



Milyen legyen egy adatbányászati szoftver elemzése?

Szommer K.

Budapesti Corvinus Egyetem, Számítástudományi Tanszék, 1093 Budapest, Fővám tér 8.

ÖSSZEFOGLALÁS

Az adatbányászati eszközök széles körben elterjedtek, szolgáltatásaik terén kezdenek homogénné válni. A szoftverelemzések és prezentálásuk nem követik a piaci igények változását. Olyan szempontokat elemeznek, amelyek kiegészítésre szorulnak. A cégek számára fontos, hogy a számukra megfelelő szoftvert válasszák, így pontosabb, részletgazdagabb elemzésekre van szükség. A jelenlegiekkel ellentétben olyanokra, amelyek egymással összevethetőek, a szoftverpiac fejlődését részleteiben képesek bemutatni. Visszatekintve az időben sokféle elemzés található, azonban nem jellemző rájuk, hogy átvesszik más módszerekből a hasznos részeket. Az elemzési módszereket megvizsgáltam és a megfelelőket egybeolvasztottam, hogy egy sikeresebb és részletgazdagabb, integrálható elemzési módszer jöjjön létre. Megadtam azokat az új szempontokat, amikkel az elemzéseket a cégek érdekében bővíteni kell. Meghatároztam az egyes kategóriák súlyait is, továbbá elkészítettem egy hozzá tartozó megfelelő grafikai megjelenítési mód alapjait is. Az így felvázolt elemzési és megjelenítési módszert más szoftverekhez is eredményesen fel lehet használni.

(Kulcsszavak: adatbányászat, elemzési módszer, ábrázolás)

ABSTRACT

What should the analysis of a data mining software be like?

K. Szommer

Corvinus University of Budapest, Department of Computer Science, H-1093 Budapest, Fővám tér 8.

The data mining tools are widely spreaded and started to become homogeneous in their field of service. The software analyses and presentations don't follow the changes of the demands on the market. Such aspects are analyzed, that must be amended. Companies want to choose the best data mining tool, that's why there is a need for more accurate and detailed analyses. Not like the current ones that cannot be compared with each other, it can show the details of the development/improvement of the software market. If we look back in time, many analyses can be found, but typically they did not mention the useful details of other methods. I examined the analysis methods and I created new criterias, that should be added to the analyses for the companies. I determined the weights of each category, furthermore I prepared the basics of a graphical presentation method as well. The aforementioned analysis technique and presentation method can also be used for other softwares.

(Keywords: data mining, analysis method, presentation)

BEVEZETÉS

A mai modern adatbáziskezelő rendszerek számára már nem jelent gondot a jelenlegi adatmennyiség gyűjtése és tárolása, így az értékes információkat tartalmazó adatokat el

tudjuk menteni. A problémát ezekből az adatokból a rejtett információ kinyerése jelenti. Az adatbányászati algoritmusnak gyorsnak, pontosnak kell lenni, a szoftvernek pedig megfelelően kell kommunikálnia a használt adatbáziskezelő rendszerrel. Mindez még nem elég, ugyanis jól szemléltethető és könnyen érthető formába kell hoznia a folyamat végén a kinyert tudást. (Cser, 2010) Különösen a publikusan elérhető adatbázisok esetén fontos a gyorsaság és a pontosság, mivel ez előnyt jelent a többi szereplővel szemben. Épp ezért fontos első lépésként a megfelelő szoftver kiválasztása, majd később a megfelelő módszer és algoritmus alkalmazása az adott problémára.

A tudásfeltárás során először megértjük az adott alkalmazási területet, valamint a felhasználási célokat (Bodon, 2010; Fayyad, 1996). Ezt követően kiválasztjuk a részadatbázisokat, amikből ki szeretnénk nyerni a tudást: ezek lesznek a forrás adatai. Ezeket a forrás adatokat megtisztítjuk, majd egy következő lépésben integráljuk. Az adatok bekerülnek egy adattárházba, ami kimondottan elemzéseket támogató adatbázis. A következőkben kiválasztjuk a megfelelő adatbányászati algoritmust, azt megvizsgáljuk, becsüljük a futási igényeit, majd alkalmazzuk. A kinyert információ alapján további finomításokat végzünk, az előző lépésekhez visszatérve. A megszerzett tudást összevetjük az elvárásokkal. A sikeres adatbányászati projektek azt mutatják, hogy van egy „egyharmad-kétharmad szabály”, amely azt mondja ki, hogy a munka akkor produkált jó eredményt, ha a kapott eredmény kétharmada az adott szakterületen dolgozók sejtéseit igazolja, egyharmada pedig teljesen új ismeretet szolgáltat (Cser, 2005). Látható, hogy a tudásfeltárás igen komplex folyamat. Amennyiben egy-egy lépést nem megfelelően támogat egy szoftver, az az egész folyamat hatékonyságát fogja rontani. Szükséges a megoldások megfelelő elemzése, hogy ki tudjuk választani a legmegfelelőbbet a rendelkezésre álló szoftverek közül.

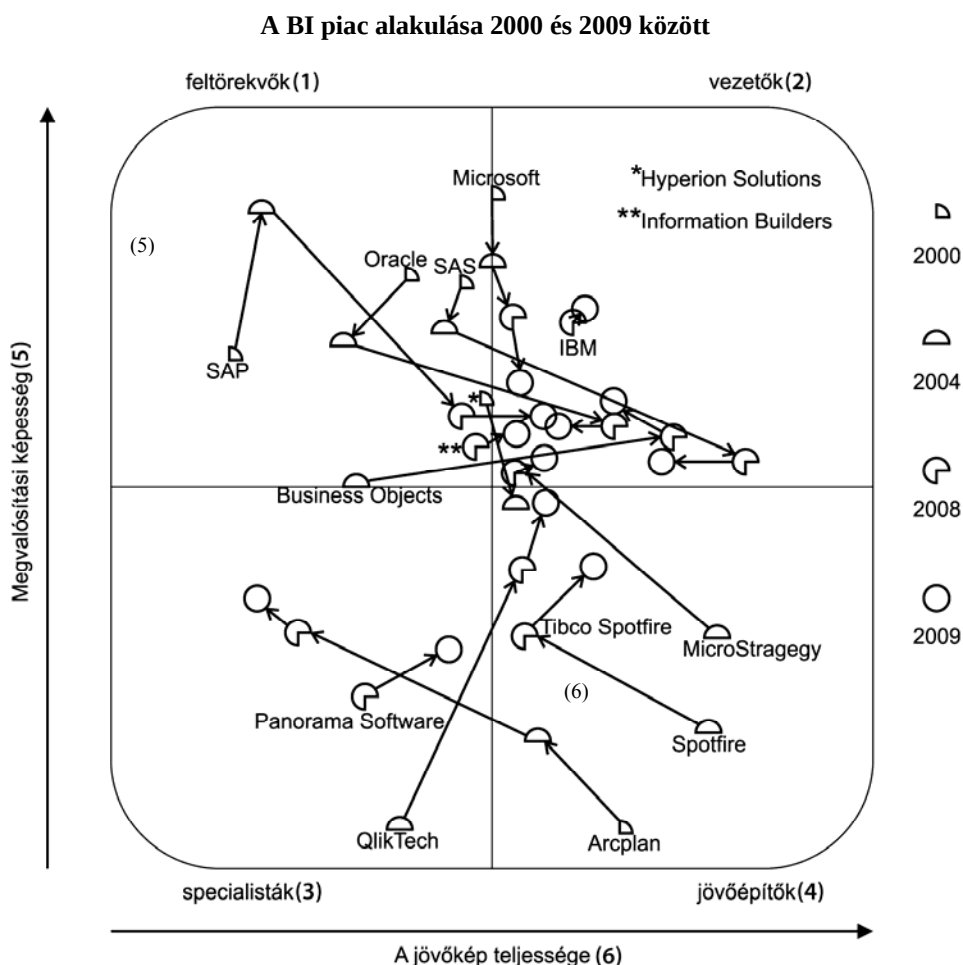
1998-ban a felmérések szerint (Elder, 1998; Goebel, 1999) még sok adatbányászati szoftver esetében az adatbázis-kapcsolatok limitáltak voltak. Mindegyik szoftver más területen volt jó: jelentéskészítés, adatbázissal való kapcsolódás, több platformon való futtathatóság, különböző algoritmusok támogatása, stb. Ma ez a különbség nem ilyen éles. Az utóbbi 10 évben óriási átrendeződés történt a BI piacon. Korábban csak a nagyvállalatok teheték meg azt, hogy korszerű BI technológiákat alkalmazzanak, azonban ma már egyre több kis cég jelenik meg a piacon, amelyek megoldásaival a kis- és középvállalkozók is magas szintű technológiához juthatnak (Cser, 2010). Az 1. ábrán a BI piac változása látható 2000. és 2009. között, amely több ábra összevetéséből jött létre. A teljes körök helyzetéből következtethető a 2009-es piac körülbelüli kiegyenlítettsége.

Az adattárházak piacán a három legnagyobb cég: az IBM, az Oracle és a Teradata versenyeznek a vezető szerepért (2. ábra). A vezetők többnyire a legkülönbözőbb méretű adattárházakhoz is teljes támogatást nyújtanak, valamint a piaci jelenlétük is stabilnak mondható. Többnyire rendelkeznek hardveres szövetségessel, így egy megrendelés esetén mind a szoftvert, mind a hardvert szolgáltatni tudják (Freinberg, 2010). A jövőépítők előrelátóan tekintenek az adattárház megoldásokra. Többnyire kevesebb ügyfelük van, a piaci jelenlétük nem annyira jelentős, mint a vezetőké, azonban itt jelennek meg azok a szolgáltatók is, amelyek valamilyen kivételes technológiával lépnek a piacra. Ezek az új megoldások a piacon akkor még nem bizonyítottak, így a szolgáltatóknak törekedniük kell arra, hogy megmutassák: az adott megoldás piacképes és előnyt jelent használóik számára. A specialisták csoportba tartozó szereplők legtöbbször kisebb, de igen jó adattárház-megoldást szolgáltatnak, azonban vagy regionális, vagy ipari specializációjuk miatt ez korlátozott.

Az adatbányászathoz kapcsolódó szoftvereket nehéz elemezni, értékelni, mivel ahogy az igények változnak a termékkel szemben, úgy cserélhetnek helyet a ranglistán,

így ha egyik szempontból nézve előnyösebb lenne az egyiket választani, egy másik szempontból ugyanaz a cég számára akár még hátrányos is lehet. Sok cikk foglalkozott már a BI piacon található megoldások vizsgálatával, azonban ez a piac nagyon gyorsan változik, amit az 1. ábra és a 3. ábra összevetése mutat be a legjobban. A 3. ábrán látható lista nem teljes, mivel speciális kritériumoknak kellett megfelelniük a cégeknek és a kínált alkalmazásaiknak, hogy bekerülhessenek az elemzésbe. (Kobielus, 2010) Nehéz lekövetni a fejlődési tendenciákat is (ki miben fejleszt, miben jobb és innovatívabb a többinél), továbbá csak egy általános 2, maximum 3 dimenzió szerinti ábrázolás a jellemző a szakirodalomban.

1. ábra

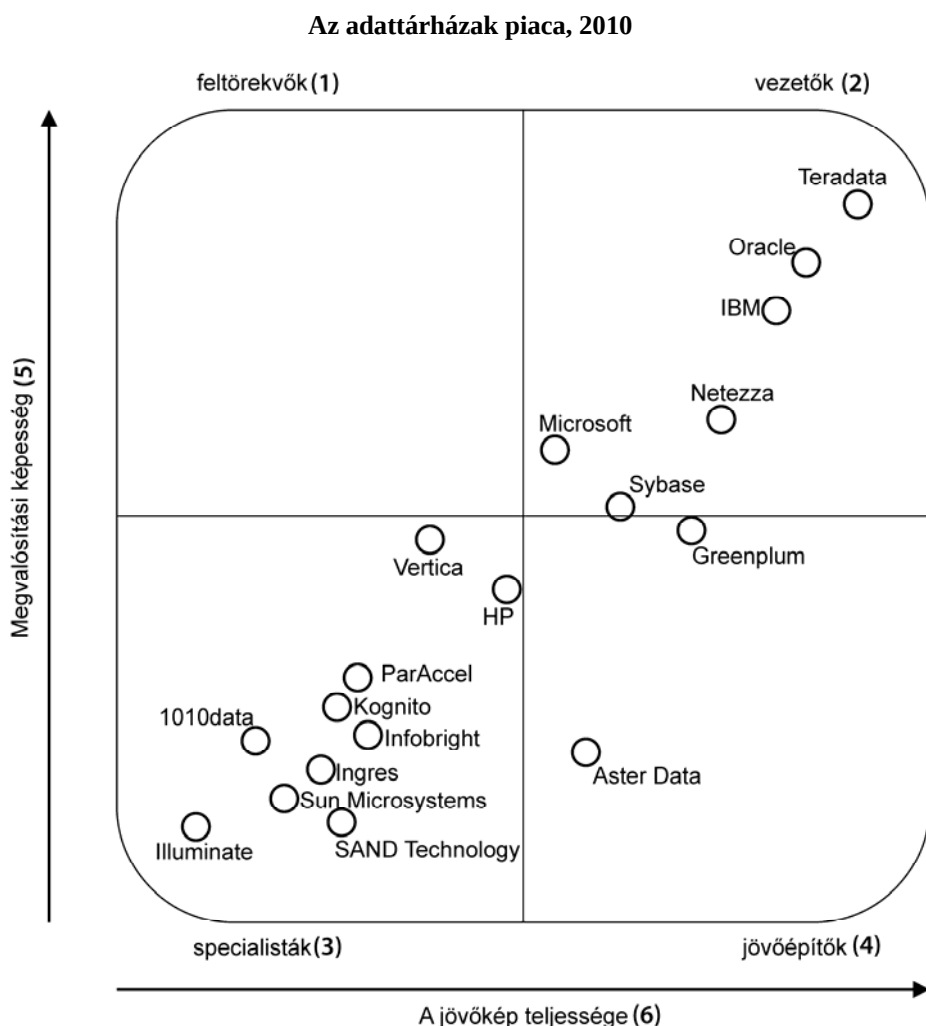


Forrás (Source): Fajszai (2010)

Figure 1: The changes on the BI market from 2000 to 2009

Challengers(1), Leaders(2), Niche players(3), Visionaries(4), Ability to execute(5), Completeness of vision(6)

2. ábra



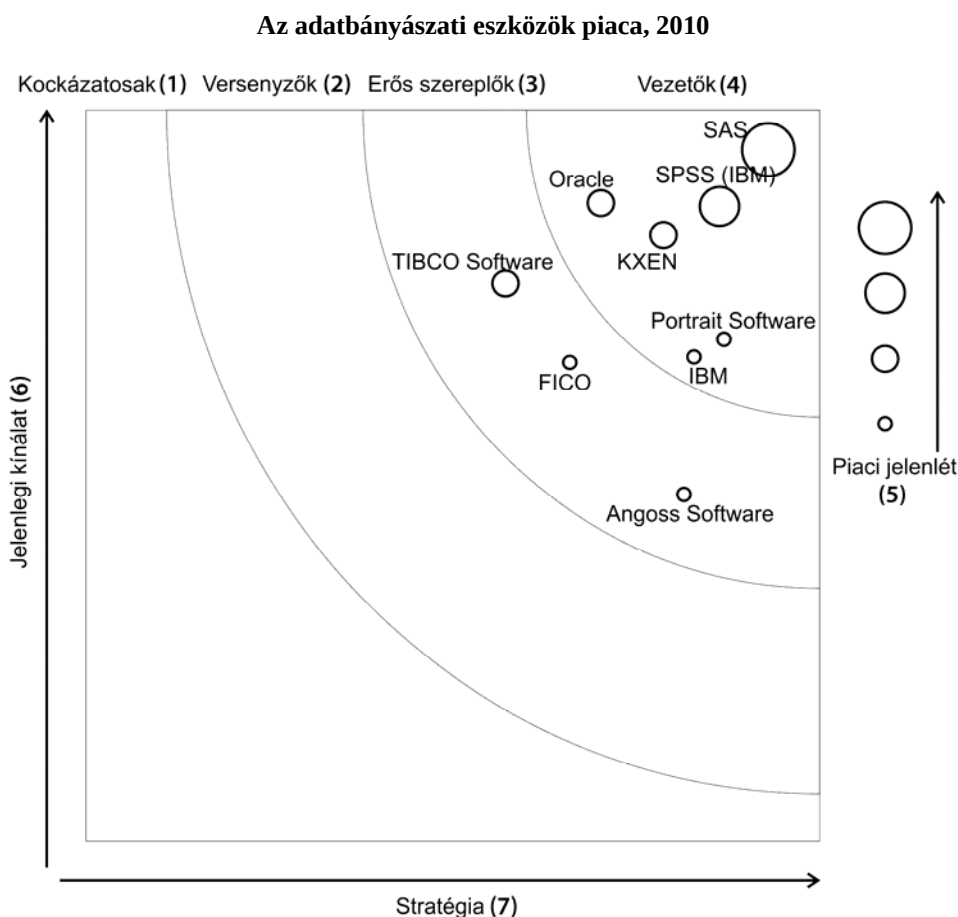
Forrás (Source): Gartner (2010)

Figure 2: The market of data warehouses, 2010

See Figure 1

Az adatbányászatot egyre több tudományterületen és iparban használják. Megtalálható többek közt az orvostudományban, az agráripárban, a banki és szolgáltatási szektorokban, és ez a szám csak nőni fog. Jellemzően az elemzés az egyre kisebb területekre is betör. A szoftverfejlesztőknek érdekében áll kiszolgálni az egyre kisebb partnereket, majd egy idő után már meg fog jelenni kkv-knál és egyéneknél is (Cser, 2010). Az elemzők elképzelhetőnek tartják, hogy néhány éven belül Software as a Service-ként is meg fog jelenni. Idővel csökkenni fog a használatukhoz szükséges képzettségi szint, valamint hosszú távon képi és videó információkat is eredményesen lehet majd bányászni. (Krauth, 2007)

3. ábra



Forrás (Source): *The Forrester Wave* (2010)

Figure 3: *The market of data mining softwares, 2010*

Risky bets(1), Contenders(2), Strong performers(3), Leaders(4), Market presence(5), Current offering(6), Strategy(7)

Egy mai cég esetében a versenyképesség fokozása érdekében fontos, hogy tudásmenedzsment szempontból is megvizsgáljuk a szoftvereket. Ebben az esetben nem elég csak a szakirodalomban található információk egyszerű összeillesztése egészszé, ugyanis többre van szükségünk: további információkra, valamint ezen információk közti kapcsolatokra. Azokra az ismeretekre is, amik csak a több szempontból, vagy ugyanabból a szempontból, de különböző időben történő értékelés során derülnek ki. A fellelhető elemzések mindegyike csak szűk látókörben vizsgálja az adatbányászati szoftvereket, valamint az elemzésekből hiányzik az egységesség, az összevethetőség.

A vizsgálatok során a főbb szempontok folyamatosan változnak, azonban ezek a változások egyre inkább a technikai oldal felé terelődnek. Régebbi cikkek (*Collier, 1999*;

Giraud-Carrier, 2003) többet foglalkoztak az akkori adatbányászati alkalmazások kipróbálhatóságával, megtanulhatóságával is, tehát a humán oldallal. Manapság ezen egyre kisebb a hangsúly, mivel a vizsgálatok feltételezik, hogy a szoftver használója az adott környezethez van szokva, pedig fontos, hogy olyan eszközt válasszon egy cég, ami a lehető legjobban megfelel a számára. Ezt pedig leginkább csak a kipróbálása után lehet eldönteni.

Jelen cikkemben az adatbányászati eszközök tudásszerzésre való alkalmasságának egy lehetséges mérési és ábrázolási módszerét dolgozom ki a szakirodalomban fellelhető anyagok segítségével. Ha az elemzők követik ezt a módszert, nem csak egymással lesz könnyebb összehasonlítani a szoftvereket, de képesek lesznek azokat időben is összevetni. Az egyes elemzésekhez tartozó igények változását is könnyebb lesz így bemutatni, továbbá így a külön elkészített elemzések is integrálhatóak lesznek. Egy, a szakirodalomból választott elemzésen keresztül mutatom be az általam elkészített módszertant. A cikk nem tartalmaz saját magam által végzett szoftverelemzést. A célom az eddig létező módszertanok vizsgálatával egy új elemzési módszer elkészítése volt, annak egy lehetséges megjelenítési megoldásával párosítva.

ANYAG ÉS MÓDSZER

A szakirodalomban többféle módszerrel elemzik a különböző szoftvereket. A következőkben az általam kiválasztott legfontosabb elemzési módszereket ismertetem, továbbá azokat a szempontokat, melyeket a szakirodalom az adatbányászati szoftverek elemzése során jelenleg vizsgál.

Ha az elemzett tulajdonságok szerint nézzük a módszereket, a legobjektívebb elemzési módszer az, amikor először meghatározzuk a vizsgált tulajdonságok főbb csoportosítási irányvonalait, majd azokat tovább bontjuk egészen addig, amíg már elegendő információ nyerhető ki úgy is, ha csak igaz-hamis válaszokkal jellemezzük őket. (*Giraud-Carrier*, 2003) Ezt az elemzési módszertant részben jónak tartom, ugyanis az ábrázolás során előnyt jelenthet az, ha előbb a főbb csoportosítás szerint láthatjuk a szoftvert, majd további lefűrés után jutunk el a részletesebb bontásig. Az igaz-hamis válaszok azonban a piac kiegyenlítetttsége miatt nem lennének megfelelőek. Minőségi jellemzést így nem tudunk adni ezzel a módszerrel szoftverekről.

A tulajdonságbontás során nem mindegy, hogy mikre térünk ki, miket részletezünk jobban, miket kevésbé. A kiértékelés egyrészt történhet technikai oldalról (*Hen*, 2008), ahol nem az a legfőbb szempont, hogy a szoftver milyen módszereket, algoritmusokat ismer, hanem az, hogy a rendelkezésre álló szűkös erőforrásokat (processzor, memória, merevlemez írási/olvasási sebesség, hálózat, stb.) milyen hatékonysággal használja, mennyire terheli le azokat. Ezen tulajdonságok elemzése kezd a háttérbe szorulni, pedig igencsak fontos tényező. Ma már elvárt, hogy egy adatbányász szoftver ne csak a tudásfeltárás lépéseit, hanem a későbbi munkákat is támogassa, így tudásmenedzsment szempontból a főbb kategóriák eltérnek a pusztán technikai oldalról történő vizsgálatoktól.

Jelenleg nem található az előbb felvázolt bontás szerinti elemzés napjaink adatbányászati szoftvereiről. Csak aggregált információkkal találkozhatunk, melyben a vizsgált főbb tulajdonságcsoporthoz a következők (*Collier*, 1999; *Hen*, 2008):

- Teljesítmény: Idetartozik a robusztusság, a más, szintén a KDD folyamatot támogató szoftverrel való kommunikáció, a lehetséges adatforrások, a szoftver architektúra, valamint hogy mely platformokon képes futni;

- Funkcionalitás: Itt az algoritmusok sokfélesége, a módszertanok, a validáció, az adattípusok kezelése, a jelentéskészítés és a modell exportálása kerül elemzésre;
- Használhatóság: Ez a kategória a felhasználókkal, a felhasználói felülettel, a hibakezeléssel, valamint a vizualizációval foglalkozik. Minél egyszerűbb a szoftver használata, minél átláthatóbb, annál hamarabb tanulhatnak meg vele az alkalmazottak dolgozni. Ezen szempont szerinti értékelés sok mindentől függ, de leginkább a vállalati kultúra határozza meg, hogy számára milyen típusú kezelőfelület, logikai működés a leggyorsabban tanulható, leginkább átlátható;
- Kiegészítő feladatok támogatása: Itt szerepel minden olyan további tulajdonság, amely a KDD egyes lépéseit támogatja, például az adattisztítást, a diszkretizálást, a szűrést, stb.;
- Erőforrás-kihasználtság hatékonysága: Nem csak a szoftver paramétereit kell vizsgálni, hanem azt is, hogy milyen hatékonysággal használja ki a rendelkezésre álló hardvert;
- Ha mind az elemzésbe, mind az ábrázolásba szeretnénk az egyes elemzett tulajdonságok fontosságát mutatni, akkor a szempontok súlyozására van szükségünk. (Collier, 1999) Ebben az esetben az egyes tulajdonságok az adott szempont szerinti elemzésnek megfelelően vannak súlyozva, így nem azonos mértékben befolyásolják a végeredményt;
- A pontszámok kialakításának a módja azonban még mindig nem teljes. Az egyes szoftverek tulajdonságokon belüli pontszámai többféle módon alakulhatnak ki: vagy választani kell egy referenciaszoftvert, amihez képest a többi viszonyítva van, vagy meg kell határozni az egyes tulajdonságon belüli lehetséges kategóriákat a hozzájuk tartozó pontszámokkal. Mindkét megoldásnál fontos az előzetes felmérés. A referenciaszoftver választása esetében azért, hogy egy olyan szoftver legyen kiválasztva, amihez képest a vizsgálatok során lesz jobb is és rosszabb is, a másik esetben pedig azért, hogy a lehetséges értékelési rekeszek teljes mértékben lefedjék a vizsgálni kívánt tulajdonságok fokozatait.

Az ábrázolás régen táblázatos formában történt. Ezekben a táblázatokban viszonylag sok tulajdonság fel volt sorolva, egyes szoftverelemzéseknel a tulajdonságok súlyai is szerepeltek. Ahogy a szoftverek száma megnőtt, a táblázat így vált számok átláthatatlan halmazává, egyszerűsítésre volt szükség. A jelenlegi ábrázolási módszer viszonylag egyszerű (Kobielus, 2010; Freinberg, 2010): kettő vagy háromdimenziós ábrán szemlélteti az egyes szoftverek tulajdonságait. A harmadik dimenzió a kettődimenziós leképezés miatt csak a szoftvert ábrázoló jelnek egy tulajdonsága (például mérete). Ezek az ábrák nem képesek több tulajdonságot használhatóan bemutatni, továbbá a tulajdonságok súlyozásai sem szemléltethetők.

EREDMÉNY ÉS ÉRTÉKELÉS

A szakirodalom elemzése során kiderült, hogy tudásmenedzsment szempontból az eddigi vizsgálatokban szerepelt kategóriák nem elegendőek. A cégeknek legtöbbször az elemzés kiértékelésén túl további vizsgálatokra is szükségük van a megfelelő választáshoz, azonban az általuk elvégzett vizsgálatok már igen specifikusak, legtöbbször nem teljes körűek, valamint egymással nehezen, vagy egyáltalán nem összevethetőek. Szükség van egy olyan elemzési módszerre, amellyel a cégek szempontjából teljes körű eredményeket lehet produkálni, és amely nem változik jelentősen évről évre, hogy a különböző felmérések időben is összevethetők legyenek.

Tudásmenedzsment szempontból azért, hogy az elemzés megfelelő legyen, tovább kell bővíteni a vizsgálati szempontokat a következőkkel:

- Kipróbálhatóság: Megfelelő véleményt a szoftverről valójában csak a kipróbálása után lehet mondani, ugyanis egy technikailag megfelelő, ám a vállalati szokásokhoz, az ottani kultúrába nem illeszkedő szoftver használatának hatékonysága csökkenhet;
- Szállító piaci helyzete: Vizsgálni kell a szoftverfejlesztő cég piaci helyzetét, fejlődésének tendenciáját, mivel igen sokáig tart berendezkedni egy adatbányászati alkalmazásra és maximális hatékonysággal üzemeltetni azt, így célszerű elkerülni azokat az eseteket, amikor az adott szoftvert szolgáltató cég jövője bizonytalan;
- Bővíthetőség: Egy olyan cégnél, ahol sok a tehetséges szoftverfejlesztő, előnyt jelent az, ha az adatbányászati szoftver igény szerint bővíthető, nagymértékben testreszabható. Ha ezt lehetővé teszi az alkalmazás, akkor a helyi fejlesztők az újonnan megjelent algoritmusokat is implementálhatják vagy a már meglévőket változtathatják. Egy kisebb szabadságot engedő megoldás lehet az, amikor az adatbányászati alkalmazást szállító cég készíti el ezeket a frissítéseket, majd vagy ingyenesen, vagy megvásárolható csomagok formájában rendelkezésre bocsátja;
- Ár: Ez egy további fontos paraméter a fenntartási költséggel együtt. Mivel hosszú távra kell tervezni, így jelentős költségkülönbségek adódhatnak az egyes opciók között;
- Egyéb: Sok egyéb lehetőséget is tartalmazhatnak az egyes szoftverek, amelyek a felhasználók munkáját segíthetik, de nem létfontosságúak. Ide tartozhat például a több nyelven elérhető kezelőfelület, az adatok mértékegységeinek automatikus átváltása, valamint az eredményekből történő automatikus prezentációgenerálás is.

Az egyes tulajdonságokat más-más súllyal kell szerepeltetni. Valamelyik fontosabb, valamelyik kevésbé fontos a vizsgálat szempontjából, de egyik sem elhanyagolható. Függetlenül attól, hogy melyik mérési módszert alkalmazzuk (abszolút, referencia-szoftveres), az egyes tulajdonságok súlyai a 4. ábrán láthatóak. Természetesen az egyes rekeszeket szoftvertípustól függően további részekre lehet bontani, azonban a túl sok rekesztől az elemzés a jelenlegi ábrázolási módszerrel már nehezebben lesz átlátható, így egy részletesebb felbontás esetén célszerű újabb diagram készítése, fókuszálva csak arra az egy tulajdonságra.

Hogy képesek legyünk még több tulajdonság szerint ábrázolni a szoftvereket, szükségünk van egy új ábrázolási módszerre. A szakirodalomban kevés szoftver összehasonlítása esetén legtöbbször táblázat formájában találhatóak a részletesebb elemzések eredményei, a különböző tulajdonságok pontszámai. A vizsgált szoftverek számának növekedésével ez a táblázat egyre átláthatatlanabbá válik, az elemzések eredményei egyre nehezebben lesznek érthetőek. Nem azt gondolom, hogy a táblázat elhanyagolható, hanem arra szeretném felhívni a figyelmet, hogy megfelelő grafikus megjelenítéssel sokkal átláthatóbbá tehetjük az elemzéseket. Egy olyan ábrázolási módszerre van szükség, amely figyelembe veszi az egyes tulajdonságok súlyozását, megkönnyíti az érthetőséget. Az ábrázolás több célnak is meg kell, hogy feleljen:

- Alkalmas kell, hogy legyen több szoftver összehasonlítására. Jól ki kell, hogy emelje az egyes szoftverek erősségeit, gyengeségeit, ezeket jól láthatóvá kell tennie;
- Be kell tudnia mutatni a pillanatnyi piaci helyzetet;
- Képes kell, hogy legyen az időbeli változások szemléltetésére;
- Meg kell, hogy oldja az idővel megnövekedett igények és a múltbeli igények szerinti elemzések összehasonlítását;
- Jól láthatóvá kell tennie az egyes kategóriasúlyok változását.

4. ábra

Az elemzés során használt egyes kategóriák súlyai

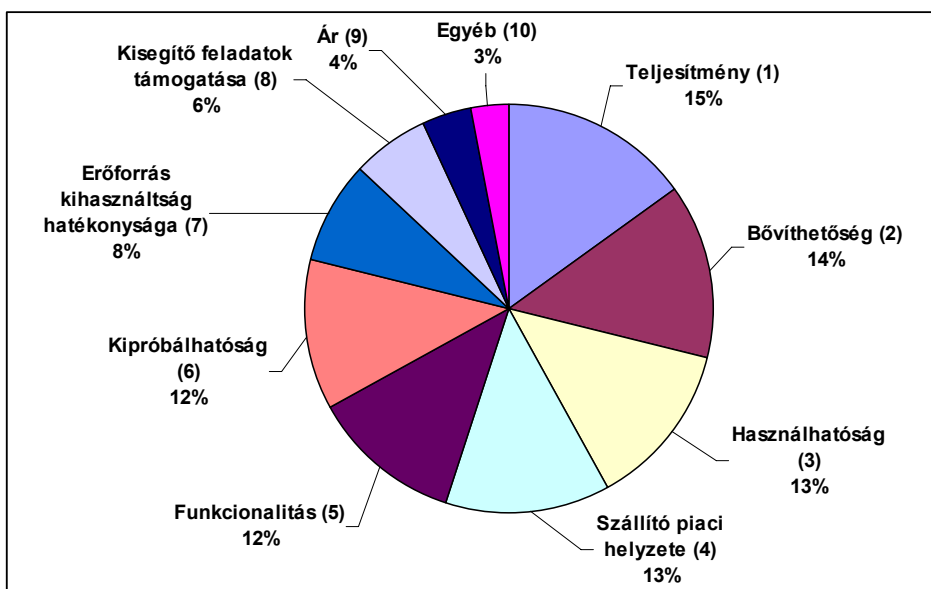


Figure 4: The weights of categories used by the analysis

Performance(1), Expandability(2), Utility(3), Supplier's market position(4), Functionality(5), Availability to try out(6), Efficiency of resource utilization(7), Availability of supporting functions(8), Price(9), Other(10)

Egy általam választott lehetséges megoldás erre az ábrázolási módra egy több rétegű pókháló-ábra készítése, amely tartalmazza a súlyozásokat. Egyes részei tetszés szerint ki/bekapcsolhatóak, átlátszó rétegei segítségével a szoftverek külön is össze-hasonlíthatóak a segítségével, vagy akár egyszerre látható rajta az összes. Ez az ábrázolási mód jól láthatóan kiemeli az egyes megoldások erősségeit, gyengeségeit. Amennyiben a felmérés során az abszolút pontozásos módszer kerül kiválasztásra, úgy képes az ábra a jelenlegi piaci kínálati minimumot is bemutatni az adott kategórián belül. Ha a pillanatnyi piaci helyzetet szeretnénk szemléltetni, tehát a szoftverek adott kategóriákon belüli teljesítőképességének szóródását, az egyes sávok határán lévő vékony vonal sűrűsödése-ritkulása ezt jól bemutatja, de a többi elemzésnél ezek a vonalak elhagyhatóak a jobb átláthatóság érdekében. Az időbeli változásokra is megoldást kínál ugyanúgy, mint a megnövekedett igények és a múltbeli igények közti különbségek bemutatására, vagy az egyes kategóriásúlyok változására. Az 5. ábrán látható pókháló-ábrázolás, amely Collier (1999) eredményeit ábrázolja, nagyon sok hasznos információt képes egyszerre megjeleníteni. A kihagyott részek információi nem álltak rendelkezésre a szoftverekről, így az ábra azon részei üresen maradtak. Azért választottam a pókhálóábra szemléltetésére Collier eredményeit, mert a szakirodalomban az adatbányászati szoftverekről az általam elkészített ábrázolási módszernek megfelelő,

pontozásos módszerrel elkészített friss elemzés nem található. Az összefésült képen jól kivehető a szoftverek pontszámának különböző tulajdonságokon belüli szóródása, valamint az a minimum, amit az elemzett szoftverek bármelyike nyújtani tud.

5. ábra

Collier (1999) eredményeinek ábrázolása pókhálóábrán

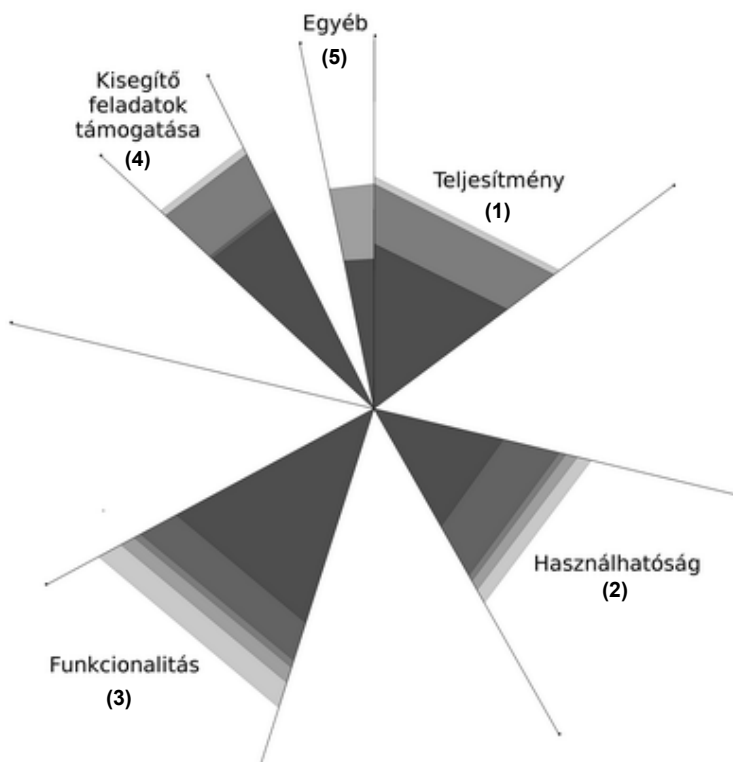


Figure 5: Presentation of the results of Collier (1999) on the spider-web diagram

Performance(1), Utility(2), Functionality(3), Availability of supporting functions(4), Other(5)

Az ábrázolásban a kategóriák összessége 360°-ot zár be, a nulla fok jelenti a függőleges tengelyen a pozitív irányt. A képpontok számozása a bal felső saroktól kezdődik, (0,0) koordinátával. Első lépésben ki kell számolni az egyes kategóriákra jutó szögeket, majd a következő vektor mutatja meg a vonalak irányát:

$$\begin{pmatrix} x + \cos(90^\circ - \sum \varphi) \cdot m \\ y - \sin(90^\circ - \sum \varphi) \cdot m \end{pmatrix} \quad (1),$$

ahol x, y a készíttendő ábra középponti koordinátái, $\sum \varphi$ az eddigi kategóriák által elfoglalt szögek összessége, m pedig az ábra nagyságának aránytényezője.

Az egyes kategóriák kezdő és záró vonalain a metszéspontra pedig a következő vektor mutat:

$$\begin{pmatrix} x + \cos(90^\circ - \sum \varphi) \cdot m \cdot \frac{p_e}{p_m} \\ y - \sin(90^\circ - \sum \varphi) \cdot m \cdot \frac{p_e}{p_m} \end{pmatrix} \quad (2),$$

ahol x, y a készítendő ábra középponti koordinátái, $\sum \varphi$ az eddigi kategóriák által elfoglalt szögek összessége, m pedig az ábra nagyságának aránytényezője, p_e az elért pontszám, p_m pedig a maximális pontszám az adott kategórián belül.

KÖVETKEZTETÉSEK

Az adatbányászati szoftverek piacán az erőhelyzet az elemzések szerint egyre kiegyensúlyozottabb lesz, így az eddig használt elemzési módszerek egyre kevésbé felelnek majd meg. Ha ez a folyamat így folytatódik tovább, négy-öt éven belül az egyéb kategória lesz a döntő a megfelelő szoftver kiválasztása során, tehát az, hogy a többihez képest milyen plusz funkciókkal rendelkezik. A kiegyensúlyozott helyzet miatt fontossá válik annak a minimális előnynek is a kihasználása, amelyet már csak egy részletesebb, pontosabb, súlyozást és egyéb szempontokat is tartalmazó elemzés során kaphatunk meg.

Az általam javasolt elemzési módszer segítségével pontosabb információkat kaphatunk az egyes tulajdonságokról. A különböző időben elkészített elemzések időben összevethetőek lesznek. Az egyes szoftverek különböző tulajdonságokon belüli fejlődési tendenciái is leolvashatóak lesznek, továbbá képesek leszünk összevetni a kereslet alakulását is (a súlyozás megváltozását). További előny lehet az is, hogy ha valaki egy-egy szempontra fókuszál csak az elemzésében, az beépíthető és felhasználható lenne. Ez a módszer nem csak az adatbányászattal kapcsolatos, hanem más típusú szoftverek összehasonlítására is használható.

A részletesebb, más szempontokra is kiterjedő elemzéseket már nem lehet megfelelően ábrázolni az eddig használt ábrázolási módszerrel, ezért szükség van egy újra, ami képes a megváltozott elvárásoknak is megfelelni. Jelen cikkhez a pókhálóábra Gimp és CorelDRAW szoftverek használatával készült.

Ez a fajta ábrázolásmód nem csak az adatbányászati, hanem más szoftverek összehasonlításainak a prezentálására is jó, például vállalatirányítási rendszerek, adatbáziskezelő rendszerek elemzésére. A Gimpel és CorelDRAW-al elkészült képek már nem módosíthatóak, rétegeik nem kapcsolhatóak ki/be, továbbá az egyes kategóriákon belüli további bontás, részletezés ezzel interaktív módon nem megoldható. Az ábrázolási bonyolultság csökkenthető lenne egy speciálisan erre a célra elkészített, mindenki által elérhető szoftverrel, amely interaktív, kinyíló diagramokat képes készíteni.

IRODALOM

- Bodon F. (2010): Adatbányászati algoritmusok. [Online] <<http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf>>
- Collier, K., Carey, B., Stautter, D., Marjaniemi, C. (1999): A Methodology for Evaluating and Selecting Data Mining Software. Proceedings of the 32nd Hawaii International Conference on System Sciences 1-11. p.

- Cser L., Fajszai B., Bodnár P. (2005): Üzleti tudás az adatok mélyén. Budapesti Műszaki Egyetem/IQSYS : Budapest
- Cser L., Fajszai B., Fehér T. (2010): Üzleti haszon az adatok mélyén. Adatbányászat. Alinea : Budapest
- Elder, J.F., Abbott, D.W. (1998): A comparison of leading data mining tools. Fourth International Conference on Knowledge Discovery & Data Mining, New York
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996): The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM., 39. 11. 27-34. p.
- Freinberg, D., Beyer, M.A. (2010): Magic Quadrant for Data Warehouse Database Management Systems. Gartner
- Giraud-Carrier, C., Povel, O. (2003): Characterising Data Mining software. Intelligent Data Analysis, 7. 181-192. p.
- Goebel, M., Gruenwald, L. (1999): A survey of data mining and knowledge discovery software tools. SIGKDD Explorations 1. 1. 20-33. p.
- Hen, L.E., Lee, S.P. (2008): Performance Analysis of Data Mining Tools Cumulating with a Proposed Data Mining Middleware. Journal of Computer Science, 4. 10. 826-833. p.
- Kobielski, J. (2010): The Forrester Wave: Predictive Analytics And Data Mining Solutions, Q1 2010.
- Krauth P. (2007): Üzleti Intelligencia, Információs Társadalom Technológiai Távlatai, 3. 109-131. p.

Levelezési cím (*Corresponding author*):

Szommer Károly

Budapesti Corvinus Egyetem, Számítástudományi Tanszék

1093 Budapest, Fővám tér 8.

Corvinus University of Budapest, Department of Computer Science

H-1093 Budapest, Fővám tér 8.

Tel.: 36-1-482-7498

e-mail: ifj.szommer.karoly@gmail.com