egy témakörhöz tartozó problémafelvetések csoportosításában.

Szintaktikai-alapú hálózatépítő módszerrel megvizsgáltuk az R-help adatbázis kapcsolati hálózatát is. A fórum üzeneteinek mind kétféle hivatkozási módját felhasználtuk az üzenetek közötti kapcsolatok feltérképezéséhez. Az RFC 4021 szabvány szerint az „in-reply-to" hivatkozások közvetlenül jelölik meg azt az üzenetet, amelyikre a szerző reagálni kíván, míg a „references" adatok további olyan üzeneteket jelölnek meg, amelyek korábbi üzenetváltások alapján kapcsolódnak a témához. Az így kapott gráf feldolgozhatósága érdekében szűkítésekre volt szükség. Mivel a hálózat szolgáltató magjának szerkezetére vagyunk kíváncsiak, ezért a vizsgálatból kihagyhatóak a csak segítséget kérők. A kapcsolati hálózat ábrázolhatóságához tovább kellett szűkíteni a csomópontok számát, és a ritkítás után csak azokat a kapcsolatok tartottuk meg, amelyek mögött legalább 25 darab üzenetváltás történt. A minimális levélváltási korlát után kapott hálózatra kiszámítottuk a PageRank presztízs index (Brin és Page 1998) értékeket. Az eredeti felfogás szerint egy tag fontosságának mértéke megítélhető annak alapján, hogy mennyien kérdeznek tőle. Azt, hogy a válaszai alapján mennyire tekinthető valaki jelentős szerepűnek, a korábbiakban kiszámított levélforgalmi adatokból állapítható meg. A 4. ábrán látható gráf csúcspontjaiban a nagy forgalmat lebonyolító személyek vannak. A csúcspontok nagysága arányos a PageRank értékkel; a csúcsokat összekötő élek vastagsága pedig a levélváltások számával.

A hálózat főbb numerikus jellemzői az alábbiak. A legaktívabb tagok gráfja 66 csúcspontot tartalmaz. Az élhálózat eléggé sűrű, két tag között az átlagos úthossz 1,2 él, és a teljes gráf átmérője 5 élnyi hosszúságú. Az információs középponti index (Stephenson & Zelen 1989), normalizált változata azt fejezi ki, hogy ki milyen részben kontrollálja az információ áramlás folyamatát. A gráfon kiemelt négy fő középponti szereplő a levelezési folyamok 10,2%-ában rendszeresen érintett. A hálózat tehát eléggé elosztott, de fontos veszteség érheti, ha a két központi szereplője Brian D. Ripley és David Winsemius elhagyja. Ripley professzor az oxfordi egyetem alkalmazott statisztika tanszékén dolgozott, ahonnan 2014 augusztusában ment nyugdíjba. Az R nyelvhez oly módon kötődik, hogy annak előzményeként tartott S nyelv egyik fejlesztője volt. Társa, aki az R-help moderátora, biztosító társaságoknál dolgozott 2015. októberi nyugdíjba vonulásáig. A kétféle tevékenység (moderálás és elméleti szaktanács) a gráfon is jól elkülöníthető, mint a két kutató profiljának egy része.

5. ábra: Az R-help fórum üzenet-alapú kapcsolattartási hálózata



Forrás: Saját szerkesztés

Az elkülönítés nemcsak a levelezési útvonalak alapján lehetséges, hanem az üzenetek tartalma alapján az is megállapítható, hogy kire milyen témakörben lehet inkább számítani. Egy ilyen adatbázis kialakítása lehetőséget nyújthat arra, hogy a fórumot egy olyan ajánló rendszerrel egészítsék ki, amelyik a kérdés szövege alapján azt automatikusan a téma szakértőihez továbbítja. Statisztikai elemzésről, programozási nyelvről és környezetről lévén szó az egyes témakörök köré csoportosuló névelemek viszonylag könnyen azonosíthatóak, hiszen azok általában függvények vagy eljárások nevei, vagy az R nyelv kiegészítésére használt speciális csomagok megnevezései. Az eddig elvégzett vizsgálatokból látható, hogy a szöveg alapú kommunikáció automatikus feldolgozásával kapott információk jelentősen hozzájárulnak a vállalati tudástőke kinyeréséhez. A tudáselemek kodifikálásához további kutatásokra van szükség, de a mérhetőségükben már az eddig elvégzett vizsgálatok is jelentős előrelépésnek bizonyultak.

ÖSSZEFOGLALÁS

Tanulmányunkban a versenyképesség különböző dimenzióit a vállalati, a területi és az egyetemi szféra mentén mutattuk be. A területi versenyképesség Florida és Lengyel-féle megközelítése alapján megállapítottuk, hogy Krugman nyomán a regionális versenyképesség a területegységen működő gazdasági egységek versenyképességéből eredeztethető. E vállalkozások, vállalatok egy speciális esetének tekintve az egyetemi versenyképességet a Triple Helix modellen keresztül vizsgáltuk. A kvalitatív jellegű indikátorokra alapozva állítjuk fel modellünket, melyben a vizsgálati egység, esetünkben az egyetem, határain kívüli és határain belüli adatforráscsoportokat azonosítunk.

A határon kívüli iparági, társadalmi és közösségi környezetet jellemző médiafelületekről származó szöveges típusú adatok begyűjtése saját fejlesztésű crawler program segítségével történik, míg az adatokból kialakított korpusz tárolása egyedi adatbázisszerkezetben valósul meg. A korpusz szövegbányászati elemzése után különböző adatvizualizációs módszerek segítségével jellemezzük a környezetet.

A határon belüli környezet vizsgálatához kutatói profilok és az egyetemen megvalósuló projektek szöveges dokumentumai szolgálják az alapot. A profilok segítségével kompetencia elemekre vonatkozóan lehetséges információk gyűjtése. A kutatói kommunikáció és a kommunikációs csatornák elemzésével kialakítható statisztikai, hálózati és tartalmi szintű tudástár segítségével az egyetemen belüli tevékenységet jellemző motivációs és attitűd háttér, kvantitatív mérőszámok és kapcsolati hálók is előállíthatók.

A kialakított modell további finomítása az adatforrásokból érkező adatok finomhangolásával fokozható, melynek egyik eszköze a fókuszált web crawler technika beépítése. Az adatforrások körének további bővítésével az egyetemhez köthető kommunikációs hubokon zajló tevékenység is monitorozható (pl. pecs.hubbub.net), amely azonban szétfeszíti a határokon belüli és a határokon kívüli distinkciónkat.

HIVATKOZÁSOK

Aghion, P., Jaravel, X. (2015), „Knowledge Spillovers, Innovation and Growth", *The Economic Journal*, **125** 583, pp.533-73

Amini, B., Ibrahim, R., Othman, M. S., Selamat, A. (2014), „Capturing scholar's knowledge from heterogeneous resources for profiling in recommender systems", *Expert Systems with Applications*, **41** pp.7945-57

Bajmóczy Z. (2005), „'Vállalkozó egyetem' vállalkozásfejlesztési szemszögből", in Buzás N. (szerk.): *Tudásmenedzsment és tudásalapú gazdaságfejlesztés*, JATEPress, Szeged, 312-27. old.

Brin, S. and Page, L. (1998), „The anatomy of a large-scale hypertextual Web search engine", *Proceedings of the seventh international conference on World Wide Web*, pp.107-17

Chang, M. and Poon, Ch K. (2009), „Using phrases as features in email classification", *The Journal of Systems and Software*, **82** pp.1036-45

Chikán A. (2006), „A vállalati versenyképesség mérése: Egy versenyképességi index és alkalmazása", *Pénzügyi Szemle*, **51** 1, 42-56. old.

Choudhary, J. and Roy, D. (2013), „Priority based Semantic Web Crawler", *International Journal of Computer Applications*, **81** 15, pp. 10-3

EC (1999): *Sixth Periodic Report on the Social and Economic Situation and Development of Regions in the European Union*, European Commission, Luxembourg

Etzkowitz, H. (2003a), „Innovation in innovation: the Triple Helix of university-industry-government relations", *Social Science Information Sur Les Sciences Sociales*, **42** pp.293-337

Etzkowitz, H. (2003b), „Research groups as 'quasi-firms': The invention of the entrepreneurial university", *Research Policy*, **32** 1, pp.109-21

Etzkowitz, H., Leydesdorff, L. (2000), „The dynamics of innovation: from National Systems and „MODE 2" to a Triple Helix of university-industry-government relations", *Research Policy*, **29** 2, pp.109-23

Etzkowitz, H., Klofsten, M. (2005), „The innovating region: Toward a theory of knowledge-based regional development", *R and D Management*, **35** 3, pp.243-55

European Union (2011), *Connecting Universities to Regional Growth: A Practical Guide, Regional Policy, Smart Specialisation Platform*, letöltve: 2015.07.28. forrás: http://ec.europa.eu/regional_policy/sources/docgener/presenta/universities2011/universities2011_en.pdf

Florida, R. (2002), *The Rise of the Creative Class: And How it's transforming work, leisure, community and everyday life.* New York, NY: Perseus Book Group

HEI (2016): *HEInnovate is a self-assessment tool for entrepreneurial higher education institutions*, letöltve: 2016.08.31., forrás: https://heinnovate.eu/

Hrubos I. (2001), "Gazdálkodó egyetem – szolgáltató egyetem – vállalkozó egyetem", *Társadalom és gazdaság Közép- és Kelet-Európában,* **23** 3-4, 7-27. old.

Efimenko, I. V., Khoroshevsky, V. F. and Noyons, E. C. M. (2016), "Anticipating Future Pathways of Science, Technologies, and Innovations: (Map of Science) Approach", in T. U. Daim, D. Chiavetta, A. L. Porter and O. Saritas (eds.): *Anticipating Future Innovation Pathways through Large Data Analysis,* Springer, pp.71-96

Kayser, V., Shala, E. (2016), "Generating Futures from Text—Scenario Development Using Text Mining", in T. U. Daim, D. Chiavetta, A. L. Porter and O. Saritas (eds.): *Anticipating Future Innovation Pathways through Large Data Analysis,* Springer, pp.229-45

Krugman, P. R. (1994), „Competitiveness: A Dangerous Obsession", *Foreign Affairs,* **73** 2, pp.28-45

Kruzslicz F. (2013), "Felsőoktatási képzéskínálat elemzési lehetőségei duo-mining eszközökkel", *Magyar felsőoktatási rangsorok, hallgatói preferenciák* ankét, Pannon Egyetem Gazdaságtudományi Kar, Veszprém, letöltve: 2016.09.07., forrás: http://kmt.gtk.uni-pannon.hu/anket/ppts/Kruzslicz%20Ferenc.pdf

Kruzslicz F. – Kovács B. – Hornyák M. (2016), *„Összehasonlító klaszterjellemzés külső, szöveges források bevonásával", Statisztikai Szemle,* **94** 11–12, 1123-48. old.

Lengyel I. (2003), *Verseny és területi fejlődés,* JATE-Press, Szeged

Lengyel, B. (2005), „Triple Helix kapcsolatok a tudásmenedzsment szemszögéből", in Buzás N. (szerk.): *Tudásmenedzsment és tudásalapú gazdaságfejlesztés,* JATEPress, Szeged, 293-311. old.

Lengyel B. – Lengyel I. – Szakálné Kanó I. – Vas Zs. (2016), „Az újraiparosodás térbeli kérdőjelei Magyarországon", *Közgazdasági Szemle,* **LXIII** 6, 615-46. old.

Li, M. (2016), "A Conceptual Framework of Tech Mining Engineering to Enhance the Planning of Future Innovation Pathway (FIP)", in T. U. Daim, D. Chiavetta, A. L. Porter and O. Saritas (eds.): *Anticipating Future Innovation Pathways through Large Data Analysis,* Springer, pp.25-44

Maeques, D. P., Simon, F. J., Caranana, C. D. (2006), „The effect of innovation on intellectual capital: An empirical evaluation in the biotechnology and telecommunication industries", *International Journal of Innovation Management,* **10** 1, pp.89-112

Mikova, N. (2016), "Recent Trends in Technology Mining Approaches: Quantitative Analysis of GTM Conference Proceedings", in T. U. Daim, D. Chiavetta, A. L. Porter and O. Saritas (eds.): *Anticipating Future Innovation Pathways through Large Data Analysis,* Springer, pp.59-70

NCEE (2014), *The University Entrepreneurial Scorecard (Reviewing the Entrepreneurial Potential of a University),* letöltve: 2015.07.28., forrás: http://ncee.org.uk/wp-content/uploads/2014/06/Entrepreneurial_University_SCORE_CARD.pdf

Némethné Gál A. (2010), „A kis- és középvállalatok versenyképessége – egy lehetséges elemzési keretrendszer", Közgazdasági Szemle, LVII 2, 181-93. old.

Polónyi I. (2005), „Zsákban táncolva, avagy az egyetemi innovációs stratégia sajátosságai", *Competitio,* **4** 1, 201-12. old.

Porter, M. E. (1990), „The Competitive Advantage of Nations. (Cover Story)", *Harvard Business Review,* **68** pp.73-93

Rapkin, D. P., Avery, W. P. (Eds.), *National Competitiveness in a Global Economy,* Lynne Rienner, London, 1995

Sinozic, T., Maier, G., Hornyák, M., Kruzslicz, F. (2015), „The Fischer-man's Catch from the ECC: An exploratory text-mining approach", *55th European Congress of the Regional Science Association (ERSA),* 2015.08.25-28., Lisbon, Portugal

Soares, D. C., Santoro, F. M., Baião, F. A. (2013), „Discovering collaborative knowledge-intensive processes through e-mail mining", *Journal of Network and Computer Applications,* **36** pp.1451-65

Stephenson, K. and Zelen, M. (1989), „Rethinking centrality: methods and examples", *Social Networks,* **11** pp.1–37

Tikk D. (2007), *Szövegbányászat,* TypoTex, Budapest

Tsui, E., Wang, W. M., Cai, C., Cheung, C. F., Lee, W. B. (2014), „Knowledge-based extraction of intellectual capital-related information from unstructured data", *Expert Systems with Applications,* **41** pp.1315-25

Varga A. (2004), „Az egyetemi kutatások regionális gazdasági hatásai a nemzetközi szakirodalom tükrében", *Közgazdasági Szemle,* **51** 3, 259-75. old.

Webster, A. and Etzkowitz, H. (2000), „The future of the university and the university of the future: evolution of ivory tower", *Research Policy,* **29** 2, pp.313-30

Wissema, J. G. (2009), *Towards the third generation university. Managing the university in transition,* Edward Elgar, Northampton

Hornyák Miklós tanársegéd
hornyakm@ktk.pte.hu

Kruzslicz Ferenc, PhD, egyetemi docens
kruzslic@ktk.pte.hu

Pécsi Tudományegyetem
Közgazdaságtudományi Kar
Kvantitatív Menedzsment Intézet

Monitoring the university ecosystem in Big Data environment

AIM OF THE PAPER
The Triple Helix model is very common in examining the competitiveness of universities (Webster & Etzkowitz 2000, Etzkowitz 2003a). The influence of the environment where the university operate is one of the most important factor in its competitiveness. (Etzkowitz & Klofsten 2005, EU 2011). The key component in the university's competitiveness is the potential of knowledge which comes from the insider research community. We can typically use quantitative indicators for measuring this potential (Etzkowitz 2003b, NCEE 2014). In our model we prefer using qualitative indicators based on different types of datasources.

METHODOLOGY
We gather data inside and outside of the University of Pécs. From inside we use researcher projects, competencies, communication channels. From outside we use regional specific blogs, social media and news. In the model we use data- and textmining technique for mining relevant information to support the calculation of our indicators.

MOST IMPORTANT RESULTS
As a result of this process we can automatically follow and analyse the information flow, social networks and researchers' profiles at the university. Based on these new indicators we would like to support the competitiveness of the university better.

RECOMMENDATION
In this paper we present the base models of university competitiveness, the groups of indicators for measuring research groups in regional environment and the ICT support for this process.

Keywords: big data, datamining, textmining, competitiveness, Triple Helix