Gradus Vol 1, No 2 (2014) 230-237
ISSN 2064-8014

Közösségi adatforrások felhasználási lehetőségei a területi kutatás támogatásában
Utilising social data sources assisting regional research

Hornyák Miklós[1]
Pécsi Tudományegyetem, Közgazdaságtudományi Kar

Összefoglalás: A regionális versenyképességi elemzésekben a területegységeket jellemző, azok objektív összehasonlítását szolgáló indexek kialakítása általános. A létrejövő indikátorok a mennyiségi (kvantitatív) típusú adatok alapján építkeznek, azonban a minőségi (kvalitatív) típusú adatok felhasználása is terjed.Bár a felhasznált adatforrások és adattípusok köre heterogén, de jellemzően kérdőíves lekérdezések és interjútechnikák alapozzák meg a minőségi (’puha’) típusú indexeket. A részvételi hajlandóság sok esetben alacsony, az adatok beszerzése így nehézségekbe ütközik. Az on-line adatforrásokon elérhető hatalmas mennyiségű adat felhasználásával e ’puha’ típusú adathiány pótolható, információvá alakítható. A minőségi tulajdonságokra reflektáló, a banális világot jellemző, ám a virtuális térben működő 'közösségi érzékelők’ a duo-mining technológia segítségével elemezhetők. Ezen szenzorok szolgáltatta adatok lokációhoz kapcsolásával, szűrésével a területegységhez kötődő adathalmaz alakítható ki. Ezen adatok további elemzésével a vizsgált térség versenyképességének jellemzését tovább finomító minőségi típusú információk nyerhetők. A big data felhasználása és az adatbányászati technológiák további támogatási lehetőséget biztosíthatnak e területen.Dolgozatunkban a duo-mining témakörében zajló kutatások adatbányászati, szövegbányászati eredményeit mutatjuk be. Első példánkban internetes hírek szövegelemzésén alapuló, a területegység médiareprezentációját (poziítv / negatív hír) jellemző HírIndex kalkulálását végezzük el. A big data alkalmazások közül a Google Trends keresési elemzésen nyugvó Jövő Orientációs Index kialakítását mutatjuk meg, mely index a területegységről indított Google keresési kulcsszavak vizsgálata alapján jellemzi a magyarországi megyéket. Utolsó példánk adatbányászati módszerekkel történőkis-ésközépvállalaticsődindexkalkulálásátmutatja, melyben Support Vector Model felhasználásával a cégek pénzügyi, területiadatairaalapozvajelezzükelőre a csődbekövetkezésénekvalószínűségét.

Abstract: In studies analyzing regional competitiveness of territories objectively, it is common to define indexes. These indicators are based on quantitative data, however, the usage of qualitative data is becoming more widespread, as well.Although data sources and data types are heterogeneous, typically interview techniques based on questionnaires are dominant to define qualitative (’soft’) indexes. Willingness to participate in a survey is low, making data collection problematic. However, with the use of huge online data sources - the so called big data- this kind of ’soft’ data can be achieved and they can be converted into information. ’Community sensors’ reflecting qualitative features of information, - and which describe real life even though they function in virtual space- can be analyzed with the help of duo-mining technology. Data achieved by these sensors can be attached to locations and by filtering them, a database related to a given territory can be created. By the further analysis of this data a more refined, qualitative type of information is gained, which contributes to the definition of a territory’s competitiveness. The application of big data and other data mining technologies open up new ways of support in this field.In our paper we present the findings of researches conducted in the domain of data- and text mining. In our first presented case we do the calculation of a Newsindex reflecting a given territory’s media representation (positive

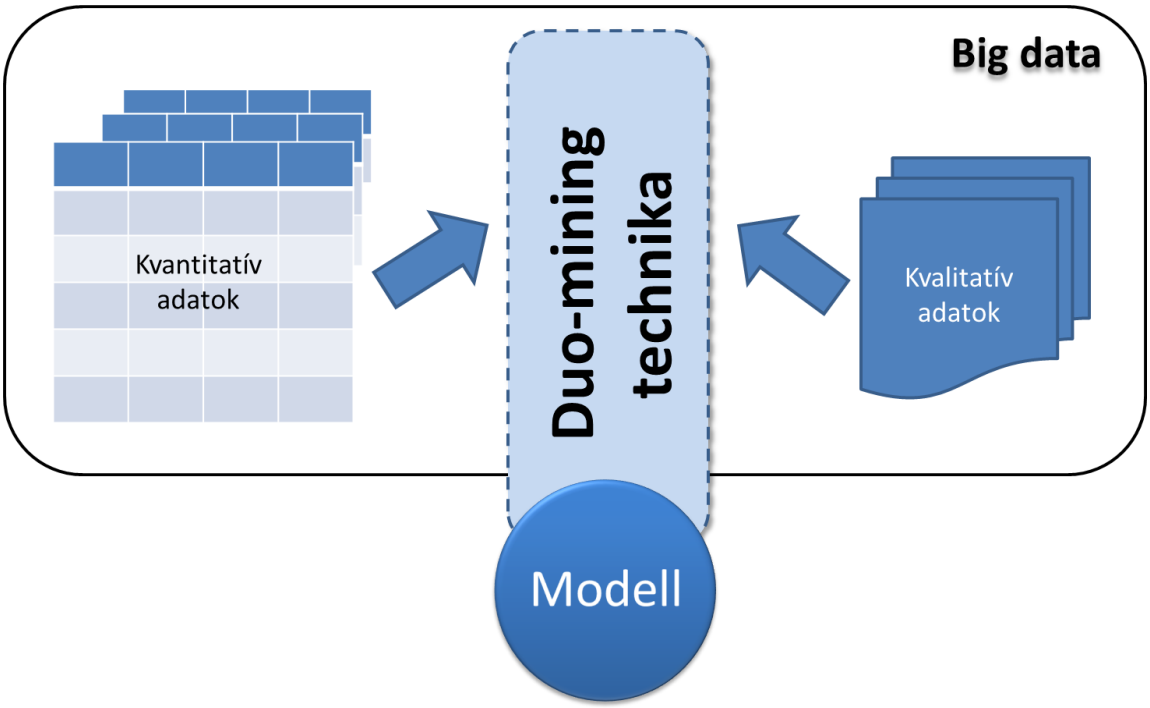
and negative) through the text analysis of online news. Among the big data applications, we highlight the creation of the Google Trends based Future Orientation Index, which characterizes Hungarian counties by processing data searches initiated from the given location. Our last example points out the calculation of small and medium size enterprises'' bankruptcy index with the help of data mining techniques, in which with the application of Support Vector Mode we forecast the possibility of companies' bankruptcy based on their financial, regional data.

Kulcsszavak: duo-mining, bigdata, szövegbányászat, versenyképesség, közösségi média

Keywords: duo-mining, bigdata, textmining, competitiveness, socialmedia

1. Bevezetés

Az információs társadalomban az internetpenetráció révén a mindennapok környezete és eseményei magas szinten dokumentáltak digitális formában. A dokumentálás elterjedtsége révén a környezet monitorozása folyamatos, melynek eredményeként hatalmas mennyiségű - jellemzően szöveges - adat halmozódik fel. A web 2.0 hatására a korábbi Business-to-Consumer (B2C) típusú egyirányú kommunikációs-modell helyett a Consumer-to-Consumer (C2C) típusú, azaz az interakciók sorozatán alapuló kommunikáció válik meghatározóvá. Ennek eredményeként a közösségi média a kölcsönös emberi viszonyok legfontosabb terepévé lép elő (Xia – Huan 2012).
A terepről érkező adatokat (szövegeket) társadalmi érzékelők (social sensors) outputjaiként felfogva a környezetünk egy speciális reprezentációja alkotható meg (Sakaki 2010). E „puha" típusú adatfolyamok feldolgozásához, a régiós kutatásokban kevéssé elterjedt technikákra van szükség. Ilyen új technika a duo-mining (1. ábra), amely az adat- és szövegbányászat párban történő használatát jelenti.A szöveg- és adatbányászat kettőse multiplikatív hatást fejt ki a vizsgálati eredmények (felismert mintázatok) minőségére. (Creese 2004, Fan 2005*)*



1. ábra: Duo-mining felhasználása a modellalkotásban
forrás: sajátszerkesztés

A következő fejezetekben a Pécsi Tudományegyetemen az adatbányászat és a szövegbányászat területén működő Duo-mining kutatócsoport regionális versenyképesség témájához köthető kutatásait mutatjuk be.

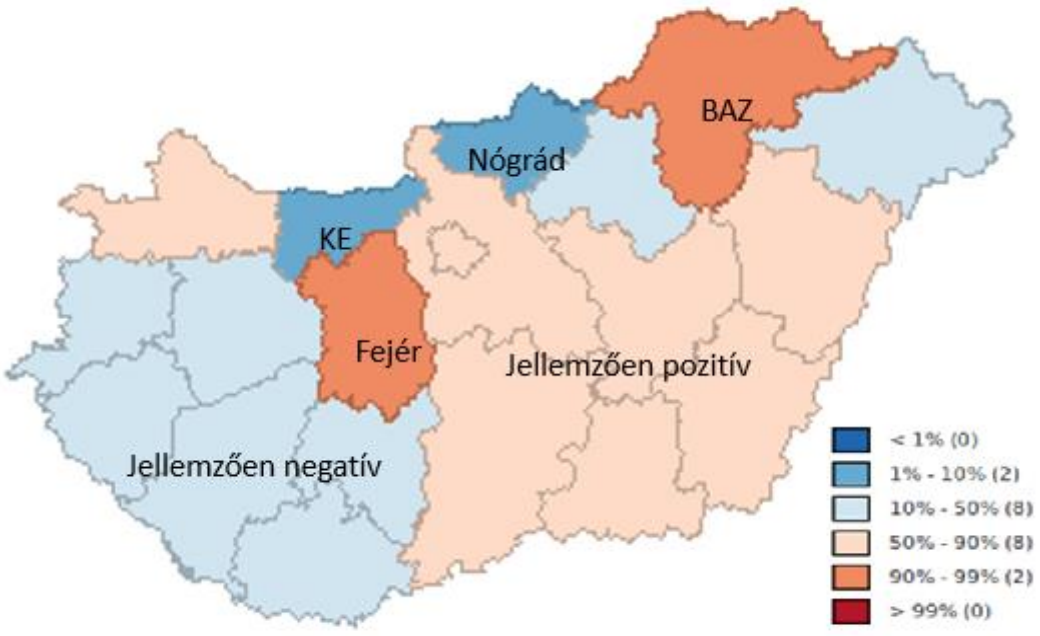
2. Szövegbányászati alkalmazás - HírIndex (HirIX) kialakítása

Az index.hu (http://index.hu) weboldal 2013. április 11 és május 13. közötti belföldi híreinek automatizált elemzését végeztük el. A szövegkorpusz kialakítása és előfeldolgozása után a hírek pozitív/negatív osztályozása a minősített szavak előfordulási gyakorisága alapján történt. Lokációk, azaz a magyarországi települések, azonosítása, majd az adatok megyénként történő aggregálása és a megyék lakosságszámának figyelembevételével kialakított hírindex (HírIX) segítségével a területegység médiában való megjelenését jellemeztük.
A korpuszunk első változata az automatizáltan begyűjtött 1000 db cikk HTML kódolású változata volt.Következő lépésben elvégzendő zajszűrés (noise filtering) célja a hiányos, hibás, szélsőséges és értelmezhetetlen adatoktól való tisztítás. Korpuszunk esetében az oldalstruktúra elemzése, a beazonosított cikktörzsek eltávolítása és HTML tagektől való tisztítás után kaptuk a korpusz nyers szöveges változatát.Szövegbányászati elemzésekre alkalmas korpusz kialakításához a szövegtartalom kisebb egységekre (szavakra) történő felbontását (tokenizálás) végeztük el, mely után vált lehetővé a jelentést nem befolyásoló, tartalmi információt nem hordozó elemek (stopwords) eltávolítása.
A következő lépésben elvégzett szótövezés (stemming) feladata a korpuszban azonosított (tokenizált) szavak módosulásainak (ragozás, toldalékolás) visszafejtése a szótőre. Ennek célja a vektortér csökkentése, lévén az azonos szavakat közös kanonikus alakba vonhatjuk így össze. Szótövezés végrehajtására a HunSpell Snowball algoritmusát használtuk.Az előfeldolgozási műveletek után a híreink osztályzását támogató szótár elkészítését végeztük el. Hu – Liu (Hu-Liu 2004) által készített szótár pozitív/negatív jelentéstartalommal bíró szavainak magyar nyelvre fordításával.
Az egyes hírek szavainak, szótövezés utáni gyakoriságát vizsgáltuk meg hírosztályok szerint. Így hírenként számíthatóa pozitív és negatív hírosztályba tartozó szavak száma.Pozitív hírnek fogadtuk el a pozitív és a negatív besorolású szavak egyenlőségének fennállását.Semleges hírnek tekintettük az egyik hírosztályba sem sorolható szavakat tartalmazó híreket (pl. képriportok, videó riportok). E lépések eredményeként a korpuszunkat alkotó híreket három csoportba sorolhattuk: pozitív, negatív és semleges.
Korpuszunkat a hírek besorolása alapján (pozitív, negatív, semleges) három részre bontottuk. A következőben a hírosztályonként alakítottuk ki a dokumentumok vektortér modellben történő reprezentációját. $D = \{d_1, \ldots, d_N\}$, ahol D vektor az egyes hírosztályokat alkotó cikkek dokumentumgyűjteményével azonos, melyben a d elemek az egyes dokumentumokra (hírek) hivatkoznak. (Tikk 2007)
Az így kapott D mátrix alapján hírosztályonként meghatározható azon szavak gyakorisága, melyek a lokációkat tartalmazó vektorunknak (magyarországi települések) is elemei.Adatainkat hírosztályonként megyei szintre összesítve kapjuk HM értékét $HM_{megye} = |\{HL_{lokáció} \in Megye\}|$ (1), amely az egyes megyék korpuszunkban történő pozitív/negatív reprezentációját mutatja.
A **2. ábra**látható szófelhő típusú ábrázolási mód segítségével érzékletesen tudjuk szemléltetni eredményeinket. Fejér és Borsod-Abaúj-Zemplén (BAZ) megyékneka hírekben való megjelenése kiemelkedően pozitív. E megyék híreiben a leggyakrabban előforduló pozitív

jelentésű szavak halmaza: támogat, kér, szeret, jog, tiszta; míg a leggyakrabban előforduló negatív jelentésű szavak: vég, fél, rossz, sért, súlyos, kár.

2. ábra: Magyarország megyéi HírIX értékeik alapján
Forrás: saját szerkesztés

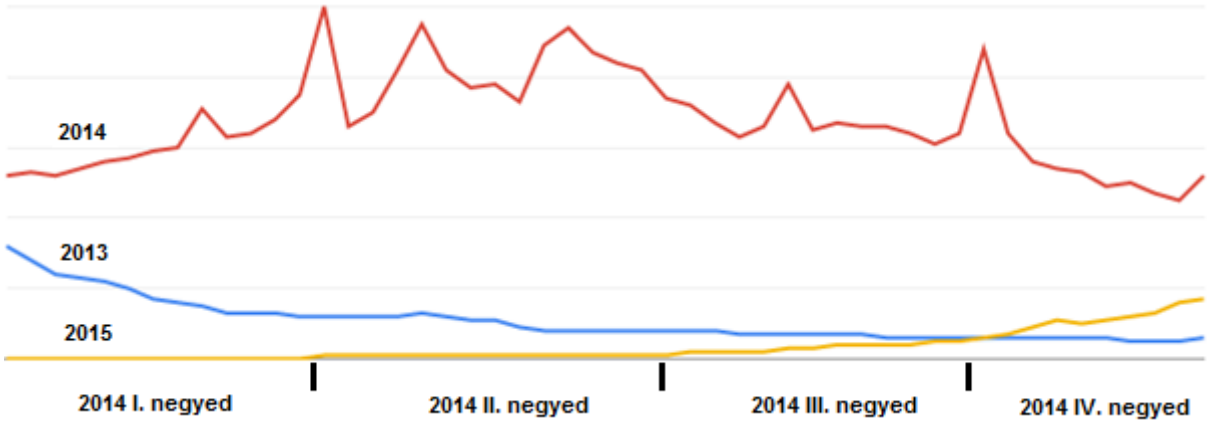
A 3. ábra az ország megyéire számított média megjelenési értékek százalékos megoszlását mutatja.Az adatok vizsgálatával az ország megyéi négy csoportba oszthatók: kiemelten negatív, kiemelten pozitív, jellemzően pozitív és jellemzően negatív. Az első csoportot Komárom-Esztergom és Nógrád megye alkotja, ahol a média reprezentáció kiemelten negatív, szemben a második csoport Fejér és Borsod-Abaúj-Zemplén megye kiemelkedően pozitív média megjelenésével. A főváros és az alföldi terület jellemzően pozitív reprezentációjú csoportja mellett a Dunántúl jellemzően negatív színben tűnik föl a vizsgált hírekben.



3. ábra:Média megjelenési értékek százalékos megoszlása megyénként

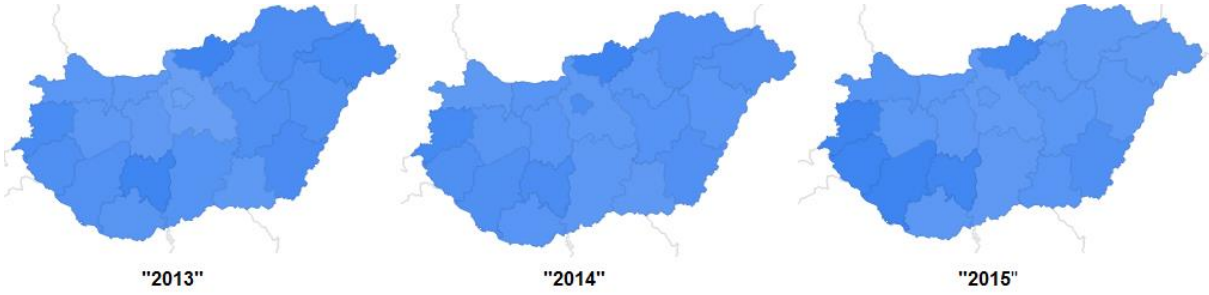
3. Big data alapúalkalmazás - Jövő-orientációs Index (FOI2014) kalkulálása

A ’big data’ napjaink egyik hívószava, mely a jelenlegi remények alapján paradigmaváltást eredményezhet a tudományos gondolkodásban. Egyes szerzők szerint jelenleg egy tudományos forradalom zajlik, melynek középpontjában a ’big data’ kínálta lehetőségek állnak. (Hey 2010)
Az egyik legkönnyebben hozzáférhető ’big data’ típusú adatlelőhely a Google Trends (http://www.google.com/trends/), ahol a Google keresőjének használati adatai alapján nyerhetők ki különböző keresési csoportok (pl. területi, téma alapú) mennyiségi jellemzői.
E lehetőséget használták ki Preis és szerzőtársai (Preis 2012) a Future Orientation Index (FOI) megalkotásakor, melynek alapgondolata egyszerű: területegységekhez (országok) kapcsolt speciális kulcsszavak használatával a keresési adatok alapján jövő/múlt orientáció meghatározása a Google adatok segítségével. E FOI mérőszám és a keresést indító területegység GDP adatai között korrelációs kapcsolatot találtak.



4. ábra:Magyarországi IP címrőlindítottGoogle keresésitrendek a “2013”, “2014”, “2015”kulcsszavakra,forrás: Google Trends

A jövő-orientációs index (FOI2014) számításához a Google Trends programot használtuk. Számításunk metódusát Preis és szerzőtársai cikke alapján végeztük. Területegységnek Magyarország megyéit tekintettük, szemben a Preis cikkben elvégzett ország szintű vizsgálattal. A 2014-es bázisévet választva a „2013” (múlt irányultság) és „2015” (jövő irányultság) kulcsszavakra történt – keresési kategória szűkítés nélkül - azon keresési darabszámok leválogatása, melyeket magyarországi IP címről indítottak a 2014-es naptári évben. A FOI2014 index számítása a jövő irányultság („2015” érték) és a múlt irányultság („2013” érték) találati elemszámok osztásával megyei bontásban számítódik. A kereséséi értéket a megye lakosságszámával arányosítottuk. Magasabb FOI érték a jövőre vonatkozó magasabb találatszámot, így jövő-orientáltságot, míg alacsonyabb FOI érték a múltra vonatkozó magasabb találatszámot, így múlt-orientáltságot jelent.



5. ábra:Google területi keresési trendek a „2013”, „2014”, „2015” kulcsszavakra (sötétebb kék magasabb keresési számot jelez), forrás: GoogleTrends

4. Adatbányászati alkalmazás - Kis- és középvállalati csődindex számítása

A feladat célja a magyar kis- és középvállalati szektorban történt rétegzetten reprezentatív felmérés adatai alapján az adatállományban megtalálható cégek esetében csődindex kalkulálása.(Szerb et al. 2014)
A feladat elvégzéséhez kialakítandó modellünknek tanuló adatbázisra, majd az eredményeink validálásához teszt adatbázisra volt szükség.A modellépítéshez használt teljes minta[i]1250 db csődbejutott cég és 3900 db működő cég 2012 és 2011-es évek mérleg és eredmény-kimutatásainak adatai, a cégekkel szemben kezdeményezett eljárások adatait és extra adatokat (TEÁOR, NUTS-2 elhelyezkedés) tartalmazta.
Az anonimizált teljes adathalmaz 1180 db cég adatait tartalmazta, amelyek között a 2013-ban csődbe jutottak megjelölésre kerültek. A teljes adathalmazt kettő részre bontottuk (tanuló, teszt), melyből a modellünk tesztelésére használt adathalmaz 506 db cég adatait tartalmazta. A teszthalmazon belül a csődbe jutott cégek aránya a minta 50%-át tette ki.
A modellünket egy neurális háló (auto multi-player perceptrons) és egy SVM (Support Vector Machine – Evolutionary, anova kernel) pontosság (accuracy) eredményének összehasonlításával választottuk ki, melyeket a tanuló adatainkon futtattunk. Első lépésben a modelleket a tanuló és a teszt adatok változatlanul hagyása mellett futtattuk. Második lépésben minden változó értéket normalizáltunk, majd harmadik lépésben az adatok mélyebb elemezése után csak bizonyos változók értékeinek normalizálását végeztük el, a többi változóérték transzformálása nélkül. A modell pontosság eredményei a normalizálás szükségtelenségét mutatták.
A teszt halmazon történő futtatások eredményei alapján az adatok normalizálása nélküli SVM modell alkalmazása mellett döntöttünk. Az így kialakított tanuló modellből (és a tesztelőből is) eltávolítottuk a cégekkel szembeni eljárásokkal kapcsolatos adatokat tartalmazó változókat, lévén azok a csődindex előrejelzésére használandó adatbázisban nem elérhetők. A végleges modellünk 82,41%-os pontossággal jelzi előre a csődeseményt.
Az előzőekben kialakított modell versenyképességi adatokon való futtatásához a modellben használt változók versenyképességi adatokból történő kalkulálására volt szükség.
Az előzőek alapján felépített adattáblánk 762 db céget tartalmaz. A kialakított csődindex optimista megközelítésű, azaz a magasabb érték a várható csődeseménnyel fordítottan arányos. Modellünk 617 cég esetében 0,5-ös csődindexnél magasabb értéket jelez, míg 144 cég esetében 0,5-ös értéknél kisebb, azaz csőd-kockázatosnak minősíti a vállalkozást. Egyetlen cég esetében nem képes a modellünk az előrejelzésre, melynek oka a kötelezettségek típusú adatok hiányában keresendő, így ez nem a modell gyengesége, hanem az előrejelzendő adatok hiányos előkészítésére utal.

Az1. táblázatban a modellünkben az öt legnagyobb súllyal résztvevő változó értékeinek átlagolásával számítottuk ki a tipikus csődkockázatú és csődkockázattól mentes cégek profilját.

Változók	Tipikus jók (index > 0,5)	Tipikus rosszak (index < 0,5)
Saját tőke aránya	0,52	-1,49
Befektetett eszközök aránya	0,38	0,30
Üzemi eredmény aránya	0,04	-0,16
Kötelezettségek aránya	0,39	2,22
Adózott eredmény aránya	0,14	0,79
Forgóeszközök aránya	0,56	0,58
Cég életkora	14	13

1. táblázat: Tipikuscsődkockázatú ésmentescégekadatai, forrás: sajátszámítás

4. Következtetések

Dolgozatunkban három különböző adatforrás és módszer (szövegbányászat, felhasználói keresések elemzése és adatbányászati tevékenység) alkalmazásával megmutattuk, hogy mind a strukturált, mind a strukturálatlan digitális adatok felhasználásával lehetőségünk van a valós világ mélyebb megismerésére. A három megközelítést a regionális versenyképesség kapcsolja egybe, vagyis annak a térségnek a vizsgálata, amelyben a kis- és középvállalkozások működnek. A használt technikák, adat- és szövegbányászat, a területi versenyképesség különböző aspektusainak vizsgálatához szükséges digitális adatforrások felhasználására, elemzésére biztosítanak lehetőséget.A „puha" típusú szöveges adatok használatával, azokat szociális érzékelőkként felfogva, a környezet egy reprezentációját alakítottuk ki a HírIndex segítségével. A kvantitatív adatokbólegyrészt a régióból indított Google keresések alapján a területegység jövő-orientációs mutatóját, másrészta térségekben működő vállalkozások csődkockázati besorolását számítottuk. Az eredményeinket térképi és szövegfelhős ábrázolással láttattuk.

5. További kutatási irányok

A hírek osztályzási modelljének korpuszát módosítani szükséges a gazdasággal, regionalitással kapcsolatos forrásokra történő fókuszálással. A szövegbányászat komplexitását további elemzési technikák bevonásával emelni szükséges.

A keresési trendek esetében további elemzési lehetőség kínálkozik a kialakított mérőszám (FOI) és a területi versenyképesség mérésére használt egyéb mutatók közti korreláció vizsgálatával. További kutatási irány mutatkozik a különböző keresési kulcsszavak bevonásának vizsgálatában.

A csődindex számításában az egyik ígéretes továbblépési lehetőség az adatkörök puha típusú (pl. szöveges) adatokkal való bővítésében kínálkozik.

Irodalomjegyzék

Creese, G.: Duo-Mining: combining data and text mining, DM Reviews, No. September, 2004

Fan, W. – Wallace, L. – Rich, S. – Zhang, Z., Tapping into the Power of Text Mining, Com. of ACM, 2005

Hey. T.: The Big Idea: The Next Scientific Revolution. Harvard Business Review. Nov 2010

Hu, M. – Liu, B.: Mining and Summarizing Customer Reviews, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Seattle, Washington, USA, 2004

Hu, X. - Liu, H.: Text analytics in social media, Mining Text Data - Aggarwal, C. – Zhai, C. (szerk.), Springer, 2012, 385-415. o.

McAfee, A., and Brynjolfsson. E.: Big Data: The Management Revolution. Harvard Business Review, 2012

Preis, T., Moat, H.S., Stanley,H.E. & Bishop, S.R.: Quantifying the Advantage of Looking Forward. Sci. Rep. 2, 350; 2012, DOI:10.1038/srep00350

Tikk D.: Szövegbányászat, TypoTex, Budapest, 2007

Sakaki, T. - Okazaki, M. – Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.

Szerb L., Csapi V., Deutsch N., Hornyák M., Horváth Á., Kruzslicz F., Lányi B., Márkus G., Rácz G., Rappai G., Rideg A., Szűcs P. K., Ulbert J.: Mennyire versenyképesek a magyar kisvállalatok? A magyar kisvállalatok (MKKV szektor) versenyképességének egyéni-vállalati szintű mérése és komplex vizsgálata, Marketing és Menedzsment 11/2014; XLVIII.(Különszám), 3-21.o.

Wu C.-H., Tzeng G.-H., Goo Y.-J., Fang W.-C.: A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. Expert Systems with Applications 32, 2007, 397–408. o.

Szerzők

Hornyák Miklós: Gazdaság-módszertani Intézet, Közgazdaságtudományi Kar, Pécsi Tudományegyetem, 7621 Pécs, Rákóczi út 80, Magyarország. E-mail: hornyakm@ktk.pte.hu

[i]Ezúton is szeretnénk kifejezni köszönetünket Pungor Gábornak a PTE KTK mesterszakos hallgatójának, aki a szakdolgozata elkészítése során begyűjtött adatokat, azok anonimizálása után a rendelkezésünkre bocsátotta.