

**Bibliomining,
új eszköz a könyvtárak szolgálatában**

Tartalomjegyzék

1. Bevezetés.....	3
2. Célok, hipotézisek.....	4
3. Módszertan, a kutatás menete.....	4
4. Kutatás.....	5
4.1. Szakirodalmi áttekintés.....	5
4.1.1. Adatbányászat.....	5
4.1.2. Könyvtári adatbányászat - Bibliomining.....	10
4.1.2.1. Alkalmazható technikák, adatforrások.....	12
4.1.2.2. A könyvtári adatbányászat folyamata.....	13
4.1.2.3. Alkalmazási területek.....	18
4.1.2.4. A jelenlegi helyzet.....	21
4.1.3. Integrált könyvtári rendszerek.....	22
4.2. Integrált könyvtári rendszerek statisztikai moduljának elemzése.....	24
4.2.1. A Corvina.....	25
4.2.2. A Szirén.....	25
4.2.3. Az Aleph.....	26
4.2.4. A Huntéka.....	31
4.2.5. A LinLib.....	31
4.2.6. A Szikla.....	32
4.2.7. Az OliB.....	34
4.2.8. Összegzés.....	34
4.3. Könyvtárakban használható adatbányászati módszerek, algoritmusok.....	36
4.3.1. Adatfeltárás, statisztikai elemzés.....	37
4.3.2. Dimenziócsökkentés.....	41
4.3.3. Csoportosítást segítő algoritmusok.....	43
4.4. A bibliomining gyakorlati alkalmazása könyvtártípusok szerint.....	45
4.4.1. A közkönyvtárak.....	45
4.4.2. A szakkönyvtárak és a felsőoktatási könyvtárak.....	47
4.4.3. Az iskolai könyvtárak.....	48
4.4.4. A nemzeti könyvtárak.....	49
4.4.5. Az integrált könyvtárak.....	50
5. Összegzés.....	51
6. Bibliográfia.....	52
7. Mellékletek.....	58

1. Bevezetés

A mai tudományos életben nem ritka a különböző inter-, multi-, illetve transzdiszciplináris tudományágak megjelenése. Ilyen tudományterületnek tekinthető az adatbányászat is, ami napjainkban a gazdasági élet területén túlmutatva bármely társadalomtudományi területen is megjelenhet, és alkalmazása nagyon hasznos lehet az e területen dolgozók, kutatók számára. Dolgozatom célja bemutatni, hogy a könyvtárakban jelenleg termelődő információmennyiség milyen mértékben hasznosítható adatbányászati eszközökkel, módszerekkel. Jelenleg a könyvtárak által használt szoftverek nagy mennyiségű információt termelnek, tárolnak, ezek nagy részét azonban nem használják fel. Ezzel párhuzamosan a más területen, illetve piaci alapon szerveződő információgazdálkodással foglalkozó intézmények az általuk létrehozott adathalmazokat folyamatosan elemzik, és az eredmények alapján sikeresebbek tudnak lenni, fejlődni, megújulni tudnak. Napjainkban folyamatosan szükség van a változásra, az igények kiszolgálásának optimalizálására, előrejelzésére ahhoz, hogy célközönségünket el tudjuk érni, és meg is tartsuk. Ezekhez mind támpontot tud adni a könyvtárakban az adatbányászat. A könyvtárak számára most jött el az idő, hogy az ilyen eszközök használatát fontolóra vegyék, mivel a könyvtárak gépesítése, az adatok adatbázisban való tárolása nagyrészt már megvalósult, valamint lehetőség van a könyvtárak közötti adatcserére, a könyvtárosok számára is megszokottá vált az informatikai eszközök használata a könyvtárban zajló alapfeladatok ellátására, ezek kiterjesztésével pedig a könyvtári munka minden területe lefedhetővé válik. Az integrált könyvtári rendszerek kiaknázatlan lehetőségeket nyújtanak, hiszen a statisztikai adatok gyűjtésével, ennek továbbfejlesztésével, adatbányászati algoritmusok használatával új eredményekre tehetünk szert. Az információs piacon zajló versenyhelyzetben, illetve a személyre szabott szolgáltatások túltengése miatt különösen nehéz elérni a potenciális olvasókat, a könyvtárnak minden lehetséges eszközt meg kell ragadnia, mely segítheti működését, és az adatbányászat pontosan egy ilyen módszer lehet.

2. Célok, hipotézisek

Kutatásom elsődleges célja, hogy megismerkedjek azokkal a módszerekkel, szoftverekkel, gyakorlatokkal, amelyek az adatbányászat területén a könyvtárakhoz kapcsolódnak, illetve hogy felmérjem milyen módon építhetők be a jelenleg használt könyvtári rendszerekbe az adatbányászati módszerek. Kutatásom további célja felmérni, hogy hogyan különülnek el az egyes könyvtártípusok adatbányászati szempontból, valamint megvizsgálni az integrált könyvtári rendszerek statisztikai moduljait, mennyire használhatóak a könyvtári adatbányászat területén. Hipotézisem, hogy napjainkban bármennyire is elterjedt a piaci szférában az adatbányászat, a magyar könyvtárügy nem ismerte fel ennek jelentőségét, így nem is terjedt el ennek használata, illetve nem épültek be a jelenleg használt integrált könyvtári rendszerek statisztikai moduljába sem az adatbányászati algoritmusok. Hipotézisként fogalmaztam meg emellett, hogy léteznek olyan adatbányászati algoritmusok, melyek hasznosíthatóak a könyvtárakban is. Dolgozatomban ezen módszereket, algoritmusokat is be szeretném bemutatni.

3. Módszertan, a kutatás menete

Kutatásom során áttekintettem az adatbányászattal, valamint a könyvtári adatbányászattal foglalkozó nemzetközi és hazai szakirodalmat, és különböző integrált könyvtári rendszerek statisztikai moduljait. A szakirodalom feldolgozása után a könyvtárakban használható adatbányászati, bibliometriai szoftvereket térképeztem fel. Vizsgálataim alapján következtetéseket vontam le arra vonatkozóan, hogy mely adatbányászati algoritmus alkalmas könyvtári felhasználásra, mely integrált könyvtári szoftver statisztikai modulja terjeszthető ki adatbányászati célokra, illetve milyen különbségek fedezhetők fel a különböző könyvtártípusok esetén az adatbányászat használata területén.

4. Kutatás

4.1. Szakirodalmi áttekintés

4.1.1. Adatbányászat

Ahhoz, hogy az adatbányászatról beszélhessünk, ismernünk kell azt a technológiai, műszaki, adattárolási környezetet, amely az adatbányászat létrejöttének időszakában uralkodott a világban.

Az adatbányászat megjelenése elsődlegesen az információtechnológia fejlődésének köszönhető. Az 1950-es évektől kezdve a számítástechnika, az információtudomány, és a hozzájuk kapcsolódó területek óriási fejlődésnek indultak. A számítástechnikai eszközök kezdetben nagyon kevés adatot voltak képesek feldolgozni, és a számítási igényük is nagy volt. A technika fejlődésével, a második, illetve harmadik generációs számítógépek megjelenésével az adatok tárolása is kifinomultabbá vált.¹

Elég csak arra gondolnunk, hogy a terem méretű számítógépeket felváltották a gyorsabb, kisebb méretű, hatékonyabb eszközök, a lyukkártyát, mágnesdobra, majd mágnesszalagra cserélték, az adatbázisok terén pedig a kezdetleges szöveg alapú fájlkezelést felváltották először a hierarchikus és hálós adatmodellt használó adatbázisok, majd az 1980-as években megjelentek a relációs adatmodellen alapuló adatbázisok is.^{2 3}

Az eszközök ilyen mértékű fejlődése ahhoz a folyamathoz vezetett, hogy azonos áron egyre jobb és jobb teljesítményű számítógéphez lehetett hozzájutni, ezzel párhuzamosan pedig egyre nagyobb tárolókapacitással számolhattunk, ennek

¹ Katona Endre: Bevezetés az informatikába. Budapest, Panem, 2004.

² Csajbók Zoltán, Sütő Gergely: A számítástechnika története. Budapest, ÉGSZI-Scola Alapítvány, 1992.

³ Relációs adatbázis: egyedi nevekkkel ellátott táblák halmaza. Mindegyik tábla attribútumok (oszlop vagy mező) egy halmaza, és szokásosan többesek (tuples) (sor vagy rekord) nagy elemszámú halmazát tárolja. Egy relációs tábla minden ilyen többese ilyen objektumot reprezentál, amelyet egy egyedi kulcs, attribútumértékek egy halmaza azonosít. A relációs adatbázishoz gyakran készítenek szemantikus adatmodellt, például az egyedkapcsolat (ED) adatmodellt (Entity-Relationship, ER), amely az adatbázist egyedek egy halmazával és a közöttük fenálló kapcsolatokkal modellezi. (Jiawei Han- Micheline Kamber Adatbányászat: koncepciók és technikák)

következménye pedig az adattárolás költségének exponenciális csökkenése lett, így az adatokhoz való hozzáférés is egyre olcsóbbá vált.⁴

A számítástechnika kezdeti időszakában a cél csupán az adatok tárolása volt, azonban az idő előrehaladtával felmerült az igény a nagy méretű adatbázisokban való keresés megkönnyítését célzó eszközökre. Az első ilyen eszközök a strukturált lekérdező nyelvek voltak. Ezek közül a legismertebb és legelterjedtebb az SQL (Structured Query Language), amelyet 1974-ben implementált Edgar F. Codd ⁵, kifejezetten a relációs adatbáziskezelőkben tárolt információk könnyebb hozzáféréséhez. Codd észrevette, hogy a kezdetleges adatfeldolgozó programokban rengeteg rutinszerűen ismétlődő algoritmikus elem van, tehát ha ezek az algoritmusok már előre definiáltak, a lekérdezés létrehozójának elég a kívánt eredményt megadni, az algoritmikus lépéseket az SQL automatikusan össze tudja állítani.⁶ Mivel így a lekérdezés létrehozása sokkal kevesebb időt vett igénybe, illetve az SQL magas számolási sebességet tudott elérni, és alacsony hibakockázattal működött, rendkívüli gyorsasággal terjedt el, 1986-ban pedig szabvánnyá vált.

Tudnunk kell azonban azt is, hogy a lekérdező nyelvek nem alkalmasak arra, hogy minden, a felhasználó számára szükséges információt kinyerhessünk a különböző adatbázisokból. Ennek oka egyrészt a hatalmas adattömegek feldolgozásának nehézségében rejlik, másrészt pedig a speciális problémák modellezésének nehézségében. Az információtechnológia fejlődése lehetővé tette ugyan nagy mennyiségű adatok tárolását, azonban az internet elterjedése elősegítette, hogy könnyen, gyorsan bárki létre tudjon hozni tetszőleges mennyiségű információt, ez azonban olyan fokú „információgyártási” folyamatot eredményezett, melynek következtében napjainkban szinte feldolgozhatatlan mennyiségű adat jön létre. (Sok nemzetközi szervezetnél például egy hét alatt több információ jön létre, mint amennyit egy ember az élete során fel tud dolgozni.⁷) Az interneten keletkező adatmennyiség, valamint a nagy cégeknél létrejövő információtömeg nagyban elősegíti azt, hogy a világon a jelenleg a tárolt adattömeg

⁴ Sramó András: Adatbányászat és statisztika . In: Statisztikai szemle, 1999. (77. évf.) 5. sz. 350. p.

⁵ Ullman, Jeffrey David: Adatbázisrendszerek: Alapvetés. Bp., Panem, 2009. 2. átdolg. kiad

⁶ Celko, Joe: SQL felsőfokon. Budapest, Kiskapu, 2002..

⁷ Adriaans, Pieter , Zantinge, Dolf: Adatbányászat. Budapest, Panem, 2002.

körülbelül másfél évente megduplázódik, csak a 2008-as évben 3 kvintillió 892 kvadrillió 179 trillió 868 billió 480 milliárd 350 millió bitnyi új digitális információ keletkezett⁸, amennyi adat körülbelül 711 milliárd 4,7 gigabájt adattartalmú DVD lemezen férne el.

Ebben a hatalmas információmennyiségben nehezen tudunk eligazodni, egyre több adat egyre kevesebb információt jelent számunkra. Emellett pedig a speciális problémák még abban az esetben sem modellezhetők le strukturált lekérdezésekkel, ha kisebb méretű adatállományról van szó. Az ilyen típusú problémák esetén lehet hasznos az adatbányászati technológiák alkalmazása.

De *mi is az az adatbányászat*? A fogalmat a szakirodalom többféleképpen határozza meg:

- Dr. Abonyi János szerint: „Az adatbányászat egy olyan döntéstámogatást szolgáló folyamat, mely érvényes, hasznos, és előzőleg nem ismert, tömör információt tár fel nagy adathalmazból.”⁹
- Pang – Ning Tan- Michael Steinbach – Vipin Kumar Adatbányászat című művükben a következő értelmezést használják: „Az adatbányászat egy olyan technológia nagy mennyiségű adat feldolgozására, amely hagyományos adatelemzési módszerek és kifinomult algoritmusok elegye.”, „Az adatbányászat az a folyamat, amellyel hasznos információ fedezhető fel automatikus módon nagy adattárakban.”¹⁰
- Jiawei Han- Micheline Kamber Adatbányászat: koncepciók és technikák című művében pedig a következő fogalom jelenik meg: „Az adatbányászat a tudás nagy mennyiségű adatból történő kiválasztása, kibányászása.”¹¹

⁸ Másfél évente megduplázódik a digitális univerzum
http://hvg.hu/tudomany/20090519_digitalis_univerzum (Letöltve: 2013.12.14)

⁹ Abonyi János: Adatbányászat a hatékonyság eszköze: Gyakorlati útmutató kezdőknek és haladóknak. Budapest, ComputerBooks, 2006. 10.p.

¹⁰ Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin: Adatbányászat: Alapvetés. Budapest, Panem, 2012. 3 p.

¹¹ Jiawei Han, Micheline Kamber: Adatbányászat: koncepciók és technikák. Bp., Panem, 2004. 26 p.

A felsorolt definíciók egyetértenek abban, hogy nagy mennyiségű adattömegből szelektáljuk az információt, nem hagyományos eszközök segítségével. Dolgozatomban az adatbányászatot a következő fogalom alapján fogom használni: Az adatbányászat gyakorlati haszonnal járó információk kinyerése adatbázisokban tárolt nagy tömegű adatokból, statisztikai és mesterséges intelligencia módszerek segítségével. Főként az üzleti, gazdasági, információtechnológiai életben alkalmazzák, de minden olyan területen használható, ahol nagy mennyiségű adatot tárolnak, és vannak olyan döntési szituációk, amelyek megoldása nem egyértelmű.

Az adatbányászatot a szakirodalomban sok helyen azonosítják a tudásfeltárás (Knowledge Discovery in Databases) folyamatával, máshol pedig a tudásfeltárás részeként jelenik meg. A tudásfeltárás a nyers adatoknak hasznos információ való átalakításának a folyamata, melynek lépéseit az alábbi ábrán láthatjuk:



1. ábra
A tudásfeltárás folyamata¹²

Az adatbányászat mintákat keres a nagy mennyiségű adathalmazban, ehhez az adatoknak egységes szerkezetűnek, megtisztítottaknak, illetve relevánsnak kell lenniük. Az adatbányászati algoritmusok ötvözik a gépi tanulás alapelveit, a statisztika, illetve a mesterséges intelligencia algoritmusokban rejlő tudásfeltárási lehetőségeket.

Az alábbiakban szeretnék több, az adatbányászathoz kapcsolódó fogalmat definiálni, melyeket dolgozatom későbbi fejezeteiben használni fogok.

Az adatbányászati algoritmusok bármely adatbázison használhatóak, melyek strukturáltan szervezik adataikat, lehetnek ezek relációs, objektumorientált vagy tranzakciós adatbázisok, azonban leghatékonyabban az adattárházak esetén

¹² Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin: Adatbányászat: Alapvetés. Budapest, Panem, 2012. 3.p.

használhatóak. Az első adattárházat Bill Inmon hozta létre az 1980-as évek végén, ezért az alapdefiníciót is ő határozta meg: „Témaorientált, integrált, az adatokat történetiségében tároló adatrendszer, amelynek fő célja az adatokból történő hatékony információkinyerés biztosítása, elsősorban a döntéshozatali folyamatok támogatása érdekében.”¹³ Ez a definíció megadja az adattárház elméleti alapjait, azonban a gyakorlati alkalmazásban hasznos fogalmat Jiawei Han fogalmazza meg: „Az adattárház egy olyan információtár, amely különböző adatforrásokat egységes séma szerint integrál, és rendszerint fizikailag nem osztott környezetben valósul meg.”¹⁴ Tehát egy adattárház egyesíti mindazon, egy szervezet különböző formában tárolt adatbázisaiban tárolt információkat, egységes szervezetbe rendezi őket, és lehetőséget biztosít együttes kezelésükre, felületet biztosít ahhoz, hogy közös kereséseket végezzünk benne.

Az adattárházban az adatok úgynevezett adatkockákban tárolódnak. Az adatkockák többdimenziós tömbként értelmezhetőek, ahol az egyes dimenziók az adattárház egy-egy változójának felelnek meg, (ezek egy vállalat esetén például a következők lehetnek: telephely, idő, termékek) a dimenziókban pedig cellák helyezkednek el, melyek az egyes dimenziók mentén létrejövő aggregált értékeket tartalmaznak.¹⁵

Az adattárház adatkocka koncepcióját használja az OLAP (Online Analytical Processing) is. Az OLAP, online analitikus feldolgozást segítő rendszerek dinamikus felületet biztosítanak a felhasználó számára, melyen egyszerű egérgattintásokkal tudják az adatstruktúrák mélyén lévő információkat feltárni, jelentéseket, reportokat készíteni. OLAP rendszerek esetében az egyik legfontosabb fogalom a **lefűrés**. Ez azt jelenti, hogy a dimenziók hierarchikus felépítésűek, az adatokat egy-egy kattintással felfelé összegezni, lefelé szétbontani lehet (például negyedévről havi bontásra, vagy termékcsoporthoz termék bontásra).¹⁶ Az OLAP technológia előnye, hogy gyors, és hatékony, valamint adatbáziskezelési, lekérdezéstervezői ismeret nélkül is használható.

¹³ Fajszai Bulcsú - Cser László: Üzleti tudás az adatok mélyén: Adatbányászat alkalmazói szemmel. Budapest, Budapesti Műszaki és Gazdaságtudományi Egyetem, 2004. 24.p.

¹⁴ Jiawei Han - Micheline Kamber: Adatbányászat: koncepciók és technikák. Bp., Panem, 2004.34 p.

¹⁵ Jiawei Han - Micheline Kamber: Adatbányászat: koncepciók és technikák. Bp., Panem, 2004.35. p

¹⁶ <http://www.lbseconet.hu/vir/vir.htm> (Letöltve: 2013.12.15)

4.1.2. Könyvtári adatbányászat - Bibliomining

De miért is olyan fontos az adatbányászat könyvtárakban való alkalmazása? A könyvtárak a tudásgazdálkodás intézményei, így létezésük kezdeteitől fogva hatalmas mennyiségű információt tároltak. Kezdetben ezek az információk maguk a gyűjtött dokumentumok voltak, a bennük rejlő tartalom. Azonban attól a pillanattól fogva, hogy elkezdtek leltárba venni a könyveket, könyvlajstromokat létrehozni, már nem csak a könyvekben rejlő információkról beszélünk, hanem magukról a könyvekről szóló információkról is. A könyvnyomtatás elterjedésével nem csak a könyvek mennyisége nőtt meg, hanem a könyvekről szóló információk mennyisége is többszöröződött. Az egyre több könyv, időszaki kiadvány létrejöttével megnőtt az igény ezek számbavételére is, erre vállalkozik a könyvészeti műfaj, a bibliográfia, a könyvleírás. A könyvlajstromok, katalógusok, bibliográfiák mellett mindig is rengeteg információ keletkezett a könyvtárakban, amelyek nagy részét a 20. század közepe előtti időszakban nem tárolták le, illetve nem használták fel elemzések készítésére, nem lett letárolva, így felhasználva sem.

Az informatika előtti időkben a keletkező információk egy részét csak nehézkesen lehetett visszakeresni. A katalógusokban minden könyvet megtalálhatunk, a benne való keresés ugyan hatékony, de időigényes, és az egyszerre több szempont szerinti keresés nem megoldható. Ebben az időben az olvasók kölcsönzéseinek visszakeresése sem volt megoldott, egyetlen alternatíva a füzetes kölcsönzés volt. Ebben az esetben az olvasó által kezdeményezett kölcsönzéseket egy füzetben külön kezelték, ezzel kimutatható volt egy olvasó összes kölcsönzése. Arra azonban ez a technika sem ad választ például, hogy milyen módszerrel találjuk meg az összes olyan művet, ami ezek alapján a kölcsönzések alapján érdekelheti az olvasót. Emellett pedig számtalan más feladat – például a könyvek beszerzéséhez kapcsolódó tevékenységek – is papír alapon szerveződtek.

A keletkező információk típusait a következőképpen határozhatjuk meg: könyvekhez kapcsolódó információk, olvasókhoz kapcsolódó információk, könyvek és

olvasók kapcsolata, kereséshez kapcsolódó, olvasók tájékoztatásához kapcsolódó információk, könyvtári munkafolyamatokhoz kapcsolódó információk.

A számítástechnika megjelenésével, majd a könyvtárakban való elterjedésével sok feladat leegyszerűsödött, a könyvekről információkat adó bibliográfiai rekordokra már rákereshetünk, az olvasók kölcsönzési tranzakciói, adatai, illetve a különböző munkafolyamatok, így a szerzeményezés, katalogizálás is elektronikusan zajlik, a hozzájuk kapcsolódó információknak birtokában vagyunk. Ezek olyan információk, melyeket a korábbi időszakban, ha nehezebben, időigényesebben is, de megtalálhattunk. Az adatbányászat könyvtári alkalmazása olyan kérdésekre is választ tud azonban adni, amit még fel sem tettünk. Olyan problémákat tudunk megoldani, amiről nem is tudjuk, hogy léteznek. Emellett pedig természetesen az általunk ismert problémákra tud hatékonyabb, gyorsabb, megfelelőbb megoldást adni. Ennek megvalósíthatósága azon alapul, hogy az adatbányászati algoritmusok felül tudják múlni az emberi agyat modellalkotási szempontból. Tudjuk azt, hogy az emberi agy azon képessége, amellyel modellt hoz létre, képes összefüggéseket találni a meglévő adatok között. Ez azonban messze elmarad az adatok tárolására, összegyűjtésére, és a nagy adathalmazok közti kapcsolatok meghatározására való képességünktől.¹⁷ Amennyiben nagy mennyiségű adatok között kell kapcsolatokat definiálnunk, egy intézmény működését kell optimalizálnunk, vagy egy ügyfél igényeit kell maximálisan kielégítenünk, az adatbányászat célravezető eszköz lehet, nem csak üzleti, hanem könyvtári környezetben is.

¹⁷ Wormell, Irene: Az informetria és a bibliográfiai adatok használata a stratégiai tervezésben. In: Mikulás Gábor (szerk.): Információból üzleti érték: az információbróker környezete és munkája. Budapest, Magyar Információbrókerek Egyesülete, 2006. 321–351. p.

4.1.2.1. Alkalmazható technikák, adatforrások

A könyvtári adatbányászat alapvetően három fő technikát használ: ezek a bibliometria, a statisztika, illetve az adatbányászat. Dolgozatomban ezek közül főként a statisztikai és adatbányászati eszközök használatára fogok fókuszálni. A könyvtárakban zajló adatbányászat adatforrásai a következők lehetnek: adattárházak, integrált könyvtári rendszerek különböző moduljai (szerzeményezési, katalogizálási, kölcsönzési, statisztikai), illetve a könyvtárak web szervereinek naplófájljai.¹⁸ Amellett az adattárházba lehet integrálni még a felhasználók demográfiai adatait, amelyet kérdőív segítségével kaphatunk meg, valamint különféle statisztikai adatokat a statisztikai hivatal adattárából. A könyvtári adatbányászatban megjelenő adatforrásokat, és ezek kapcsolatait az alábbi ábra mutatja be¹⁹:

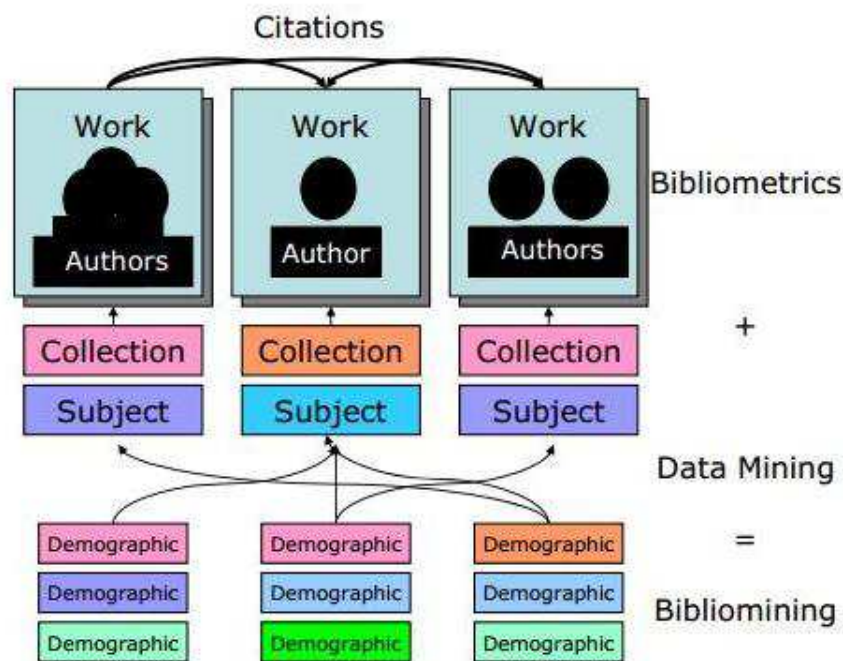


Figure 4: Data for Bibliomining

2.ábra

A könyvtári adatbányászat adatforrásai, kapcsolatai

¹⁸ Nicholson, Scott: The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services In: Information Processing & Management, 2006. (42.évf.) 3. sz. 785-804. old.

¹⁹ U.o.

4.1.2.2. A könyvtári adatbányászat folyamata

Jiawei Han és Michaline Kamber a tudásfeltárás folyamatát (melyet sokan adatbányászatként definiálnak) a következőképpen adja meg²⁰:

1. Adattisztítás
2. Adatintegrálás
3. Adatkiválasztás
4. Adat-transzformáció
5. Adatbányászat
6. Minta kiértékelése
7. Tudásmegjelenítés

Ahogy látjuk, e szerint a meghatározás szerint maga az adatbányászat is a folyamat egyik lépcsőfoka. Kutatásom során a könyvtári adatbányászat folyamatának meghatározásakor ebből az alapdefinícióból indultam ki, és ezt bővítettem ki, úgy, hogy a könyvtárakban zajló tudásfeltárás folyamatának meghatározására is alkalmas legyen. Az általam definiált folyamat megalkotása emellett elsősorban Scott Nicholson fogalommeghatározásán alapul. Nicholson a Syracuse-i egyetem Információs tudományok karának professzora foglalkozott eddig a legtöbbet ezzel a területtel, és ő említette ezt először 2002-ben egyik előadásában. Ő volt az első, aki a könyvtárakban zajló adatbányászatot bibliominingnak nevezi, a 2003-ban megjelent *The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making* című cikkében²¹. Dolgozatom további részében a könyvtári adatbányászat kifejezés helyett én is e terminust fogom használni. Ebben a cikkben Nicholson a bibliomining folyamatát a következő lépésekben határozta meg:

1. Hasznosítási területek definiálása
2. Külső és belső adatforrások kiválasztása
3. Adatok összegyűjtése, megtisztítása, adatok anonimmá tétele, adattárház feltöltése.
4. Megfelelő elemző eszköz kiválasztása

²⁰ Jiawei Han - Micheline Kamber: Adatbányászat: koncepciók és technikák. Bp., Panem, 2004. 27.p.

²¹ Nicholson, Scott: *The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making*. In: *Information Technology and Libraries*, 2003 (22.évf.) 4. sz. 146-151.p.

5. Mintázatok keresése az adatok között adatbányászati eszközökkel, jelentések készítése hagyományos elemző eszközökkel
6. Eredmények implementálása

A célterületek meghatározásánál a legfontosabb, hogy eldöntsük, hogy létezik-e valamilyen konkrét probléma, amit meg szeretnénk oldani, vagy csak egy specifikus terület van, aminek a működésén javítani szeretnénk. Az előbbit feltáró, az utóbbit bejósoló típusú adatbányászatnak nevezzük.²² Bejósoló típusú adatbányászatot használunk például, ha meg szeretnénk vizsgálni, hogy könyvtárunk távoli elérésű szolgáltatásait mennyire használják az olvasók, vagy hogy a gyűjteményünk megfelel-e az olvasók igényeinek. Nincsen egységes válasz arra, hogy melyik típust érdemes használni, de azt mindenképpen el kell mondani, hogy a bejósoló típusú adatbányászat abban az esetben célravezető csak, amennyiben nagyobb mennyiségű adattal rendelkezünk, vagy adattárházunk több, akár kisebb adatforrásból származó adat összességéből nagy mennyiségű adatot képes elemzésre szolgáltatni. Kis mennyiségű adat esetén előfordulhat, hogy téves következtetéseket vonunk le az algoritmus eredményei alapján, mivel az adatbányászati algoritmusok alapvetően nagyméretű adathalmaz esetén képesek nagy biztonságú információt szolgáltatni.

A folyamat következő lépése az adatforrások kiválasztása. Az adatforrások esetében belső és külső forrásokról beszélhetünk. Belső forrás lehet a könyvtár integrált könyvtári rendszere, webservereiről érkező adatok, külső forrás pedig más könyvtárak adatbázisai, statisztikák, demográfiai adatok, kutatási jelentések. Ezek után következik az adatok előfeldolgozása. Ez két részből áll, az adattisztításból, és az adatintegrációból. A folyamat egyik legfontosabb lépése az adatok megtisztítása. Ugyanis sokszor előfordul, hogy az egyes adatbázis rekordok üres mezőket, elgévelt információkat tartalmaznak, ezeket ki kell szűrni, hogy később megfelelő eredményeket kaphassunk. Elég ha csak arra gondolunk, hogy az olvasó lakcímének megadásakor Budapest-et szeretnénk tárolni, hányféle adat lehet az adatbázisban: budapest, bp. Bp., Budapes, Budpest, Budapest II. ker., Budapest II. Kerület, Bp. 2. kerület stb. Már ebből az egy példából is láthatjuk

²² Somfai Zoltán: Az adatbányászat – az információbróker elemző eszköze In: Mikulás Gábor (szerk.): Információból üzleti érték: az információbróker környezete és munkája. Budapest, Magyar Információbrókerek Egyesülete, 2006.

mennyire nehéz lenne ez alapján, akár csak a Budapesten élő olvasók olvasási szokásaira választ kapni. Az adatok tisztítása után következik az adatok adattárházba történő implementációja. Ehhez a több különböző adatforrásból származó adatokat integrálnunk kell, hogy az adattárház struktúrájába illeszthessük a más-más, feltehetően eltérő koncepciójú relációs adatbázisokból, felmérésekből stb. érkező adatokat. Ennél a pontnál el kell döntenünk, hogy valós idejű rendszerben gondolkodunk, vagy már lezárt adatokat importálunk bizonyos időközönként (negyedévente, havonta) az adattárházunkba. A valós idejű rendszerek hirtelen megjelenő problémákat is képesek kimutatni, azonban megvan az a hátrányuk, hogy nehézkes valós időben a különböző forrásokból érkező adatok átkonvertálása, adattárházba illesztése, elemzése.

A másik probléma, hogy az általunk használt adatokat, mivel az olvasók személyiségi jogait védenünk, anonimizálnunk kell, valamint az összes olyan adatot, amelyből az olvasó egyénileg azonosítható lehet, törölnünk kell, és egyéni azonosítót kell generálnunk, amely alapján mi el tudjuk ugyan dönteni, hogy az adatok mely egyénhez tartoznak, de az azonosítás már nem lehetséges.²³ Ez a gyakorlatban annyit jelent, hogy a demográfiai változókat a számítógép által generált azonosítókkal kombináljuk, az IP címeket kódokkal helyettesítjük, demográfiai csoportokat hozunk létre, így egy nem visszakövethető, anonim felhasználói azonosító jön létre. Az adatok integrációja során ezt is meg kell tennünk, ez is bonyolítja a valós idejű rendszer megvalósításának folyamatát.

A folyamat harmadik lépése a megfelelő elemző eszköz kiválasztása. Ez magától a problémától is függ. A menedzsmenttel kapcsolatos információk kinyerésére a nagyrészt már az integrált könyvtári rendszerek részét képező menedzsment információs rendszerek a megfelelőek. Az elemző eszközök amennyiben adatbányászati algoritmust szeretnénk futtatni, lehetnek adatbányászati szoftverek, úgy mint a SAS²⁴, SPSS²⁵, WEKA²⁶, WORLDCAT. Ha ezeket az eszközöket használjuk, akkor definiálnunk kell, milyen adatbányászati algoritmust szeretnénk használni. Erről részletesebben a 4.3 fejezetben fogok írni. Amennyiben feladatunk a bibliomining-on belül valamilyen

²³ Nicholson, Scott: The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. In: Information Technology and Libraries, 2003 (22.évf.) 4. sz 146-151.p

²⁴ http://www.sas.com/en_us/software/sas9.html (Letöltve: 2014.01.09.)

²⁵ <http://www.spss.hu/> (Letöltve: 2014.01.09)

²⁶ <http://www.cs.waikato.ac.nz/ml/weka/> (Letöltve: 2014.01.09)

bibliometriai elemzés elvégzése, akkor vagy valamilyen bibliometriai elemző szoftver használatát kell igénybe vennünk, vagy saját szoftvert kell készítenünk. Ilyen bibliometriai elemzést biztosító szoftver a svéd Umeå város egyetemének könyvtára által készített szabadon hozzáférhető szoftver, a Bibexcel²⁷: A szoftver segítségével bibliometriai elemzéseket végezhetünk, szerzők közötti hivatkozási rendszereket térképezhetünk fel. A Bloomington-i egyetem fejlesztése, a Network Workbench²⁸: Software és a Pajek²⁹ elnevezésű szoftver a szerzők kapcsolati hálózatát jeleníti meg vizuális elemző eszközök segítségével. A Publish or Perish³⁰:hivatkozáselemző eszköz (University of Melbourne) a Google Scholar adatbázisát, a SAINT³¹ pedig a Web of Science adatait használja fel az elemzéshez. Bár létezik megannyi szoftver, saját eszköz készítése azért előnyös, mert bár több munkával jár a bevezetése, azonban igényeinkhez szabható elemzéseket hajthatunk végre a segítségével.

A bibliomining utolsó lépése az adatok elemzése, információk kinyerése, majd a kapott információk beépítése a könyvtár működésébe. Fontos, hogy a munka megkezdése előtt a kiválasztott algoritmust próbaadatokkal tesztelni kell, majd miután már a tesztelt adathalmazon megfelelően működnek az algoritmusaink, csak akkor használhatjuk az éles, valós adatokon.³² Miután a kapott információkat használatba vettük, érdemes bizonyos időközönként újabb elemzést elvégezni, hogy a kapott válasz segítette-e a könyvtár működését. Emiatt a bibliomining soha véget nem érő, rekurzív, kisebb módosításokkal megújuló folyamat.

A Scott Nicholson által megadott definíciót én a következőkkel egészíteném ki:
az adatbányászat alapvetően egy műszaki, informatikai feladat, és mint a legtöbb műszaki feladat esetében, létezik szabvány, amely leírja mit hogyan kell csinálnunk.
Mivel a bibliomining szerves része az adatbányászat, ezért fontos ezt a szabványt

²⁷ <http://www8.umu.se/inforsk/Bibexcel/> (Letöltve: 2014.01.07)

²⁸ <http://nwb.cns.iu.edu/index.html> (Letöltve: 2014.01.09)

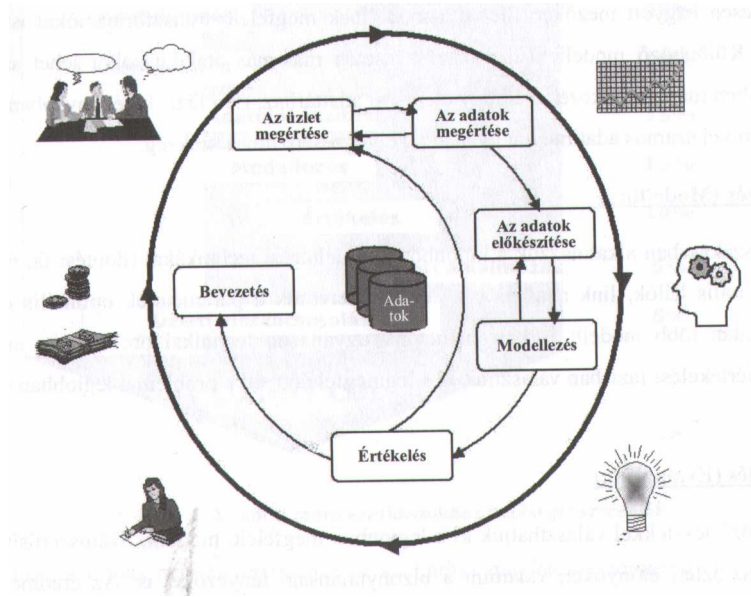
²⁹ <http://pajek.imfm.si/doku.php> (Letöltve: 2014.01.09)

³⁰ <http://www.harzing.com/resources.htm#pop.htm> (Letöltve: 2014.01.09)

³¹ <http://www.rathenau.nl/en/themes/theme/project/bibliometric-software-tools/saint.html> (Letöltve: 2014.01.07)

³² Nicholson, Scott: The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services In: Information Processing & Management, 2006. (42.évf.) 3. sz. 785-804. p.

használunk. A jelenleg hatályban lévő szabvány a CRISP (Cross Industry Standard Process for Data Mining), amely átfogóan leírja az adatbányászat folyamatát³³.



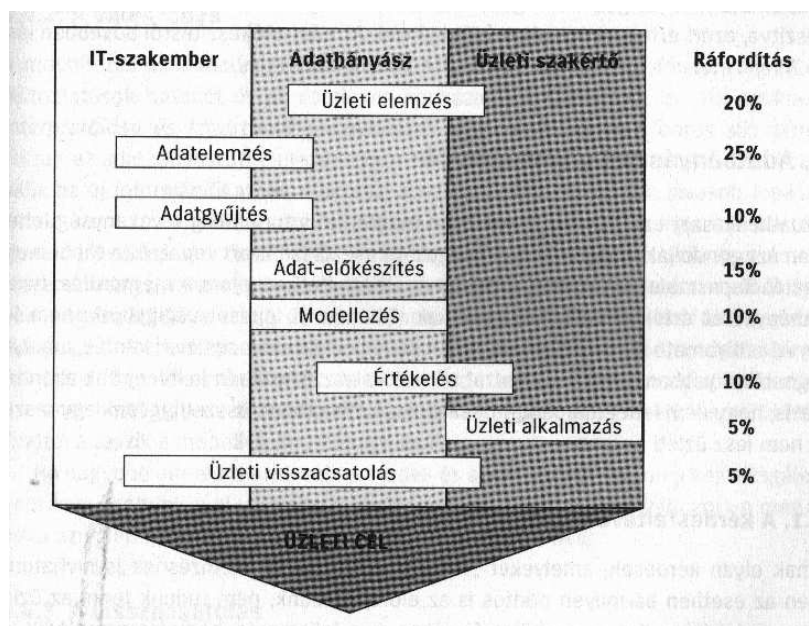
3. ábra
A CRISP-DM módszertan³⁴

A szabvány meghatározza azt is, hogy egy adatbányászati projekt időben hogyan alakuljon, az egyes folyamatok milyen időigénnyel rendelkeznek. A folyamat elején lévő üzleti elemzés: 20%, az adatelemzés: 25%, tehát maga a problémakörnyezet elemzése teszi ki az egész folyamat majdnem felét. Ezek után következik a környezet létrehozása és elemzése (adatgyűjtés: 10%, adat-előkészítés: 15%, modellezés: 10%. A folyamat végén pedig következik az értékelés: 10%-kal, az üzleti alkalmazás: 5%-kal, az üzleti visszacsatolás pedig 5%-kal jelenik meg.³⁵

³³ Fajszai Bulcsó- Cser László, Fehér Tamás: Üzleti haszon az adatok mélyén: Az adatbányászat mindennapjai. Budapest, Alinea IQSYS, 2010. 95.p.

³⁴U.o.

³⁵ Fajszai Bulcsó, Cser László, Fehér Tamás: Üzleti haszon az adatok mélyén: Az adatbányászat mindennapjai. Budapest, Alinea IQSYS, 2010. 97.p



4.ábra
Az adott munkaszakaszok időigényei, szereplői³⁶

4.1.2.3. Alkalmazási területek

Az adatbányászat a könyvtári munka minden területén megjelenik. Scott Nicholson 2003-ben megjelent cikkében határozza meg³⁷, hogy mely könyvtári területeken használható az adatbányászat, magyarul Mátyás Melinda foglalta össze ezeket az ismereteket szakdolgozatában³⁸. A könyvtárakban az adatbányászat három fő területen használható: az állománygyarapítás-gyűjteményszervezés, a menedzsment és döntéstámogatás, illetve a tájékoztatás.³⁹

A könyvtár gyűjteményéről optimálisan minden adat megtalálható az integrált könyvtári szoftverekben, így az összes, a gyűjteményhez kapcsolódó folyamat, tehát az állománygyarapítás, állományapaszítás, karbantartás, állományelemzés, elemezhetővé válik, ha valamilyen statisztikai, adatbányászati algoritmussal az adatok mélyére

³⁶ Fajszí Bulcsó, Cser László, Fehér Tamás: Üzleti haszon az adatok mélyén: Az adatbányászat mindennapjai. Budapest, Alinea IQSYS, 2010. 97.p.

³⁷ Nicholson, Scott: The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. In: Information Technology and Libraries, 2003 (22.évf.) 4. sz. 146-151.p.

³⁸ Mátyás Melinda: Adatbányászat a könyvtárakban. [szakdolgozat] Budapest, Eötvös Lóránd Tudományegyetem, 2009. <http://mek.oszk.hu/08000/08059/08059.pdf> (Letöltve: 2013.12.15)

³⁹ Nicholson, Scott: Approaching Librarianship from the Data: Using Bibliomining for Evidence-Based Librarianship In: Library Hi-Tech, 2006. (24. évf.) 3. sz. 369-375.p.

tekintünk⁴⁰. Ezekből a bibliográfiai adatokból információkat kinyerve optimalizálhatjuk a könyvtár gyűjteményét, valamint nem elhanyagolható szempontként költséghatékonyabban tudják a szerzeményező munkatársak végezni a munkájukat. Alapszintű statisztikák segítségével fényt deríthetünk arra, hogy gyűjteményünk egyes részeit milyen mértékben használják, melyek azok a dokumentumok, amelyeket nem kölcsönöznek, melyből lenne szükség több példányra a kölcsönzési, illetve előjegyzési statisztikák alapján. Megvizsgálhatjuk a gyűjteményünket olyan szempont szerint is, hogy a meglévő állományban lévő dokumentumok mennyire fedik le a könyvtárunk profilját, a meglévő állomány mérete mennyire illeszkedik a könyvtár szakterületéhez. Ennek megállapítását segítheti, ha több könyvtár közösen épít adattárházat, és így összehasonlíthatjuk saját gyűjteményünket a hasonló profilú könyvtárak gyűjteményével, és ezek alapján kiegészíthetjük az állományunkat.⁴¹ Ezen összehasonlító elemzések elősegíthetik a könyvtárközi kölcsönzések optimális működését, például melyik könyvtár állományából kérjük az adott dokumentumot, hol van a legkisebb szükség az adott műre. Emellett vizsgálhatjuk a művek műfaj szerinti megoszlását, vagy hogy mely könyvet mivel kölcsönöznek együtt.

Az állomány elemzése után dönthetünk arról, mely könyvet lenne érdemes selejtezni, mire nincs szüksége már a könyvtárnak, illetve kiválasztjuk, hogy milyen dokumentumokat kellene beszerezni. Érdemes azonban magát a szerzeményezés folyamatát is megvizsgálnunk, hiszen itt is érdekes információkra bukkanhatunk. A folyamatot végigkövetve megvizsgálhatjuk, mikor, kitől milyen kedvezménnyel rendeltünk, honnan érdemes megrendelnünk egy adott művet, mennyi ideig tart, míg a rendelés leadásától a polcra kerül a dokumentum, és olyan logisztikai problémákra is fényt deríthetünk, hogy miért késik egy adott rendelés.⁴²

⁴⁰ Mátyás Melinda: Adatbányászat a könyvtárakban. [szakdolgozat] Budapest, Eötvös Lóránd Tudományegyetem, 2009. 18.p. <http://mek.oszk.hu/08000/08059/08059.pdf> (Letöltve: 2013.12.15)

⁴¹ Nicholson, Scott - Stanton, Jeffrey: Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.1835&rep=rep1&type=pdf> (Letöltve: 2014.01.14)

⁴² Mátyás Melinda: Adatbányászat a könyvtárakban [szakdolgozat]. Budapest, Eötvös Lóránd Tudományegyetem, 2009. <http://mek.oszk.hu/08000/08059/08059.pdf> (Letöltve: 2013.12.15)

A következő nagy csoport a **menedzsment, döntéstámogatás** területe. Itt az egyik fontos elem a személyi döntésekkel kapcsolatos feladatok segítése. A munkaidő beosztás elkészítése, például milyen időintervallumban van szükség több olvasószolgálatos kollégára, felsőoktatási könyvtár esetében a tanév mely időszakában kéri a tájékoztató munkatárs segítségét, milyen időszakban legyen hosszabb ideig nyitva a könyvtár. Az adatbányászat emellett a költségvetés összeállításában is segítséget nyújthat, a szerzeményezés folyamatát is segítheti, ahogy már említettük, illetve a felhasználói szokások megértésében is hasznos eszköz lehet. A felhasználói igények ismerete elengedhetetlen ahhoz, hogy megfelelő szolgáltatást nyújthassunk a számukra. Adatbányászati algoritmusok segítségével felhasználói csoportokat hozhatunk létre, különböző korcsoportok kölcsönzési szokásait vizsgálhatjuk meg, és személyre szabott szolgáltatásokat nyújthatunk olvasóink számára. A felhasználói szokások megértéséhez a kölcsönzési adatok mellett használhatjuk az olvasók online katalógusban végzett kereséseit, összefüggéseket vizsgálhatunk, hogy a keresésekből született-e kölcsönzés, megtalálták-e a keresett dokumentumot. Érdekes lehet annak vizsgálata milyen keresőkifejezéseket használtak, illetve ezek mennyire korreláltak az általunk használt tárgyszavakkal. Ezen a területen nehézséget az jelenthet, hogy az olvasók tevékenysége nem csak az online térben zajlik, hanem offline, a könyvtár épületében. Megoldandó problémát képez, hogyan rögzítsük a helyben használathoz kapcsolódó adatokat, a helyben feltett kérdéseket, és az ez alapján keletkezett kölcsönzési információkat. A felhasználói szokások elemzése közben folyamatosan figyelniünk kell arra, hogy az olvasók személyiségi jogait tiszteletben kell tartanunk, és csak anonim módon használhatjuk fel a hozzájuk kapcsolódó adatokat.

A könyvtári munkafolyamatok közül talán a tájékoztatás az, ahol a legnehezebben hasznosítható az adatbányászat. A tájékoztatás egy rendkívül egyedi feladat minden esetben, nagyon ritka hogy két olvasó teljesen ugyanolyan kérdéssel keresse meg a tájékoztató munkatársat. Azonban már az is segítséget jelenthet a könyvtári dolgozó számára, ha a korábbi kérdéseket és válaszokat, ajánlott forrásokat egy adatbázisban tárolják, mivel ekkor különböző algoritmusok segítségével megtalálhatjuk a konkrét kérdéshez legközelebb álló korábbi kérdéseket, és ez segítheti a munkánkat.

Nincs különbség aközött, hogy hagyományos könyvtárban, digitális vagy elektronikus könyvtárban hajtunk végre az elemzéseket, mivel ugyanúgy adatbázisokban tárolják az adatok, és ugyanolyan módon készíthetünk jelentéseket, statisztikákat, csak abban van különbség, hogy elektronikus könyvtár esetében nem a kölcsönzéseket vizsgáljuk, hanem a letöltések eloszlását, a különböző dokumentumtípus szerinti kereséseket. Scott Nicholson Bibliomining for Automated Collection Development⁴³ című cikkében remek példát mutat, hogyan építhető be az adatbányászat a digitális könyvtárak gyűjteményszervezésébe.

4.1.2.4. A jelenlegi helyzet

A könyvtári adatbányászat már az 1990-es években megjelent, több könyvtárban kísérleti alkalmazásokat hoztak létre. A legtöbb ilyen rendszer az Egyesült Államokban született. Sok, elsősorban kutatóegyetemi könyvtár készített már mintaalkalmazásokat, amelyek beváltak, hasznos információkkal segítik a könyvtárak működését. A szakirodalomból látható, hogy az utóbbi 10-15 évben egyre nagyobb hangsúlyt kap a könyvtárak életében, elsősorban más technológiákkal, webometria, bibliometria, informetria közösen alkalmazva. Hazánkban Mátyás Melinda és Jezerniczky Istvánné készített szakdolgozatot ebben a témában. Mátyás Melinda elsősorban elméleti hátterét mutatta be a kérdésnek, Jezerniczky Istvánné pedig (Az adatbányászat egy könyvtári alkalmazása⁴⁴) egy lehetséges könyvtári adatbányászati alkalmazás TextLib környezetben való bevezetését vázolta. Ebből a dolgozatról kiderül, hogy megfelelően alkalmazható hazánkban is a módszer, esetlegesen a hozzáállás szabhat gátat elterjedésének. Tóth Erzsébet 2002-es cikkében⁴⁵ azonban felhívja a figyelmet a bibliomining bevezetésében rejlő lehetőségek melletti veszélyekre is, amelyek következők: technikai akadályok, kevés alkalmazási példa, illetve a szabványok hiánya. Ezek alapján látható, hogy Magyarországon is felmerült már az adatbányászat szükségessége, azonban a köztudatba nem épült be, és bár világszerte több alkalmazás is született, egységes használatról még

⁴³ Nicholson, Scott: Bibliomining for Automated Collection Development In: Journal of the American Society for Information Science and Technology, 2003. (54. évf.) 12. sz. 1081-1090.p.

⁴⁴ Jezerniczky Istvánné: Az adatbányászat egy könyvtári alkalmazása [szakdolgozat] Dunaújvárosi Főiskola. 2009

⁴⁵ Tóth Erzsébet: Mire jó az adatbányászat? In: Tudományos és Műszaki Tájékoztatás, 2002. (48. évf.), 3. sz. 527-532.p.

nem beszélhetünk. Ahhoz, hogy ez megtörténhessen egyaránt szükség van a technológiai fejlesztésekre, a szakemberek képzésére és működő alkalmazások létrehozására.

4.1.3. Integrált könyvtári rendszerek

Az integrált könyvtári rendszerek története az 1960-as években az Egyesült Államokban vette kezdetét. ⁴⁶Mivel a számítástechnikai eszközök fejlődése is az Egyesült Államokból indult, természetes, hogy a különböző szakterületek céljainak megfelelő szoftverek is itt készültek el először. A '60-as évek végén létrehozott könyvtárakban használt programok célja elsősorban a kölcsönzések és az állomány nyilvántartása volt. ⁴⁷Ebben az időszakban még ezek a programok a nagy számítóközpontokban található gépeken futottak, később az eszközök fejlődésével terjedhetett el külön szerverek vásárlása a könyvtárak számára, és ezen szerverek alkalmassá tétele, a könyvtárak számára készített szoftverek futtatására. Az elsődleges feladat a bibliográfiai rekordok szoftveres környezetben történő megjelenítése volt. Ahhoz, hogy a számítógép kezelni tudja a bibliográfiai adatokat, szabványosan létrehozott kódokat kellett az adatokhoz hozzárendelni, emiatt jött létre a MARC (MAchine Readable Cataloging; géppel olvasható katalogizálás) a könyvtárakban és a nemzeti bibliográfiai központokban általában alkalmazott adatcsere-formátum. A 1960-as évek elején a Kongresszusi Könyvtár fejlesztette ki az első MARC formátumot. A gépi katalogizálás az egyes MARC rekordok létrehozásakor a katalogizáló munkatárs a bibliográfiai leírás létrehozása során meghatározza, hogy az egyes adatelemeket mely csereformátumbeli kódhoz kell rendelnie, és ezt a számítógépes rendszerben rögzítve hozza létre a bibliográfiai rekordokat, majd a szoftver elvégzi a megfelelő adatfeldolgozást, indexkészítést, a bibliográfiai tételek megjelenítését.

Az 1970-es évek végére már Európában is megjelentek az integrált könyvtári rendszerek előfutárai. Ekkora már elterjedtek a miniszámítógépek, és előtérbe került a modularitás gondolata is, tehát a különböző feladatok más-más modulban kaptak helyet,

⁴⁶ Kerecsendi András: Integrált könyvtári rendszerek. 2011.

http://www.tankonyvtar.hu/hu/tartalom/tamop425/0005_28_int_konyvtari_rendszerek_pdf/adatok.html
(Letöltve: 2014.01.05.)

⁴⁷Bakonyi Géza - Kokas Károly: Könyvtári integrált rendszerek és hazai alkalmazásuk
<http://mek.oszk.hu/03000/03088/html/> (Letöltve: 2014.01.05.)

egy szoftveren belül. Ebben az időszakban az egyes szoftverek még nagyon erősen meghatározott hardver és operációs rendszer függőek voltak, azonban már kialakultak a ma is legfontosabb modulok: katalogizálás, kölcsönzés, online katalógus. Az egyes modulok külön kezelése lehetővé tette, hogy különböző jogosultsági szinteket állapítsanak meg, egy dolgozó csak a saját munkaköréhez tartozó modulhoz kapott hozzáférést, így az adatbiztonság, ellenőrizhetőség jelentősen megnőtt. Ebben az időben kezdték el az integrált könyvtári rendszerek elnevezés használatát. Az integráltság abból adódik, hogy míg bár az egyes modulok külön is elérhetőek, használhatóak, azonban a háttérben az adatok egy, elsődleges relációs adatbázisokban tárolódnak, így az adatok egymásra épülhetnek, lehetővé válik az egész könyvtári rendszer együttes kezelése.

Az 1980-as években újabb változáson mentek keresztül az integrált könyvtári rendszerek, az IBM által piacra dobott személyi számítógépek könyvtári elterjedésével a szoftvergyártóknak alkalmazkodniuk kellett az új helyzethez, az eddigi szerver – terminál kommunikációt felváltották a grafikus felületet használó integrált könyvtári rendszerek, illetve a modulok kibővültek, megjelentek a szerzeményezési, folyóiratkezelési modulok is.

A 90'-es évekre a UNIX rendszerek elterjedésével megoldódtak az operációs rendszerek okozta kötöttségek, illetve az internet elérte a könyvtárakat is, és bevezetése az integrált könyvtári rendszer használatának elterjedését is előrevetítette. Az integrált könyvtári rendszerek már a kezdetektől rendelkeztek bizonyos statisztikai adatok lekérdezését segítő funkciókkal, például kölcsönzés nyilvántartás készítése, azonban csak a 90'-es végén kezdtek el a rendszerekbe különálló statisztikai modulokat építeni. Mivel piaci alapon szerveződött az integrált könyvtári rendszerek fejlesztése, rengeteg vállalat kezdett el könyvtári rendszereket gyártani Amerikán kívül is, így nem beszélhetünk egységes szoftverkörnyezetről, az országok nagy részében, így nálunk is több szoftver használata terjedt el, amely megnehezíti a könyvtárak közötti információcserét, kommunikációt.⁴⁸

⁴⁸ Kerecsendi András: Integrált könyvtári rendszerek. 2011.

http://www.tankonyvtar.hu/hu/tartalom/tamop425/0005_28_int_konyvtari_rendszerek_pdf/adatok.html

(Letöltve: 2014.01.05.)

4.2. Integrált könyvtári rendszerek statisztikai moduljának elemzése

Az integrált könyvtári rendszerek statisztikai moduljainak elemzése kulcsfontosságú abból a szempontból, hogy ezek képezik az alapját a későbbi adatbányászati elemzéseknek. Mivel a bibliomining statisztikai eszközökkel is dolgozik, tisztában kell lennünk azzal, hogy az általunk használt rendszerek milyen információkat szolgáltatnak számunkra.

Hazánkban több mint 15 integrált könyvtári rendszer van használatban. Nincsen egységes szoftverhasználat sem régiók, sem könyvtártípusok esetén. Erre az elemzésre azért is van szükség, hogy megállapítsuk mely szoftver lenne a legjobban használható adatbányászati szempontból, illetve melyiket lenne legegyszerűbb kibővíteni úgy, hogy adatbányászati feladatokat is ellásson.

A következőkben a Magyarországon használt, elérhető dokumentációjú integrált könyvtári rendszereket elemeztem a következő szempontok szerint:

- Rendelkezik-e statisztikai modullal a rendszer?
- A statisztikák mennyire statikusak?
- Beépített lekérdezések futtatására van csak lehetőség, vagy saját lekérdezések, jelentések elkészítésére is.
- Lehetőség van-e az adott rendszerben időben zajló folyamatok, trendek felismerésére, összehasonlítására, vagy csak egy időintervallumra vonatkozó információkat kapunk?

Mivel a bibliomining három fő könyvtári területen alkalmazható, ezért az integrált könyvtári rendszereket is ez alapján vizsgáltam, hogy a statisztikák mennyire segítenek ezen a három területen felmerülő kérdések megválaszolására, tehát:

- Milyen statisztikák készítésére van lehetőség a gyűjteményszervezés, tájékoztatás, és a menedzsment támogatás területén?

4.2.1. A Corvina

A Corvina Magyarország egyik legelterjedtebb integrált könyvtári rendszere. Külön statisztikai modullal rendelkezik (Perl nyelven íródott), amely vezetői információk néven érhető el. Ebből is láthatjuk, hogy ezek általános statisztikák, nem célspecifikusan az egyes részterületen dolgozók számára készültek, hanem az intézmény vezetője számára adnak információkat a könyvtár működéséről. A lekérdezések háttérében lefutó SQL lekérdezések módosíthatóak ugyan, szűrőfeltételekkel elláthatóak, azonban ehhez az SQL nyelv, és a mögöttes adatbázisszerkezet pontos ismerete szükséges. A gyűjteménnyel kapcsolatban lekérdezhető egy adott intervallumban elindított kölcsönzési tranzakciók száma, illetve hogy hány dokumentumot kölcsönöztek. Emellett a beiratkozások száma, valamint az egyes felhasználók kölcsönzései, kifizetései tekinthetők meg. A bibliomíng *harmadik* célterületének, a tájékoztatásnak a támogatása nem jelenik meg ebben az integrált könyvtári rendszerben. A lekérdezések egy adott időintervallumra vonatkoznak, így trendek vizsgálatára nincs lehetőség.

4.2.2. A Szirén

A Szirén teljesen magyar fejlesztésű integrált könyvtári rendszer, 1985-ben kezdődött el a fejlesztése. Jelenleg hét országban használják, összesen 1600 könyvtár integrált könyvtári rendszere. 2003-ban építették be a rendszerbe a közös katalógizálást, amelynek lényege, hogy ha bármelyik Szirén rendszert használó könyvtárban elkészítik egy dokumentum bibliográfiai leírását, az feltöltődik az összes Szirén által használt könyvtár központi szerverére. A rendszerben több külföldi könyvtár adatai is megtalálhatóak (mint például a Library of Congress vagy az Oxford University könyvtára). A külföldi könyvtárak elérése nem csak az idegen nyelvű dokumentumok leírását segíti, hanem az adatbányászat során összehasonlító szerepet is betölthet.

A szoftver nem rendelkezik különálló statisztikai modullal, az egyes modulokban kiegészítő funkcióként jelenik meg az adott modulhoz tartozó statisztikák készítésének lehetősége. A statisztikák előre definiáltak, nem hozhatunk létre új lekérdezéseket, azonban paraméterezhetők, kezdő és végdátum megadására van lehetőség, az egyes lekérdezések különböző csoportok mentén szűrhetők is.

A gyűjteményszervezéshez kapcsolódó statisztikák: az adott dokumentum története (keresés leltári szám és cím szerint), előjegyzések listája, iskolai könyvtár esetén a dokumentumok kölcsönzési statisztikája osztályok szerinti bontásban. Lehetőség van az egyes ETO főosztályok szerinti kölcsönzések, előjegyzések megtekintésére is. Információkat kaphatunk a felhasználókról, kölcsönzéseik, előjegyzéseik, tartozásaik is megtekinthetők, valamint az olvasókat osztály szerinti csoportosításban is lekérdezhetjük. A költségvetés összeállításánál hasznos lehet a befizetéseket és a pénzforgalmakat mutató statisztika, mely csoportosítható a könyvtár egyenlege, egyes olvasók egyenlege szerint is, valamint szűrések készíthetők a rendelési bizonylatok alapján is, beszerzési forrásonként. Emellett külön lekérdezés áll rendelkezésre a közművelődési könyvtári statisztikai adatok megadásához is.

A fentiek alapján látható, hogy a lekérdezések két területen, a *gyűjteményszervezés*, és a *menedzsment, döntéstámogatás* területén jelennek meg, valamint egyértelmű, hogy trendek vizsgálatára nincs lehetőség, csak egy időintervallum adatainak lekérdezésére. Más külső programot használva és ezeket az adatokat kigyűjtve lehetséges csak időben zajló folyamatok vizsgálata.

4.2.3. Az Aleph

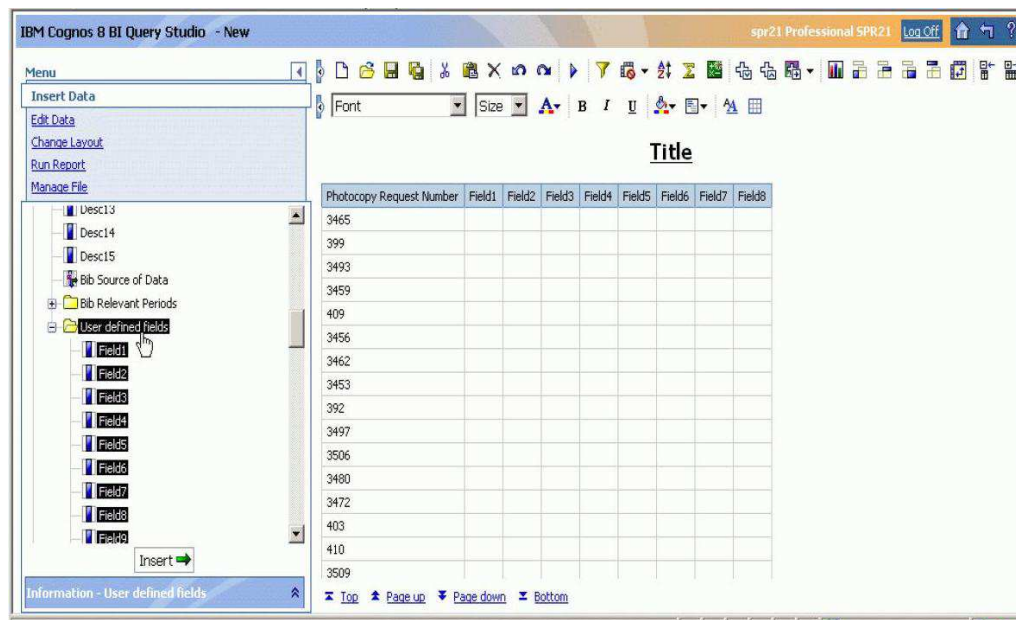
Az Ex-Lh Kft. által forgalmazott ALEPH⁴⁹ (Automated Library Expandable Program) Magyarországon nagyon elterjedt könyvtári szoftver, több, mint 60 könyvtár használja. Magát a szoftvert a jeruzsálemi Hebrew egyetem programozó és rendszerszervező csoportja kezdte el fejleszteni 1978-ban, azóta világszerte elterjedt, jelenleg az ALEPH 500-as verzió található a piacon.

Az Alephben készíthetünk alapstatisztikákat az adatbázis adatai alapján, úgy, mint beiratkozások, kölcsönzések száma, hány példánnyal bővült a könyvtár katalógusa egy adott évben, valamint felhasználókhöz kapcsolódó információk lekérése. Ezekben a statisztikákban az olvasók csoportosítására is van lehetőség életkor, illetve lakhely alapján. Az Aleph beszerzési moduljának egyik leghasznosabb tulajdonsága, hogy elutasítja a megrendeléseket, ha az adott dokumentumok ára már nem finanszírozható a

⁴⁹http://www.tankonyvtar.hu/hu/tartalom/tamop425/0005_28_int_konyvtari_rendszerek_scorm_10/1038_az_alephrendszer_1talnos_bemutatsa.html (Letöltve: 2014.01.12.)

költségvetési keretből. Az olvasók tájékoztatását segítő funkció a rendszerben a felhasználóknak nyújtott ama lehetőség, hogy saját maguk beállíthatják milyen témák érdeklik őket, és erre témakövetési szolgáltatást igényelhetnek.

Az alapstatisztikák mellett statisztikai modulként működik a rendszerben az ARC (Aleph Reporting Center), amely gyakorlatilag egy jelentésgeneráló felület, ahol az adatbázis ismerete és egyéb SQL ismeretek nélkül tud a könyvtáros statisztikákat készíteni. A modulban több, mint 100 beépített lekérdezés található, amely mélyebb információkat szolgáltatnak a könyvtár működéséről, az egyes munkafolyamatokról, ezen lekérdezések az IKR adatbázisából nyerik forrásadataikat. Ezeket a lekérdezéseket teljesen egyénre szabhatjuk, nem statikusak, egy úgynevezett *Jelentésszerkesztő* felületen adhatunk újabb mezőket a lekérdezésekhez, szűkíthetjük a lekérdezendő kört, az adathalmazt, adott felhasználói csoportokat állíthatunk be, illetve több csoportosítási változó mentén elemezhetjük az adatokat.

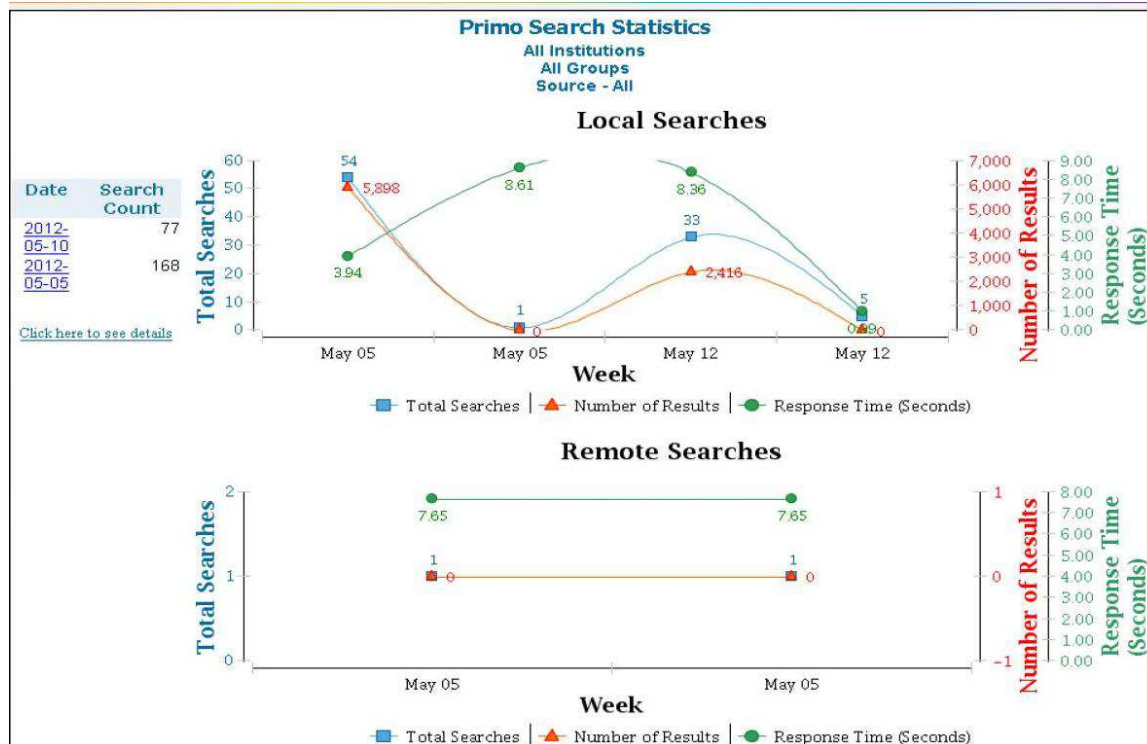


5.ábra
Aleph jelentésgeneráló felület

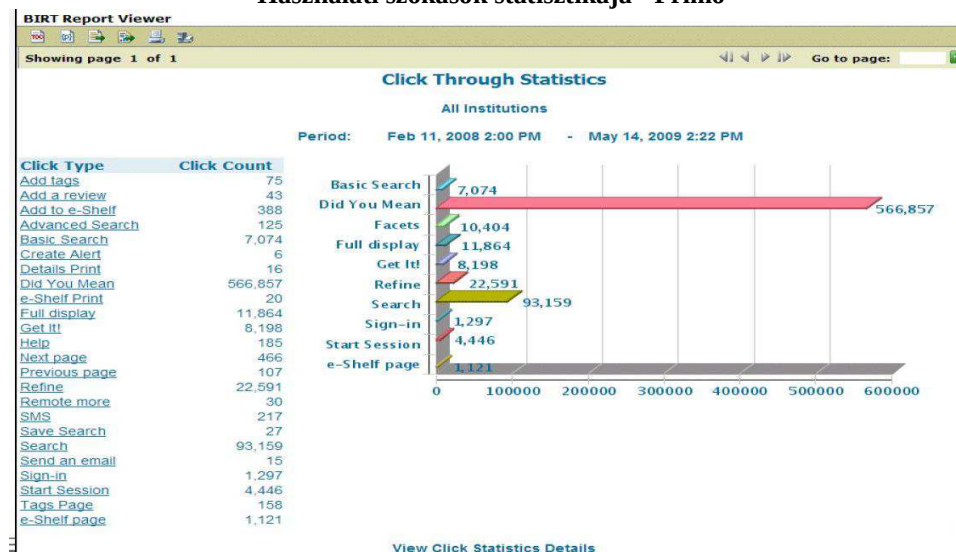
A döntéshozatalt segítő egyszerű, felhasználóbarát felületen grafikonok állnak a dolgozók rendelkezésére, amelyeket szintén módosíthatjuk, egyénre szabhatjuk. A több, mint száz lekérdezés a könyvtári munkafolyamatok minden területét lefedi. Az állomány elemzéséhez nagy segítséget nyújthat az adott évben legtöbb kölcsönzött

dokumentumok lekérdezése, az egyes dokumentumok forgási sebessége, akár alkönyvtárak, felhasználók, dokumentumtípusok, nyelv szerinti bontásban. Emellett diagramos kimutatás készíthető a kölcsönzések, előjegyzések eloszlásáról adott időszakban. Az állomány karbantartását segítheti az a funkció, hogy maga az Aleph visszakeresi az állományból már törölt könyveket vonalkód, cím, leltári szám vagy szerző alapján, amennyiben igény merülne fel a beszerzésére, hasznos információ lehet, hogy korábban milyen használati mutatókkal rendelkezett, és ez alapján dönthetünk róla, hogy visszakerüljön-e az állományba az adott dokumentum. Az ARC segítségével elemezhetjük az online katalógusban zajlott kereséseket, illetve egyéb felhasználók által indított műveleteket, úgy mint előjegyzés, témakövetés, fénymásolás stb. Az adatokat nem csak az egyes olvasók alapján csoportosítva láthatjuk, hanem a kérést elindító gép IP címe szerinti bontásban is, így például ha a külső IP címről érkező kéréseket vizsgáljuk, akkor az olvasók távoli könyvtárhasználatának szokásairól kaphatunk információkat. A jelentések készítésekor beállíthatunk több időintervallumban való keresést is, ezáltal képesek vagyunk trendek vizsgálatára is.

Az Aleph integrált könyvtári rendszert használó könyvtárak bővíthetik palettájukat, más az Ex Libris által kifejlesztett szoftverrel, úgy mint a Metalib, Primo, SFX, Rosetta. Az SFX linkszervere, valamint a MetaLib segíti a felhasználói kereséseket, hozzáférést biztosít több bibliográfiai és szakirodalmi adatbázishoz, elektronikus dokumentum tárhelyhez, webes információforrásokhoz, ezáltal a felhasználói élmény növelésében játszik nagy szerepet. A MetaLib háttéradatbázisából a könyvtár statisztikai adatokat kaphat a felhasználói keresésekről akár órás bontásban is. A Primo szoftvere segítségével a könyvtárak részletes használati, keresési információkkal gazdagodhatnak.



6. ábra
Használati szokások statisztikája - Primo



7. ábra
Keresési szokások - Primo

Szintén az Ex Libris fejlesztéseként hozták létre 2013-ban az Alma⁵⁰ elnevezésű integrált könyvtári szoftvert, amely következő generációs könyvtári rendszerként hirdeti magát, mivel a legújabb technológiákat beépítve próbál megfelelni a 21. században megjelenő kihívásoknak. Elsősorban a döntéselőkészítés és a munkafolyamat orientáció fejlesztését célozta meg, így a könyvtárakban a menedzsment, döntéstámogatás területén nyújthatja a legtöbbet. Interaktív felületen mutatja be az éppen aktuális, valós idejű adatokon alapuló statisztikákat, diagramokat, mint például a legkevesbé használt folyóiratok tízes listája, az elektronikus folyóiratok használata diagramon ábrázolva, valamint az aktuális dokumentumhasználat. Az automatikusan beállított lekérdezések átszerkeszthetők, szűréseket, aggregációkat végezhetünk rajtuk, új mezőket adhatunk hozzájuk, beállíthatjuk milyen diagramon szeretnénk megjeleníteni az eredményeket. Az Alma erőssége a munkafolyamat orientáció fejlesztése, ez a rendelések menedzselésénél is megjelenik, ahol a rendszer képes a beérkező kérésekre automatikus rendelési folyamat elindítására, a leghatékosabb kereskedő kiválasztására, sőt a dokumentumok tényleges megrendelésére is. (Természetesen beállítható, hogy ehhez engedélyt kérjen egy beszerző munkatárstól, aki tisztában van azzal, hogy az adott dokumentumra szüksége van-e a könyvtárnak.) Az egyes rendelések az egész munkafolyamat során automatikusan továbbítódnak ahhoz a munkatárshoz, akinek dolgoznia kell a dokumentumokkal, tehát a beszerzés után beérkező dokumentumokról automatikus értesíti azt, aki katalogizálni fog, a polcra kerülés után pedig értesítést küld a tájékoztató munkatársaknak az újonnan beérkezett dokumentumokról. Ezek a jelentések állandó információt szolgáltatnak számunkra a könyvtár aktuális működéséről, emellett pedig lehetőségünk van a múltbeli adatok elemzésére, több időintervallum összehasonlítására is. Ezek lehetnek számszerű egyszerű statisztikák (hány kölcsönzés zajlott egy adott időszakban), de lehetnek bonyolultabb, több változó alapján generálódó analitikák, melyek egy adott témakörhöz kapcsolódó információkat szolgáltatnak, időintervallum, vagy tagkönyvtár alapján csoportosítva. A statisztikák adatai exportálhatóak Excel, XML, CSV vagy tabulátorral tagolt formátumban, ez a későbbi elemzés során hasznosítható lehet, akár egy adatbányászati szoftver forrásfájljai lehetnek így az adatok, valamint exportálhatunk pdf,

⁵⁰ <http://www.exlibrisgroup.com/category/AlmaOverview> (Letöltve: 2014.01.12.)

ppt, vagy webarchívum formátumban, ezek elsősorban eredményeink prezentálásakor lehetnek hasznosak.

Összességében elmondható, hogy az Ex Libris által fejlesztett szoftverek rendkívül magas képességekkel rendelkeznek már most is, az általuk használt technológiák alkalmasak lehetnek arra, hogy adatbányászati irány alapján továbbfejlesztve segíthessék a könyvtárak működését.

4.2.4. A Huntéka

Az MTA SZTAKI és a Ikron Kft. együttműködésében jött létre a magyar fejlesztésű Huntéka integrált könyvtári rendszer, melyet jelenleg a Monguz Kft. fejleszt tovább. A HunTéka nem rendelkezik külön statisztikai modullal, az egyes modulokban találhatunk lekérdezések készítésére lehetőséget.⁵¹ A kölcsönzés modulban például statisztikákat készíthetünk a kölcsönzések, előjegyzések számáról egy adott időintervallumban. A statisztikákat különböző felhasználói csoportok, például osztályok alapján csoportosíthatjuk. A *Gyarapítások* modulban az aktuális rendelésekhez kapcsolódó információkról, a könyvtár költségvetéséről, a dokumentumok igénylőiről készíthetünk napi statisztikákat. Az egyes jelentések paramétereizhetők, azonban a jelentéskínálat statikus, saját általunk kreált lekérdezések nem építhetők be a rendszer működésébe. Trendek vizsgálatára nincs lehetőség, csak a napi, illetve egy adott intervallumra vonatkozó információk kérhetők le a rendszerből.⁵²

4.2.5. A LinLib

A LinLib egy többnyelvű, ingyenes integrált könyvtári szoftver, melyet a bajai városi könyvtár fejlesztett ki. A program rendelkezik különálló statisztikai modullal, emellett pedig más modulokban is találhatunk statisztikák készítésére lehetőséget. A statisztika menüpont alatt a forgalomra, beiratkozásokra, katalogizálásokra vonatkozó statisztikák érhetőek el. A kölcsönzés modulban találhatóak egy adott olvasó

⁵¹ <http://www.hunteka.hu/index.pl/progism> (Letöltve: 2014.01.12.)

⁵² Király László – Tóth Kornél: HunTéka: egy új integrált rendszer a magyar könyvtári piacon In: Tudományos és Műszaki Tájékoztatás, 2004. (51.évf.) 8. sz.
http://tmt.omikk.bme.hu/show_news.html?id=3677&issue_id=453 (Letöltve: 2014.01.12.)

kölcsönzései, egy dokumentum története, keresés dátum alapján (év, ezen belül hónap szerint).

A gyűjtemény forgalmára vonatkozóan havi és évi lebontású lekérdezés áll rendelkezésünkre, a havi bontásban napokra szűkítve kaphatunk adatokat. Külön megjelennek a 14 éven aluli és felüli kölcsönzők adatai, a kölcsönzött dokumentumok száma, beiratkozott olvasók száma, a gyerek és felnőtt látogatók száma, az internet használat, távhasználat, illetve a referenzs kérdések száma. A beiratkozó olvasókra vonatkozó adatokat hónapokra, évekre bontva szűkíthetjük, és csoportosítani tudjuk az olvasókat életkor, lakhely, nem, kor, foglalkozás alapján. A katalogizálás menüpont évenkénti, és azon belül havi bontásban mutatja az újonnan beszerzett dokumentumokat dokumentumtípusok szerinti (könyv, régi könyv) bontásban.

Látható, hogy a rendszer statikus lekérdezéseket tartalmaz, egyedül a lekérdezés intervalluma állítható be személyre szabottan, emiatt nyilvánvaló, hogy trendek vizsgálatára nincs lehetőség. Pozitívum, hogy a felhasználók adatait többféle csoportosításban is megtekinthetjük, így lehetőséget biztosít a rendszer felhasználói csoportokat célzó szolgáltatások létrehozására.

4.2.6. A Szikla

A Szikla a NETLib Kft. Fejlesztésében létrejött integrált könyvtári szoftver, 1998-ban jelent meg a piacon. A fejlesztő cég elsősorban a városi és szakkönyvtárak számára fejlesztette ki rendszerét. Három Szikla terméket fejlesztenek jelenleg, a SZIKLA 21-et, a Szikla-Minit, illetve a SZIKLA-Basic-et. Az egyes termékek funkcionalitásukban különböznek elsősorban, és emiatt az elérni kívánt célközönségben is.⁵³

Nem rendelkezik különálló statisztikai modullal, de van lehetőségünk statisztikák készítésére, a statisztikák a különböző modulokhoz kapcsolódóan készíthetők el, a kölcsönzési, a személyek, és a statisztikai adatbázis adatainak felhasználásával. A felhasználói szokások megértéséhez nyújtanak segítséget a kölcsönzési űrlapon elérhető lekérdezések. A következő lekérdezések találhatóak itt: az olvasók kölcsönzéseit tekinthetjük meg, illetve azt, hogy korábban milyen kölcsönzési, előjegyzési

⁵³ <http://www.szikla.net/hun/termekek.php> (Letöltve: 2014.01.13)

tranzakciókat, tartozások hajtottak végre. Ezeket az adatokat rendezni lehet, időintervallumra szűrni nem lehet belőlük, keresni viszont lehet bennük. Felhasználói csoportoknál statikusan beállított csoport létezik csak, a 14 év alatti és feletti olvasókra vonatkozó lekérdezések. Az olvasók igényeinek maximális kiszolgálása érdekében témafigyelési szolgáltatás építettek a rendszerbe, ahol a felhasználók maguk állíthatják be milyen téma, szerző, szakjelzet vagy adott mű érdekli őket, és erről a rendszer automatikusan értesíti őket email, vagy sms formájában.⁵⁴

Állományelemzési szempontból a következő lekérdezések érhetőek el: egy dokumentum kölcsönzései, kölcsönzött dokumentumok száma, letétben lévő állományokra vonatkozó adatok lekérése. A forgalomra vonatkozó statisztikákat részlegek szerinti csoportosításban is megtekinthetjük.

A rendszer lehetőséget biztosít az olvasók számának meghatározására, valamint a bejelentkezett felhasználók megtekintésére, illetve a naplófájl tartalmában is kereshetünk⁵⁵, ami egy későbbi adatbányászati projekt során nagy előny lehet. Menedzsment döntéseknél, döntéstámogatásnál a tartozások statisztikái segíthetik a munkát, emellett pedig a könyvtár megrendeléseit, és az ezekhez kapott kedvezményeket is tárolja a rendszer, ez segítheti a későbbi állománygyarapítás során a döntést, hogy mely forgalmazótól rendeljük meg az adott dokumentumokat. Nincs statisztika, mellyel a dolgozókra vonatkozó műveletek hajthatnánk végre.

Az úgynevezett *Top statisztikák* gyakorlatilag a már elkészült lekérdezéseket csoportosítják szerzők vagy cím szerint, és a találati halmazt aszerint jelenítik meg sorrendben melyikből van a legtöbb. Ezen belül pedig dokumentumtípus és műfaj (szak és szépirodalom) szerinti szűrésre is van lehetőségünk.

Saját lekérdezések létrehozására és trendek vizsgálatára nincs lehetőség, azonban pozitívum ennél a rendszernél, hogy az egyénre szabott témakövetés az olvasók tájékoztatását segítő opcióként jelenik meg.

⁵⁴ <http://www.szikla.net/szikla6kezikonyv.pdf> (Letöltve: 2014.01.13)

⁵⁵ <http://www.szikla.net/szikla7ujdonsagok.pdf> (Letöltve: 2014.01.13)

4.2.7. Az OliB

Az OliB (Oracle Libraries) integrált könyvtári rendszer Angliában, a Fretwell-Downing Informatics nevet viselő cég fejlesztésében jött létre, Magyarországon jelenleg a T-Systems forgalmazásában érhető el.

A szoftver rendelkezik egy külön statisztikákat létrehozó, úgynevezett kimutatáskészítő modullal. Ebben a modulban előre definiált jelentéseket találhatunk, több, mint 150 darab áll a rendelkezésünkre, melyeket egyénre szabhatunk könyvtárunk igényei mentén, emellett pedig saját jelentések definiálására is lehetőségünk van. A beépített lekérdezések a könyvtári munkafolyamatok minden területére vonatkozóan megtalálhatóak. Lehetőségünk van az állományban lévő dokumentumok forgásának vizsgálatára, kölcsönzési statisztikák készítésére, egyes olvasókhöz kapcsolódó adatok megtekintésére⁵⁶. Emellett vezetői jelentések készíthetők a szerzeményezés részfolyamataihoz kapcsolódóan, ezeket szállítók, megrendelések, dokumentumtípusonként csoportosíthatóak. A döntéshozatalt segítheti a különböző dokumentumigénylések iktatása, majd automatizált folyamatmenedzsment ezekhez a kérésekhez kapcsolódóan (felelős munkatársak kiválasztása, prioritásbeállítás). Az információkutatás, tájékoztatás folyamatát segítheti az olvasóhoz rendelt témakövetés beállítása, valamint a rendszer tudásbázis funkciója, mely véleményem szerint alkalmas lehetne adatbányászati keresésekre is. A lekérdezések egy adott időszakra vonatkoztathatók, viszont pozitívum, hogy lehetőség van az egyes jelentések bizonyos időszakonkénti lefuttatására.

4.2.8. Összegzés

Az integrált könyvtárak statisztikai moduljának elemzése után elmondható, hogy a Magyarországon használt integrált könyvtári rendszerek segítségével képesek vagyunk a könyvtár működéséhez kapcsolódóan alapstatisztikákat készíteni. A statisztikai modulok elemzése során kapott eredményeket két táblázatban összegeztem. Az 1. táblázatban összefoglaltam, hogy az egyes könyvtári rendszerek tartalmaznak-e statisztikai modult,

⁵⁶ Sajó Andrea: Az ORACLE LIBRARIES felhasználói szemmel. Budapest, Kossuth Lajos Tudományegyetem Természettudományi Kar, 1998.
<http://mek.oszk.hu/03100/03107/html/> (Letöltve: 2014.01.13)

egyenre szabható, illetve saját lekérdezéseket, valamint lehetőség van-e bennük trendek elemzésére. A második táblázatban (1. melléklet) összesítve jelenítem meg, hogy az egyes rendszerekben létrehozható lekérdezések mennyire felelnek meg a bibliomining célterületeinek, tehát milyen lekérdezések hozhatóak létre a gyűjteményszervezés, menedzsment és döntéstámogatás, valamint a tájékoztatás területén.

	Statisztikai modul	Egyénre szabható lekérdezések?	Saját lekérdezés?	Trendek
Corvina	Igen	Csak SQL-ben	Csak SQL-ben	Nem
Szirén	Nem	Igen	Nem	Nem
Aleph	Igen	Igen	Igen	Igen
Primo	Nem	Igen	Igen	Igen
Alma	Nem	Igen	Igen	Igen
HunTéka	Nem	Igen	Nem	Nem
LinLib	Igen	Nem	Nem	Nem
Szikla	Nem	Nem	Nem	Nem
Olib	Igen	Igen	Igen	Nem

1. Táblázat
Integrált könyvtári rendszerek statisztikai moduljai

Ahogy a fenti táblázatban is láthatjuk, a statisztikák sok esetben nem külön modulban jelennek meg, hanem egy-egy modul kiegészítéseként találkozhatunk velük. A legtöbb rendszerben a lekérdezések egyénre szabhatóak, főként időintervallumot tudunk beállítani, illetve szűréseket hajthatunk végre. A legtöbb esetben az látható, hogy beépített statisztikákat tartalmaz a rendszer, saját lekérdezések létrehozására nincs lehetőségünk. Ez alól kivételt képeznek az Ex Libris Kft. termékei (Aleph, Primo, Alma), valamint az Olib, melyek rendelkeznek saját lekérdezés létrehozására alkalmas felülettel, ezek mellett pedig a Corvina integrált könyvtári rendszerben tudunk SQL nyelven lekérdezéseket készíteni. Trendek vizsgálatára kevés integrált könyvtári rendszer esetében van lehetőségünk (Aleph, Primo, Alma), a rendszerek többségében ez csak úgy oldható meg, ha több különböző időintervallumra lefuttatjuk a lekérdezéseket, és ezeket később külön összehasonlítjuk.

Az eredmények összegzése után arra a következtetésre juthatunk, hogy jelenleg az Aleph könyvtári rendszere tartalmazza a legtöbb dinamikus jelentésgenerálásra alkalmas felületet, emiatt ez a rendszer bővíthető ki a legegyszerűbben, és ez lehet a legalkalmasabb az adatbányászati algoritmusok, eszközök beépítésére.

4.3. Könyvtárakban használható adatbányászati módszerek, algoritmusok

A *bibliomining* harmadik lépése a megfelelő elemző eszköz kiválasztása (ld. 4.1.2.2 fejezet) Ekkor választjuk ki milyen adatbányászati szoftvert szeretnénk használni, vagy hogy milyen adatbányászati algoritmust kell beépítenünk az általunk használt integrált könyvtári rendszerbe, hogy az általunk definiált problémákra megfelelő választ kaphassunk. Ehhez ismernünk kell a piacon lévő adatbányászati szoftvereket, illetve azokat az algoritmusokat, amelyek könyvtári területen alkalmazhatók.

A *bibliomining* során, amennyiben úgy döntünk, hogy külön adatbányászati szoftvert fogunk használni az elemzéshez, használhatjuk az általános elterjedt szoftvereket, mint például a kereskedelmi forgalomban elterjedt SPSS, vagy a SAS, de léteznek szabadon hozzáférhető ingyenes programok is, mint a WEKA (Waikato Environment for Knowledge Analysis), amely képes közvetlenül SQL lekérdezések segítségével kinyerni a feldolgozni kívánt adathalmazt az adatbázisból – jelen esetben a könyvtár integrált könyvtári rendszeréből.

Ahogy már említettük a *bibliomining* három eszközzel dolgozik, ezek a statisztika, az adatbányászat és a bibliometria. Az előző fejezetben tárgyaltak alapján láthatjuk, hogy az integrált könyvtári rendszerek már tartalmazznak statisztikákat. Ha úgy döntünk, hogy külön szoftvereket használunk a statisztikai, bibliometriai, adatbányászati elemzésekhez, tisztában kell lennünk azzal, hogy ebben az esetben, annak aki az elemző munkát végzi három különböző szoftver használatát kell megtanulnia, illetve a kinyert eredmények is három különböző platformon tekinthetők meg, valamint nem elhanyagolható szempontként három szoftvert kell megvásárolnunk, ha nem ingyenes programokat választottunk. Ha mindezeket el szeretnénk kerülni, és a már régóta használt integrált könyvtári rendszerünket bővítenénk ki egy olyan statisztikai modullal, amely képes statisztikai, bibliometriai és adatbányászati elemzésre, akkor viszont tisztában kell lennünk azokkal az algoritmusokkal, amiket az elemzés során használnunk kell. Emiatt célszerű a könyvtár egy-két munkatársának továbbképzése, és miután ők már az

alapbeállításokat megtették rendszerünkben, más könyvtárosok is használni tudják azt. Célszerű, hogy a háttérben zajló folyamatokat szakismeretekkel rendelkezők állítsák be, és az információkinyerést, a konkrét elemzést pedig egy interaktív, grafikonokkal, táblázatokkal gazdag felületen minden könyvtáros el tudja látni.

A következőkben be szeretném mutatni azokat a statisztikai, adatbányászati algoritmusokat, amik hasznosak lehetnek a *bibliomining* folyamata során, illetve bemutatom milyen feladatokra használhatóak, milyen célterületen tudjuk hasznosítani őket.

4.3.1. Adatfeltárás, statisztikai elemzés

Az egyszerűbb statisztikai módszerek egy része már az integrált könyvtári rendszerek részét képezi, a következőkben olyan statisztikai algoritmusokat szeretnék bemutatni, melyek ugyanolyan fontosak, mint a már beépített módszerek, de már közelebb állnak az adatbányászati algoritmusokhoz, mint az egyszerű alapstatisztikákhoz.

A **kovariancia** a változók együttmozgását vizsgáló statisztikai módszer. Azt vizsgálja, hogy változók két csoportja között milyen hatás érzékelhető. A kovariancia értéke háromféle lehet: pozitív, negatív vagy nulla. Ha a változók között pozitív kovariancia érték lép fel, abban az esetben az egyik változó átlag feletti pozitív értékéhez a másik változó átlag feletti értéke tartozik. Ugyanez negatív kovarianciánál úgy jelenik meg, hogy átlag feletti értékhez átlag alatti negatív érték társul. Amennyiben a kovariancia nulla, a két változó semmilyen hatással nincs egymásra, értékük egymástól függetlenül alakul. A kovarianciát úgy számoljuk, hogy mindkét változócsoporthoz összes elemének az adott változócsoporthoz átlagától vett különbségét vesszük, majd ezeket összeszorozzuk, végül a szorzatok összegét osztjuk az elemszám mínusz 1-gyel.

A **kovariancia segítségével** a könyvtárral, az olvasókkal, a könyvekkel kapcsolatos tulajdonságok, változók egymásra hatását tudjuk vizsgálni, és ezek alapján kaphatunk képet könyvtárunk működéséről, és javíthatjuk azt. A kovarianciát remekül tudjuk használni az olvasók igényeinek feltérképezésére, szokásaik megértésére. Emellett gyűjteményünk elemzése és gyarapítása során is használható. Képet kaphatunk arról, hogy olvasóközönségünket érdeklő könyveket szerzünk-e be, ha kiszámítjuk a könyvek

számának változásának/növekedésének és a kölcsönzések számának kovarianciáját. Amennyiben pozitív értéket kapunk, akkor rendben működik a beszerzésünk. Témafigyelési szolgáltatásunk működését vizsgálhatjuk azzal, ha újonnan beszerzett könyvekről kiküldött értesítések és a könyvtárlátogatás vagy a kölcsönzés kovarianciáját mérjük. Emellett a könyvtárról általános információkat is szerezhethetünk, például ha a nyitvatartási időt vizsgáljuk, bizonyos időszakokban a meghosszabbítása eredményez-e több könyvtárlátogatást.

A **korreláció** a kovarianciával rokon statisztikai fogalom, a kovariancia skálázott változata, mely a változók között lineáris kapcsolatot feltételez. A változók együttmozgását vizsgáló együttható, mely azt határozza meg, hogy az egyik változó növekedése maga után vonja-e egy másik változó növekedését. Tehát ebben az esetben nem csak azt vizsgáljuk, hogy van-e valamilyen kapcsolat két változó között, hanem azt is, hogy tendenciális kapcsolatban állnak-e egymással. Míg a kovariancia elsősorban két változó egymásra hatását tudja mérni, addig a korreláció több változóra is kiterjeszthető, úgynevezett korrelációs mátrixra végezzük el ebben az esetben a korreláció értékének a meghatározását. Ugyanúgy, ahogy a kovarianciánál, a korreláció is pozitív, negatív és nulla értékeket vehet fel.

Bármilyen területen is dolgozunk, az átgondolt, megtervezett munkafolyamat összeállításához hasznos, ha ismerjük azt, hogy egyes cselekedeteink, bizonyos döntések, értékek változása milyen következményekkel jár. A könyvtári munkában sokszor lehet segítségünkre az, ha tudjuk, hogy mi motiválja olvasóinkat. Hasznos lehet például annak vizsgálata, hogy a könyvtárba beiratkozottak könyvtárlátogatása és lakóhelyük távolsága között milyen összefüggés van. Bármilyen szolgáltatás időbeli vizsgálatához is eredményes lehet a korreláció, a korrelációs mátrix egyes oszlopaiban tárolt, egy adott szolgáltatás egyes hónapokban való igénybevételének elemzésével például. Ha számszerűsítjük a különböző korcsoportba tartozók online katalógus használatát, akkor megfigyelhetnénk milyen korosztályt kell megcéloznunk annak érdekében, hogy kiterjesszük az opac népszerűségét olvasóink körében. Ha valamilyen tulajdonságokkal rendelkező, például egy adott szakra járó hallgatók online katalógus használatánál

negatív korrelációt tapasztalunk, tanfolyam szervezését is megfontolhatjuk, ahol fejleszteni tudjuk a hallgatók könyvtárhasználati ismereteit, információs műveltségét is.

A regressziószámítás, vagy **regresszióanalízis** során két vagy több véletlen változó között fennálló kapcsolatot vizsgáljuk. A regresszióanalízis olyan modellezési módszer, melynél a megbecsülendő változó értéke folytonos.⁵⁷ A kovarianciával és a korrelációval ellentétben itt nem két, vagy több változó közötti kapcsolat felderítése a cél, hanem több változó értéke alapján egy előre meghatározott változó értékének megbecslése. A regresszióanalízisen belül megkülönböztetünk egy- vagy többváltozós, illetve súlyozott *lineáris* regressziót, valamint *nemlineáris* regressziót, emellett pedig az adatok milyensége alapján *időszori*, *keresztmetszeti*, és *panel* regresszióanalízisről beszélhetünk.

Könyvtári környezetben megvalósult példa lehet regresszióanalízisre, tudományos tevékenységet folytató könyvtárak esetében, szakfolyóiratok befolyásának vizsgálatakor regresszióanalízis segítségével felmérni a folyóirathoz tartozó szerzők, hivatkozók és előfizetések földrajzi eloszlását.⁵⁸ A regresszióanalízist üzleti környezetben gyakran használják az egyes termékek eladási rátájának meghatározására a hirdetésre költött pénz függvényében. Ehhez hasonlóan az információgazdálkodási intézményekben vizsgálható az egyes szolgáltatások árának meghatározása annak függvényében mennyi pénzt számunk annak megismertetésére, elterjesztésére.

A különböző adatfeltárást segítő eszközök közül talán a leghatékonyabb a többdimenziós adatelemzést segítő **OLAP**, melyet az adatbányászatról szóló fejezetben már vázoltam (ld. 4.1.1). Az eszköz több általánosan elterjedt táblázatkezelő szoftverbe is beépítették már, így egy könnyen használható módszerről beszélünk. Amennyiben adattárházban tároljuk adatainkat, készen állnak számunkra azok az *adatkockák*, amik az OLAP módszerben használt dimenziókat tudják szolgáltatni. Ha relációs adatbázissal rendelkezünk, akkor az adatbázistáblák egyes mezőit tudjuk dimenzióként használni. Ezen mezők metszeteit tudjuk vizsgálni, aggregálni, szűrni. Az OLAP könyvtári

⁵⁷ Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin: Adatbányászat: Alapvetés. Budapest, Panem, 2012. 767.p.

⁵⁸ Wormell, Irene: Az informetria és a bibliográfiai adatok használata a stratégiai tervezésben. In: Mikulás Gábor (szerk.): Információból üzleti érték: az információbróker környezete és munkája. Budapest, Magyar Információbrókerek Egyesülete, 2006. 252.p.

használatára remek példa a norvég statisztikai hivatal honlapja⁵⁹, ahol az ország könyvtárügyével kapcsolatos statisztikákat egy dinamikus analitikus feldolgozó felületen tudjuk elemezni, ahol az előre definiált kérdés statisztikai adatait általunk kiválasztott dimenziók, szűrőfeltételek segítségével elemezhetjük, dinamikus jelentéseket hozhatunk létre, majd az oldal automatikusan generál számunkra diagramot is. A norvég adatbankban olyan statisztikák egyénre szabására lehetőség van, mint a közkönyvtárak látogatottsága nem, kor, és iskolai végzettség szerint, a különböző könyvtártípusok kölcsönzési adatai a kölcsönzések típusai, és időpont szerinti bontásban, vagy a könyvtárban dolgozókra vonatkozó statisztikák.⁶⁰

Norwegian cultural barometer

Create your own tables and graphs

1 Select the table containing the variables you want 2 Select values from different variables 3 See, export or save your custom table

Table: 05300: Distance to nearest place where different kinds of culture activities are available (per cent)

My table Select via search Select via groups Information Log on

Contents 1 of 1

Distance to nearest place - Unit: per cent

Distance 3 of 6 Cultural activities 1 of 8 Year 2 of 7

Under 1 km
1-4.9 km
5-9.9 km
10-24 km
25-49 km
50 km or more

Concert hall/locality used for concerts
Art gallery/art collection
Museum
Sports ground/sports hall
Library/public book collection
Book shop

2012
2008
2004
2000
1997
1994

So far you have selected 6 figures - at most, Select values from the listboxes and click 'Show table'.

Show table >>

8.ábra

A Norvég Statisztikai Hivatal jelentésgeneráló felülete

Norwegian cultural barometer

Create your own tables and graphs

1 Select the table containing the variables you want 2 Select values from different variables 3 See, export or save your custom table

Table: 05300: Distance to nearest place where different kinds of culture activities are available (per cent)

Log on

Rotate table Sort table Edit table Graphical presentation Save as

Show code/text OK Line OK Excel OK

Distance to nearest place (per cent), by distance, cultural activities, time and contents

	2000	2012
Distance to nearest place		
1-4,9 km		
Library/public book collection	51	50
10-24 km		
Library/public book collection	10	11
50 km or more		
Library/public book collection	0	1

Open in new window Print

9.ábra

A Norvég Statisztikai Hivatal jelentésgeneráló felület – diagramgenerálás

Saját rendszerünkbe építve az egyes metszetek a következők lehetnek: időintervallum, tagkönyvtár, korcsoport, kölcsönzések száma, gyűjteményhasználat, dokumentumok forgási sebessége, szolgáltatások használata, dolgozókra vonatkozó adatok. Ezeket kombinálva tetszőleges mélységű adatokat generálhatunk, és a saját rendszer előnyét kihasználva nem csak előre definiált kérdésekre találhatjuk meg a választ, hanem mindig a könyvtár számára legaktuálisabb kérdéskör felderítésére használhatjuk a rendszert.

4.3.2. Dimenziócsökkentés

Az eddig bemutatott algoritmusok nagy része különböző, a vizsgált rendszerre jellemző változókat, és az általuk leírt tulajdonságok egymásra hatását vizsgálta. Ezekkel a módszerekkel az a probléma, hogy nagy méretű modell esetében olyan nagy mennyiségű változóról beszélünk, hogy szinte lehetetlen minden egyes változónak az összes többivel való kapcsolatát megvizsgálni, és valójában értelmetlen is. Azonban nem mindig egyértelmű, hogy melyik változó bír információértékkel, melyik lesz hasznos számunkra, és hogy mely változók hatnak egymásra. Amennyiben konkrét kérdésekre keressük a választ, hasznosak az eddig bemutatott algoritmusok, amikor pedig már nem rendelkezünk konkrét tudással, akkor le kell szűkítenünk a változóink körét, ehhez használható a *főkomponens analízis*, és a *faktoranalízis*.

A **főkomponens analízis** egy olyan, az adatbányászatban használt statisztikai eljárás, amely változók egy csoportját alakítja át egy lineáris transzformáció segítségével (A koordináta-rendszer olyan elforgatásával, amely azt eredményezi, hogy a tengelyek rendre az adathalmaz legnagyobb szórásainak irányába állnak be⁶¹) egy kevesebb elemből álló változócsoporthá. A transzformáció után megkapott változókat nevezzük főkomponenseknek, ezek nem korrelálnak egymással, és az eredeti információk legnagyobb részét megőrzik. A főkomponensek kiválasztásánál az egyes változók korrelációját mérik, és amennyiben ez közel nulla, akkor a változót nem tekintik relevánsnak, és kihagyják az elemzésből.

⁶¹ <http://www.konkoly.hu/~veres/grb/d1/node51.html> (Letöltve: 2014.01.15)

A főkomponens analízis segít kiválasztani számunkra azokat a változókat, amelyek leginkább fontosak a könyvtár működése szempontjából, amiket stratégiai céljaink meghatározásakor leginkább figyelembe kell vennünk. Emiatt gyakorlatilag a *bibliomining* mindhárom területén alkalmazható válik, mert a gyűjteményszervezés során is meg kell keresnünk mik azok az alapelvek, ami alapján beszerzünk vagy leselejtezzünk műveket, a felhasználókhöz való elérést is megkönnyíti, ha tudjuk milyen preferenciáikat kell megcéloznunk, és a tájékoztatási apparátus megszervezéséhez is segítséget tud nyújtani.

A **faktoranalízis** a többváltozós statisztika módszerei közé tartozik, a megfigyelt változók lineáris kombinációjából képez új változókat, úgynevezett faktorokat. A faktoranalízis elve szerint ezekkel a faktorokkal a teljes adattömeg viszonylag kis információvesztéssel reprezentálható.⁶² A faktorok készítése során a változók egymásra hatását mutató korrelációs vagy kovariancia mátrixot készítünk, és megkeressük az egymással magas korrelációban lévő változókat, és ezeket egy faktorban egyesítjük. A faktorok megfelelő azonosításához az algoritmus elvégzi a faktorok rotációját, melynek során a magas korrelációjú változók még inkább előtérbe kerülnek, és ezek alapján határozza meg a végleges faktorokat. Egy elemzés során akár több faktor használata is szükséges lehet, amennyiben a feladat megoldásához több változócsoporthoz tartozó tulajdonság elemzése szükséges. A faktoranalízis nem csak üzleti területen használatos, használják a pszichológiában a személyiség modelljének meghatározásához, szociológia területen, operációkutatásnál, de mindenhol használható, ahol nagy változómenyiség leredukálására, összevonására van szükség.⁶³

A főkomponensanalízissel ellentétben ez az eszköz nem csak kiválaszt egy fontosnak vélt változócsoporthoz, hanem azokat egy kitüntetett változóban egyesíti. A könyvtár szempontjából ez az egyesítés lehető teszi, hogy az intézmény által megcélózni kívánt területeket különválasszuk. A könyvtári munka során fontos, hogy az olvasóknak az őket érdeklő dokumentumokat szolgáltatassuk. Ebben nyújt segítséget, ha a felhasználói csoportokat egy faktorral azonosítjuk, és a későbbiekben az egyénre szabott

⁶² Fajszí Bulcsú, Cser László: Üzleti tudás az adatok mélyén: Adatbányászat alkalmazói szemmel. Budapest, Budapesti Műszaki és Gazdaságtudományi Egyetem, 2004. 148.p.

⁶³ <http://hu.wikipedia.org/wiki/Faktoranal%C3%ADzis> (Letöltve: 2014.01.14)

szolgáltatások kialakításakor már egyszerűbben érhetjük el őket. Ilyen faktorok lehetnek a tudományos irodalom iránt érdeklődő kutatók, közkönyvtár esetén a sikerkönyvet kereső fiatalok, a szolgáltatásokat maximális igénybe vevő olvasó, a könyvtártól távol élő, ezért ritkán könyvtárba járó nyugdíjas stb. A faktorok segíthetik gyűjteményszervező munkánkat is, meghatározhatjuk vele azokat a tényezőket, amik alapján egy dokumentum egyezik a felhasználói igényekkel, gyakran használják, kölcsönzik, előjegyzéseket kell felvenni rá. A könyvtári programok szervezése során is hasznos lehet, ha tudjuk milyen tényezők befolyásolják azt, hogy a könyvtárhasználókat érdeklik-e az adott programok, vagy sem. A faktoranalízis tehát hasznos lehet a könyvtár számára önmagában is, valamint segítséget nyújthat későbbi elemzések során, például amikor felhasználói csoportokat hozunk létre.

4.3.3. Csoportosítást segítő algoritmusok

Mindennapi életünk során is gyakran hajtunk végre csoportosító algoritmusokat, kiválasztjuk a meghirdetett tantárgyak listájából azokat, melyeket az adott félévben fel szeretnénk venni, a könyvtárban kiválasztjuk azoknak a könyveknek a csoportját amit kölcsönözni szeretnénk, és ezzel egy időben létrehozuk azoknak a könyveknek a csoportját is, amit nem szeretnénk használni. Könyvtárak esetében a csoportosító adatbányászati algoritmusok elsősorban a gyűjteményszervezés és a szolgáltatásmenedzsment területén használhatók.

A **klaszteranalízis** az egyik legelterjedtebb csoportosító eljárás, a meglévő adattömbben meglévő adatokból jól elkülöníthető, homogén csoportokat képez, ezeket a csoportokat nevezzük klasztereknek.⁶⁴ Az egyes klasztereken belüli adatok valamilyen tulajdonság alapján jobban hasonlítanak egymáshoz, mint a többi klaszterekben lévő elemekhez. A hasonlóságot különböző távolság és hosszsmértékegységekkel állapítják meg. A klaszterező algoritmusok megbízhatósága nagyban függ az adattömb méretétől, és a megfelelő algoritmus kiválasztásától.

A dimenziócsökkentő, klaszterező eljárások közül az egyik legmegbízhatóbb a **k-közép klaszterezés**. A prototípus alapú klaszterező módszerek közé tartozik, ami azt

⁶⁴ Abonyi János: Adatbányászat a hatékonyság eszköze: Gyakorlati útmutató kezdőknek és haladóknak. Budapest, ComputerBooks, 2006. 133.p.

jelenti, hogy a csoport elemei mind közelebb vannak a klasztert definiáló prototípushoz, mint az összes többi klaszter prototípus eleméhez.⁶⁵ A módszer véletlenszerűen létrehoz n számú klasztert, meghatározza ezek prototípus elemét, mégpedig úgy, hogy olyan elemet választ prototípusnak, amely egy adott pontcsoport átlaga. Ezek után az algoritmus minden egyes elemet ahhoz a klaszterhez sorol be, amelynek középpontjához legközelebb esik az adott elem. Ahhoz, hogy megfelelő megbízhatóságú legyen az algoritmus, a prototípusok kiválasztása több iteráció során alakul ki, az algoritmust egymás után többször le kell futtatni, a kiválasztást egészen addig kell folytatni, amíg a besorolás nem változik. Az algoritmus előnye egyszerűségében és gyorsaságában rejlik, hátránya az egyes futások során más számú csoportok létrehozása, ami az algoritmus többszöri futtatásával kiküszöbölhető.

A k-közép klaszterezés segítségével a könyvtár felhasználói csoportokat tudna létrehozni érdeklődési kör, életkor, lakóhely, végzettség vagy ezek kombinációi alapján. A csoportosítás végrehajtásában segíthetnek egy korábbi elemzés során létrehozott olvasói rétegeket leíró faktorok. A létrehozott klaszterek olvasóinak személyre szabott értesítések küldhetőek, programok szervezhetőek számukra, illetve egyénre szabott szolgáltatások kialakításánál, állományunk egyes részeinek népszerűsítésénél használhatók.

A **diszkriminancia-analízis** olyan statisztikai módszer, amely azt vizsgálja, hogy mely változóknak különböznek legszignifikánsabban az adatokat leíró csoportok.⁶⁶ Alkalmazásával tesztelni tudjuk a klaszteranalízis során kapott eredményeket, mivel az algoritmus eredménye megadja, hogy egy adatelem tagja-e valamely csoportnak. A **territoriális térképek** segítségével meghatározhatjuk, hogy mi az a változó (nem, kor, iskolai végzettség, érdeklődési kör, lakóhely távolsága, kölcsönzések száma stb.), ami alapján leginkább elkülönülnek a csoportok, így ez is segítségünkre lehet csoportok létrehozásában.

⁶⁵ Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin: Adatbányászat: Alapvetés. Budapest, Panem, 2012 520 p.

⁶⁶ Fajsi Bulcsú, Cser László: Üzleti tudás az adatok mélyén: Adatbányászat alkalmazói szemmel. Budapest, Budapesti Műszaki és Gazdaságtudományi Egyetem, 2004. 167.p.

4.4.A bibliomining gyakorlati alkalmazása könyvtártípusok szerint

Az eddigiekben általánosítva mutattam be milyen területeken használható a bibliomining, milyen adatbányászati algoritmusok alkalmazhatóak könyvtári területen. Dolgozatom záró fejezetében azonban arra szeretném felhívni a figyelmet, hogy a bibliomining alkalmazhatósága a különböző könyvtártípusok esetén eltérő mértékben és eltérő módon jelenhet meg. Úgy, ahogy más-más módon kell eljárni az adatbányászat bevezetésével a különböző üzleti területeken, nem létezik általános séma a könyvtárakban való alkalmazására sem. A különböző profilú intézmények számára a bibliomining eltérő lehetőségeket kínál, mást tudnak felhasználni az adatbányászat nyújtotta lehetőségekből. Egyes könyvtáraknak a gyűjtemény elemzésében nyújt segítséget, van ahol a profilhoz való illeszkedés vizsgálata szükséges, vagy a munkafolyamatok optimalizálása, a tájékoztatás segítése. Az alábbiakban be szeretném mutatni, hogy egyes könyvtártípusok esetén az adatbányászat mely szelete használható fel a leghatékonyabban.

4.4.1. A közkönyvtárak

„...A közkönyvtár feladata a gondolatok szabad kifejezésének és befogadásának elősegítése, a közjó szolgálata. Legfőbb kommunikációs eszköze az olvasás, mint a kultúra megszerzésének egyedülálló módja, miközben minden médiumot befogad, és lehetőséget nyújt mindenféle információ és ismeret megszerzésére”⁶⁷. A közkönyvtár legfőbb feladata az információhoz való hozzáférés biztosítása az állampolgárok számára, a lakosság felől érkező igények kielégítése. Emellett közművelődési, nevelési, szórakoztató és szociális funkcióval rendelkezik.

A bibliomining szempontjából a különböző közkönyvtárak között elsősorban méretük, és esetleges további funkciójuk alapján tudunk különbséget tenni. A kis méretű, elsősorban falusi közkönyvtárak esetében felmerül az a kérdés, hogy egyáltalán szükség van-e az adatbányászat használatára. Mivel kis méretűek ezek a könyvtárak, kevés

⁶⁷ Horváth Tibor, Papp István (szerk.): Könyvtárosok kézikönyve. 3. kötet. Budapest, Osiris Kiadó, 2003. 69.p.

dolgozót foglalkoztatnak, kisebb gyűjteménnyel rendelkeznek, valamint kisebb olvasóforgalmat bonyolítanak le, számítógépes segítség nélkül is képesek átlátni a könyvtár mozgatórugóit, megszervezni saját működésüket, kiszolgálni az olvasói oldal felől érkező kívánságokat. A felmerülő igényeket ki tudják szolgálni, azonban nem tudnak olyan szolgáltatásokat nyújtani, amik potenciálisan érdekesek lehetnek olvasóiknak, ám ők nem kérik ezeket. Az ilyen típusú könyvtárakban a munkafolyamatok menedzselésének és a tájékoztatási munkának bibliomining-gal való segítése elhanyagolható előnnyel jár, hiszen a tapasztalatok információértéke ebben az esetben felülmúlja az adatbányászat nyújtotta előnyt. Ezen típusú könyvtárakban a bibliomining a gyűjteményszervezés terén nyújthat segítséget, elsősorban azzal, hogy közös célt kitűző, azonos profilú (gyűjtemény, felhasználói csoport) könyvtárakban zajló folyamatokat hasonlítjuk össze sajátjainkkal. Ezen könyvtárak hasonló tulajdonságokkal, érdeklődési körrel rendelkező olvasói csoportjai által kölcsönzött dokumentumoknak helye lehet saját gyűjteményünkben is, így a kis méretű közkönyvtár úgy tudja bővíteni saját állományát, hogy nagy megbízhatósággal tudja azt, hogy a megrendelt dokumentum használatba fog kerülni. Közös adattárházban tárolva könyvtárunk adatait, az összehasonlító elemzéseket könnyen végre tudjuk hajtani, egy közös jelentésgeneráló felületen pedig valós idejű információkat szerezhetnénk közös célunk megvalósítása érdekében.

A nagyobb méretű közkönyvtárak, úgy mint a városi, megyei könyvtárak esetén egyértelmű, hogy sokkal több munkatárs látja el ugyanazokat a munkafolyamatokat. Több munkatárs eltérő időpontokban ül az olvasópultban, ezért más-más olvasói igényekkel találkozhatnak. Ebben az esetben az adott könyvtárosok rendelkeznek bizonyos képpel, hogy milyen olvasói igények keletkeznek, azonban globális képet nem kapnak erről. Emiatt szükséges, hogy az igényeket valamilyen módon dokumentáljuk, majd elemzéseket hajtsunk végre rajtuk, és ez alapján biztosan fogjuk tudni hogyan kell jobbra tennünk szolgáltatásainkat. Ezt kiegészítendő a megemlített feladatok mellett megyei könyvtárak esetén, akik a könyvtárközi kölcsönzés menedzselésében részt vesznek, dokumentumaik elosztásának optimalizálásához is használni tudják az adatbányászatot. A feladatukként megjelenő különböző típusú könyvtárak koordinációjának segítségét is egyszerűsítheti a bibliomining. Ezek mellett pedig a

felhasználókat célzó programok kialakítását segítő algoritmusok használata ezekben a könyvtárakban használható a legnagyobb mértékben. Emiatt az ilyen típusú könyvtárak a legalkalmasabbak a bibliomining bevezetésére, itt a bibliomining mindhárom célterületén eredményeket érhetünk el.

4.4.2. A szakkönyvtárak és a felsőoktatási könyvtárak

A *szakkönyvtárak* feladata az egyes tudományterületeken megjelenő információk összegyűjtése és szolgáltatása, a tudományos kutatás támogatása. Ha áttekintjük azokat a területeket, ahol megjelenhet a bibliomining, láthatjuk, hogy itt a gyűjteményszervezésben kisebb szerepet kap, mivel a könyvtár egy olyan szűk, körülhatárolt területen megjelenő dokumentumokat gyűjt, ahol az olvasói igények előre definiáltak, konstansak. Esetlegesen a más szakkönyvtárak gyűjteményével való összehasonlítás alkalmazható, ennek segítségével kiegészíthetjük, ellenőrizhetjük állományunkat. Emellett a gyűjtemény bibliográfiai adatait elemezve megállapítható, hogy az adott szakterület legismertebb, leghasznosíthatóbb dokumentumai közül melyek vannak könyvtárunkban leginkább használatban, és melyek azok, amelyek ugyan a rangsorokban nem kiemelkedő helyen szerepelnek, de nálunk sikeresek⁶⁸. A szakkönyvtárak esetében elsődlegesen a tájékoztatás területén használható a bibliomining, mivel ezek azok a könyvtárak, ahol a legtöbb ilyen típusú kérdés megjelenik, így egy tudásbázist építésére nyílhat lehetőség, melyben kereső algoritmusok futtathatóak.

A *felsőoktatási könyvtár* feladata az oktató, kutatómunka támogatása, valamint egyéb szakkönyvtári feladatokat is el kell látnia.⁶⁹ A könyvtár gyűjteményében a felsőoktatási intézmény által oktatott összes tudományág alapidokumentumainak meg kell lennie, valamint tájékoztatási szolgáltatást kell biztosítania ezekhez a területekhez

⁶⁸ Nicholson, Scott. - Stanton, Jeffrey.: Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries.
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.1835&rep=rep1&type=pdf> (Letöltve: 2014.01.14)

⁶⁹ Horváth Tibor, Papp István (szerk.): Könyvtárosok kézikönyve. 3. kötet. Budapest, Osiris Kiadó, 2003. 133.p.

kapcsolódóan. Ezek alapján két fő funkciója van, az egyetemi közösség kiszolgálása, valamint a könyvtári rendszerbeli feladatok végrehajtása.

Az oktatómunka támogatása bibliometriai elemzések végrehajtásával is megvalósítható, amely a bibliomining eszköztárának egyik eleme. Ilyen elemzést hajtott végre a svéd könyvtári és információs tanintézet, az északi könyvtárak tematikus portáljának vizsgálatakor.⁷⁰ A vizsgálat keretein belül a felsőoktatási intézmény oktatóit, és az oktatóknak az adott portálon megjelent tanulmányait vizsgálták, céljuk a tanszéken megjelent kutatások trendjének feltérképezése volt. Egy másik lehetséges felhasználása a bibliominingnak a könyvtár szakfolyóirat használatának vizsgálata. A használat olyan szempontú elemzése, hogy az egyetemen szakdolgozatot írók mely folyóiratokat használták, és a kutatók mely folyóiratokat használták fel kutatásaikhoz. A szakdolgozatok, publikációk hivatkozásait elemezték, és ez alapján hoztak döntéseket arról, hogy a könyvtár által megrendelt folyóiratok közül melyek vannak használatban, és melyek nincsenek. Emellett természetesen a kutatók publikációs tevékenységének vizsgálata is a könyvtár feladata. A különböző bibliometriai szoftverek ebben segítheti a felsőoktatásban dolgozó könyvtárosokat.

Az adatbányászati algoritmusok közül a felsőoktatási könyvtárakban, ugyanúgy, ahogy a szakkönyvtárak esetében is, a tájékoztatási adatbázis elemzését végző algoritmusok lehetnek hasznosak.

4.4.3. Az iskolai könyvtárak

Az iskolai könyvtár feladata az iskolai oktatás, nevelés támogatása, a tanulással, tanítással kapcsolatos dokumentumok és információk szolgáltatása. Az iskolai könyvtáraknak is egyre nagyobb adatmennyiséggel, és egyre több kihívással kell megküzdeniük. Egyrészt ki kell szolgálniuk az oktatói, hallgatói igényeket, emellett pedig ők azok, akik legközelebb állnak a tanulókhöz, megfelelő hozzáállással elősegíthetik, hogy a potenciális olvasókból az olvasást szerető felnőtt válhasson. Az igények kiszolgálása mellett az iskolai könyvtár megfelelő infrastrukturális háttérrel

⁷⁰ Wormell, Irene: Az informetria és a bibliográfiai adatok használata a stratégiai tervezésben. In: Mikulás Gábor (szerk.): Információból üzleti érték: az információbróker környezete és munkája. Budapest, Magyar Információbrókerek Egyesülete, 2006. 252.p

rendelkezik ahhoz, hogy az iskola működéséhez is segítséget nyújtson. Jelenleg a különböző iskolai könyvtári rendszerek nem alkalmasak az adatvezérelt döntéshozatalra, trendek feltérképezésére, adatbányászati elemzésekre. A Mankoto Area Public School igazgatója vetette fel 2004-ben megjelent cikkében⁷¹ az adatbányászat iskolai könyvtári szükségességét. Doug Johnson egy amerikai iskolában dolgozik, ahol a tanulók teljesítményének mérésére minden tantárgyból különféle teszteket használnak. Ő vetette fel annak lehetőséget, hogy ezen tesztek eredményét a könyvtár elemzés alá vethetné, és az eredmények alapján célzottan kereshetné meg a diákokat könyvajánlásokkal, felzárkóztató programokkal, olvasóklub lehetőségével. Ezekkel a lehetőségekkel a diákok teszteredményei javíthatóak, valamint a diákok egyénre szabott szolgáltatásokat kaphatnak, és ha érzik, hogy foglalkoznak az igényeikkel, szívesebben látogatnak el az iskolai könyvtárba. Ezen könyvtártípus nem csak a saját beszerzéseit, döntéseit tudja adatbányászati eszközökkel menedzselni, hanem magának a fenntartó intézményt, tehát az iskolát is. Amennyiben az iskolai menedzsmenttel kapcsolatos döntéseket adatbázisban tárolják, adattárház létesíthető, amelyben együttesen tárolhatóak a könyvtárral, és az iskolával kapcsolatos információk, és az iskolával kapcsolatos döntések meghozatalában is segítséget nyújthat a könyvtár.

Ezek alapján egyértelmű, hogy az iskolai könyvtárat elsősorban a döntéshozatallal, és a tanulókat célzó programok kialakításának segítségével érheti el a bibliomining.

4.4.4. A nemzeti könyvtárak

A nemzeti könyvtárak elsődleges feladata minden országban a nemzeti dokumentumállomány összegyűjtése és megőrzése. Emiatt a bibliomining elsősorban a dokumentumelemzés területén hasznosítható, valamint a gyűjteményünk kiegészítésében játszhat szerepet. A más országok gyűjteményében való keresés segítséget nyújthat, hogy az esetlegesen a gyűjtőkörünkbe tartozó, de még nem meglévő dokumentumok is a látókörünkbe kerüljenek. Az ilyen elemzéshez használható az OCLC által fejlesztett Collection Evaluation névre hallgató elemző eszköz, mely a OCLC tagkönyvtárak

⁷¹ Johnson, Doug: A data mining primer and implications for school library media specialist In: Knowledge Quest, 2004. (32. évf.) 5. sz. 32-35 p.

kollekcióiban keres, saját gyűjteményünket hasonlíthatjuk össze a tagkönyvtárak dokumentumállományával, olyan kérdésekre választ kapva, mennyire kurrens a dokumentumállományunk, milyen hiányzó dokumentumaink vannak, gyűjteményünk melyik része számít egyedinek az OCLC tagkönyvtárainak körében.

4.4.5. Az integrált könyvtárak

Napjainkban sok olyan könyvtár létezik, amelyben több könyvtártípus egyszerre van jelen, ilyen a Pécsen található Dél-dunántúli Regionális Könyvtár és Tudásközpont is, mely egy épületben biztosít helyet gyermekkönyvtár, városi könyvtár, egyetemi könyvtárak számára. Az ilyen, több könyvtártípust integráló könyvtárak ideálisak a bibliomining használata szempontjából, hiszen minden eddig bemutatott módszer alkalmazható itt, nagyszerű segítség lehet a munkafolyamatok, programok, állomány összehangolásában.

5. Összegzés

A szakirodalom áttanulmányozása után kiderült számomra, hogy az adatbányászat használata napjainkban még csak elszórtan jelent meg, léteznek ugyan az Egyesült Államokban alkalmazási példák, azonban tömeges elterjedésről még nem beszélhetünk. A téma feldolgozása sem teljes még, rengeteg szakirodalom szól adatbányászatról, azonban a könyvtárak adatbányászati szokásainak feltérképezése még hiányos.

Kutatásom során arra a következtetésre jutottam, hogy a könyvtári élet minden területén, a gyűjteményszervezés, tájékoztatás, menedzsment és a döntéstámogatás területén is hasznosítható az adatbányászat. Az integrált könyvtári rendszerek statisztikai moduljának elemzése igazolta hipotézisemet, hogy a Magyarországon használt könyvtári szoftverekbe nincsenek beépülve adatbányászati algoritmusok, így az adatbányászat könyvtárakban való használata sem terjedt még el. Az adatbányászat bevezetésének első lépése a már létező rendszerek statisztikai moduljának kiterjesztése lehet, ebből a szempontból az Aleph könyvtári rendszert használó könyvtárak állnak a legjobban, mivel ez az a rendszer amely már tett lépéseket az interaktív, döntéseket segítő jelentésgenerálás bevezetése felé, és ennek továbbfejlesztése egyszerűen megoldható lenne az adatbányászat segítségével. Kutatásom során meghatároztam azokat az adatbányászati algoritmusokat, melyek használhatóak a könyvtárakban, ezzel igazolódott hipotézisem, hogy az adatbányászati módszerek egy része felhasználható könyvtári területen. Több olyan bibliometriai, adatbányászati szoftver létezik, mellyel kiegészítve meglévő rendszereinkkel eredményeket tudunk elérni. Ezek mellett pedig az adatbányászati algoritmusok könyvtári rendszerekbe való integrálása az a lépés, amellyel akár Magyarországon is elterjedté válhat az adatbányászat. A kutatás eredményeképpen elmondható, hogy ugyan láthatunk különbségeket a különböző könyvtártípusok esetében, azonban minden könyvtár fel tudja használni saját céljai eléréséhez, szolgáltatásainak javításaihoz az adatbányászati módszereket.

6. Bibliográfia

Abonyi János: Adatbányászat a hatékonyság eszköze: Gyakorlati útmutató kezdőknek és haladóknak. Budapest, ComputerBooks, 2006.

Adriaans, Pieter , Zantinge, Dolf: Adatbányászat. Budapest, Panem, 2002.

Bakonyi Géza, Kokas Károly: Könyvtári integrált rendszerek és hazai alkalmazásuk

<http://www.mek.iif.hu/porta/szint/tarsad/konyvtar/automat/intsys/html/libsys.htm>
(2014.01.04)

Berendt, B., Krause, B., Kolbe-Nusser, S. : Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks In: Information Processing & Management, 2010 (46. évf.) 1. sz. 1–10. p.

Celko, Joe: SQL felsőfokon. Budapest, Kiskapu, 2002.

Cvejic, Radoje, Mijailovic, Jelena: Knowledge Management and Changes Management In University Libraries, In: Analele Universitatii "Eftimie Murgu" Resita, 2009 (1. évf.) 16. sz. 74-82 p.

Csajbók Zoltán, Sütő Gergely: A számítástechnika története. Budapest, ÉGSZI-Scola Alapítvány, 1992.

Demaine, Jeffrey, Dumouchel, Bernard: Knowledge Discovery in the Digital Library: access tools for mining science. In: Information Services and Use - Electronic Tools for Knowledge Discovery and Health, 2006. (26. évf.) 1. sz. 39-44. p.

Dickey, Timothy J.: Books as Expressions of Global Cultural Diversity: Data Mining for National Collection Analysis. In: Library Resources & Technical Services, 2011. (55. évf.) 3. sz. 148-162.p.

Eszenyiné Borbély Mária: A Magyarországon alkalmazott könyvtári szoftverek értékelése a többtényezős döntéshozatal módszerével
http://tmt.omikk.bme.hu/show_news.html?id=1800&issue_id=49 (2014.01.13)

Fajsi Bulcsó, Cser László, Fehér Tamás: Üzleti haszon az adatok mélyén: Az adatbányászat mindennapjai. Budapest, Alinea IQSYS, 2010.

Fajsi Bulcsú, Cser László: Üzleti tudás az adatok mélyén: Adatbányászat alkalmazói szemmel. Budapest, Budapesti Műszaki és Gazdaságtudományi Egyetem, 2004.

Fischerné Dárdai Ágnes: Szolgáltatásintegráció a Tudásközpontban
http://tmt.omikk.bme.hu/show_news.html?id=5481&issue_id=525
(2014.03.30)

Genoni, Paul, Wright, Janette: Australisa's national research collection overlap uniqueness and distribution. In: Australian Academic & Research Libraries, 2011. (42. évf.) 3. sz.162-178 p.

Groth, Robert: Data mining: Building competitive advantage. Upper Saddle River, Prentice Hall PTR, 2000.

Jiawei, Han, Kamber, Micheline: Adatbányászat: koncepciók és technikák. Bp., Panem, 2004.

Johnson, Doug: A data mining primer and implications for school library media specialist In: Knowledge Quest, 2004. (32. évf.) 5. sz. 32-35 p.

Katona Endre: Bevezetés az informatikába. Budapest, Panem, 2004.

Kerecsendi András: Integrált könyvtári rendszerek. Eger, Eszterházy Károly Főiskola, 2011.

http://www.tankonyvtar.hu/hu/tartalom/tamop425/0005_28_int_konyvtari_rendszerek_pdf/adatok.html (Letöltve: 2014.01.05.)

Király László, Tóth Kornél: HunTéka: egy új integrált rendszer a magyar könyvtári piacon In: Tudományos és Műszaki Tájékoztatás, 2004. (51.évf.) 8. sz.

Horváth Tibor, Papp István (szerk.): Könyvtárosok kézikönyve. 3. kötet. Budapest, Osiris Kiadó, 2003.

Mátyás Melinda: Adatbányászat a könyvtárakban. Budapest, Eötvös Lóránd Tudományegyetem, 2009. <http://mek.oszk.hu/08000/08059/08059.pdf> (Letöltve: 2013.12.15)

Nicholson, Scott, Smith, C.A.: Using lessons from health care to protect the privacy of library users: Guidelines for the de-identification of library data based on HIPAA. In: Journal of the American Society for Information Science and Technology, 2007. (58. évf.) 8. sz. 1198-1206 p.

Nicholson, Scott., Stanton, Jeffrey.: Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.1835&rep=rep1&type=pdf> (Letöltve: 2014.01.14)

Nicholson, Scott: Approaching Librarianship from the Data: Using Bibliomining for Evidence-Based Librarianship. In: Library Hi-Tech, 2006. (24.évf.) 3. sz. 369-375 p.

Nicholson, Scott: Avoiding the Great Data-Wipe of Ought-Three. In: American Libraries, 2003. (34. évf.) 9. sz. 36-39. p.

Nicholson, Scott: Bibliomining for Automated Collection Development In: Journal of the American Society for Information Science and Technology, 2003. (54. évf.) 12. sz. 1081-1090.p.

Nicholson, Scott: Proof is in the Pattern: : Librarians Follow the Corporate Sector toward More Data-Driven Management In: Library Journal, 2006. (131. évf.) 1. sz. 2-6 p.

Nicholson, Scott: The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services In: Information Processing & Management, 2006. (42.évf.) 3. sz. 785-804. p.

Nicholson, Scott: The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. In:Information Technology and Libraries, 2003. (22.évf.) 4. sz 146-151.p.

Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin: Adatbányászat: Alapvetés. Budapest, Panem, 2012.

Pyle, Dorian: Data preparation for data mining. San Francisco, California [etc.], Morgan Kaufmann, 1999.

Sajó Andrea: Az ORACLE LIBRARIES felhasználói szemmel. Budapest, Kossuth Lajos Tudományegyetem Természettudományi Kar, 1998. <http://mek.oszk.hu/03100/03107/html/> (2014.01.13)

Somfai Zoltán: Az adatbányászat – az információbróker elemző eszköze In: Mikulás Gábor (szerk.): Információból üzleti érték: az információbróker környezete és munkája. Budapest, Magyar Információbrókerek Egyesülete, 2006.

Sramó András: Adatbányászat és statisztika . In: Statisztikai szemle, 1999. (77. évf.) 5. sz. 350. p.

Tosaka, Yuji, Weng, Cathy: Reexamining Content-Enriched Access: Its Effect on Usage and Discovery. In: College & Research Libraries, 2011. (72. évf.) 5. sz. 412-427 p.

Tóth Erzsébet: Mire jó az adatbányászat? In: Tudományos és Műszaki Tájékoztatás, 2002. (48. évf.), 3. sz. 527-532.p.

Ullman, Jeffrey David: Adatbázisrendszerek: Alapvetés. Bp., Panem, 2009.

Will, Nicole: Data-mining: Improvement of university library services. In: Technological Forecasting and Social Change, 2006. (73. évf.) 8. sz. 1045–1050 p.

Wormell, Irene: Az informetria és a bibliográfiai adatok használata a stratégiai tervezésben. In: Mikulás Gábor (szerk.): Információból üzleti érték: az információbróker környezete és munkája. Budapest, Magyar Információbrókerek Egyesülete, 2006.

7. Mellékletek

Integrált könyvtári rendszerek statisztikai moduljának elemzése

	Gyűjteményszervezés	Menedzsment,döntéstámogatás			Tájékoztatás	Felhasználói szokások megértése		
	Dokumentumállományi statisztikák	Költségvetés Pénzforgalmi keret	Automatizált Megrendeléstámogatás	Döntéselőkészítés Munkafolyamat- orientáció		Beiratkozások/kölcsönzések/előjegyzések statisztikái	Felhasználói csoportok létrehozása	Témakövetés beállítása, elemzése
Corvina	X	X				X		
Szirén	X	X				X	X	
Aleph	X + forgási sebesség	X	X			X	X	X
Primo, MetaLib	X +adatok továbbbontása					X	X	
Alma	+ forgási sebesség	X	X	X	X	X	X	X
HunTéka	X	X				X	X	
LinLib	X	X				X		
Szikla	X	X				X		
Olib	X	X		X		X		X