Adatbányászatra irányuló törekvések könyvtári területen

Tóth Erzsébet

Mire jó az adatbányászat?

Napjaink információrobbanásának köszönhetően egyre nagyobb gondot jelent a minőségi, rejtett információk kinyerése a nagyméretű adatbázisokból. A korábbi, hagyományos adatbáziskezelési módszerek nem alkalmasak ennek a feladatnak a teljesítésére, hiszen az SQL (Standard Query Language) vagy más, hasonló lekérdező nyelv az adatok listaszerű felsorolását eredményezi és további érdemi információt nem szolgáltat az adatbázisban található adatokról. Az adatbányászat kiszűri a hasznos információkat az adatbázisban található adatelemek változásából. Az ilyen szintű elemzés már túllép az SQL-lel végrehajtott felszínes lekérdezéseken és a gépi tanulás irányába mutat.[1] Az adatbányászatban alkalmazott algoritmusok optimális klasztereket és érdekes szabályszerűségeket fedeznek fel az adatbázisokban. Ezen kívül képesek az adatbázis érdekes részeit közelebbről is megvizsgálni.[2] Az adatbázisban található információk jobb értelmezését és a jövőbeli előrejelzések elkészítését biztosítják.[1]

Az adatbányászati eszközökben számos kiaknázatlan lehetőség rejlik, ezek közül néhányat könyvtári területen is hasznosítanak. Ha egy könyvtár több nagyméretű vagy szakterületi adatbázissal rendelkezik, akkor a könyvtár vezetőjének érdemes elgondolkodnia azon, hogy könyvtára befektessen-e egy korszerű döntéstámogató rendszerbe avagy sem. Az adatbányászati technikák előnyösebbek olyan könyvtárak számára, amelyek a hangsúlyt főként az adatbázisok közvetlen elérésére helyezik, nem pedig a nyomtatott dokumentumok szolgáltatására. A teljes szövegű adatbázisok az on-line kataló-

gushoz képest jobban megfelelnek az adatbányászati technológiák követelményeinek, hiszen ez utóbbi frissítése meglehetősen nehézkes és költséges.[3] Mivel a könyvtári adatbázisok mérete folyamatosan növekszik és tartalmuk változik, ezért szinte lehetetlen manuálisan megtalálni bennük a jövőbeli fejlődési irányokat és a változó tendenciákat (mintákat). Nem hagyhatjuk figyelmen kívül azt a tényt sem, hogy ezeknek az előrejelzéseknek a felhasználása milyen változást eredményezne a gyűjteményfejlesztés támogatásában vagy egy naprakész, teljesen új információs szolgáltatás működtetésében.[1]

Az adatbányászat előnyei és hátrányai

Az adatbányászat két előnyös lehetőséget biztosít a könyvtárak számára, amelyek a következők:

- Gyorsabb és alaposabb dokumentum hozzáférés megvalósítása a hagyományos katalógushoz képest.
- A keresett információ könnyen megtalálható a gyűjteményben anélkül, hogy a felhasználók külön segítséget kérnének a könyvtárostól.

Az adatbányászat *hátrányait* is érdemes számba venni, amelyek az alábbiak:

- Nem alkalmaznak szabványokat az adatbányászati eszközök tárolásával és lekérdezésével kapcsolatban. Kezdetben ez a hiányosság azt eredményezte, hogy problémás volt azoknak a programoknak a megvalósítása, amelyek a könyvtárak közötti rekordcserére és információ megosztásra épültek. Azonban a MARC formátum széles körű használata mára már leegyszerűsítette a rekordcserét és megkönnyítette a könyvtárközi kölcsönzést. Az adatbányászati technológiákba komoly pénzeket befektető könyvtárak jobban ki vannak téve annak a veszélynek, hogy költséges és nehézkes adatkonverziót végeztessenek el velük a szoftver forgalmazói. Ráadásul komoly adatveszteségeik is lehetnek, amennyiben a szoftverszolgáltató további technikai támogatást nem nyújt termékéhez.
- Jelenleg még nem tesztelték az adatbányászati technikák sikeres alkalmazását könyvtári területen. Ezek a technikák sikeresnek bizonyultak olyan üzleti és tudományos területeken, ahol jól strukturált, statisztikai adatokat tartalmazó, rövid dokumentumokat használnak. Ezzel szemben a könyvtárak nagyobb méretű, főként strukturálatlan szövegekkel dolgoznak, amelyek különféle forrásokból származhatnak. Számos „szöveg bányászatra" specializálódó eszköz minimálisan strukturált szöveges dokumentumokat tesz elérhetővé, viszont az általuk nyújtott információmennyiség rendkívül csekély a nagy könyvtárakban található információ mennyiségéhez képest. Tehát a rövidebb terjedelmű dokumentumokra alkalmazott indexelési és keresési mechanizmusok nem alkalmasak a könyvtárakban található nagyobb méretű anyagok és multimédiás dokumentumok hatékony visszakeresésére.
- A technikai akadályok leküzdése továbbra is gondot okoz. Ha az adatok jelentése nincs pontosan meghatározva az adatbányászati eszközök számára, akkor azok nem képesek az információs objektumok közötti relációk felismerésére. Gyakorlatilag nem lehetséges az összes feldolgozási technikát alkalmazni egy lefuttatott keresés dokumentumaira, ez kizárólag a kisebb méretű adatcsoportok esetében oldható meg. A felhasználók keresőkérdéseik megfogalmazásával fontos információkat árulhatnak el önmagukról és keresési célkitűzéseikről. Ezeket a hasznos információkat az adatbányászati eszközök működésük közben felhasználják. Ennek ellenére nehéz meghatározni azokat a technikákat és módszereket, amelyek képesek optimálisan kielégíteni a felhasználói igényeket.

Mindezeket figyelembe véve a könyvtáraknak óvakodniuk kell attól, hogy olyan adatbányászati tech-

nikát válasszanak maguknak, amelytől hosszabb távon nem várható el a dokumentumok közvetlen elérése.[3]

Az adatbányászat lehetséges könyvtári alkalmazási területei

Döntéstámogató rendszerekben az adatbányászatot a felhasználói szokások jobb megismerésére használják. Ha a gépi rendszer egy egyedi azonosítót ad felhasználóinknak a távoli adatbázisok vagy bizonyos adatbázisok eléréséhez, akkor nyomon tudjuk követni könyvtári erőforrásaink és szolgáltatásaink kihasználtságát.

Rendelkezésre állnak olyan háromdimenziós megjelenítő eszközök is, amelyekkel az adatbázisból vett előrejelzések minden aspektusból megtekinthetők. Miután a várható tendenciákat (mintákat) a korábbi felhasználói szokások alapján megjósoltuk, a jövőbeli felhasználói igények még sikeresebben kielégíthetők.

Nemcsak a gyűjteményfejlesztés finomítható további speciális igények meghatározásával, hanem olyan könyvtári programok is tervezhetők, amelyek ismert felhasználói szokásokra és összekapcsolódó elemekre épülnek. Egy közművelődési könyvtárban megjósolható például az is, hogy mely könyvek iránt mutatkozik a legnagyobb érdeklődés, valamint a korábbi felhasználói szokások alapján megállapítható előre az is, hogy hányan jegyzik elő a könyveket és mennyi ideig.[1]

Az adatbányászati algoritmusok felhasználási lehetőségei

Az adatbányászattal kapcsolatos algoritmusok az alábbi feladatok elvégzésére alkalmasak könyvtári környezetben:

1. *Osztályozás és klaszterálás*: az adatok közötti további relációk feltárására jól alkalmazható mindkét módszer. Klaszterálásnál különböző osztályok határozhatók meg az adatok természetes csoportosítása alapján.
2. *Hivatkozáselemzés*: általában a színvonalasabb, igényesebb dokumentumokra gyakrabban hivatkoznak, mint a kevésbé színvonalas írásokra. A hivatkozások mindig elárulnak valamit a dokumentum tartalmáról. A hivatkozáselemzés a gyakrabban hivatkozott dokumentumokat a lista elejére helyezi, illetve felismeri azokat a dokumentumokat, amelyek más dokumentumokhoz kapcsolódnak.
3. *Szekvencia elemzés:* statisztikai elemzéseken keresztül megtalálja azokat az egymással össze nem kapcsolódó dokumentumokat, amelyeket a felhasználók valószínűleg együtt szeretnének elolvasni. Felismeri azokat az útvonalakat, amelyeket a felhasználók információkeresés közben követnek.
4. *Automatikus tartalmi kivonat készítés*: a különféle automatikus módszerekkel előállított tartalmi kivonatok általában nem tekinthetők olyan mértékben igényesnek, mint az emberi munkaerővel előállított összefoglalások. Azonban ez az automatikus megoldás segíti a felhasználót annak eldöntésében, hogy milyen dokumentumra van szüksége. Egy tartalmi kivonatok előállítására alkalmas szoftver sokféle módszert alkalmazhat működésében, mint például: fontosabb szavak, kifejezések automatikus felismerése a szövegben elfoglalt hely alapján, a program képes megkeresni, hogy egyes kifejezésekkel milyen szavak, kifejezések fordulnak elő szorosan összekapcsolódva, különböző szintaktikai és nyelvtani elemzések végzése, stb.[3]

Adatbányászatra irányuló fejlesztések

A Floridai Egyetem könyvtárai olyan kutatási programban vettek részt, amely egy döntéstámogató rendszer kifejlesztésére irányult. A rendszer elsődlegesen a könyvtárvezetőket segítette a döntéshozatalban. A program első részében a különböző adatbázisokból származó adatokat egy relációs adatbáziskezelő rendszerbe (Microsoft Access) importál-

ták, ahol az új, létrejövő adatbázis adattárházként működött. Általában adattárháznak tekinthető olyan nagyméretű adatbázis, amelyet már meglévő, főként működési adatokat tartalmazó adatbázisokból hoznak létre kizárólag döntéshozatal céljából. Az adattárház rugalmas lekérdezésére grafikus felhasználói felületet fejlesztettek ki, ahol előre meghatározott vagy „ad hoc" SQL lekérdezések összeállítására volt lehetőség. A program második részében különböző adatbányászati technikák használatát tervezik a létrehozott adattárházon. Egy olyan neurális hálózati technológiára épülő adatbányászati eszközt vizsgálnak, amely képes asszociációs szabályokat felfedezni az adattárházban található adatok között.[4]

A Kansas Állambeli Egyetemi Könyvtárak egy olyan adattárházat fejlesztettek ki, amely a gyűjtemény használatát méri egy bizonyos időintervallumon belül és különböző felhasználói szokások szerint. Az adattárház hatékony gyűjteményfejlesztést biztosít a tartalmi átfedések megszüntetésével és a felhasználói igények teljes körű figyelembe vételével. Az adattárházat ORACLE relációs adatbáziskezelővel valósították meg, ahol az adatok lekérdezésére egy Java-ban kifejlesztett programot használtak. Az adattárházban található működési adatokat három különböző helyről gyűjtötték össze: az on-line katalógusból, a pénzügyi adatbázisból és a dokumentum-szolgáltatásból. A programban külön adatelemzésre alkalmas szoftvereszközöket is alkalmaztak.[5]

A kanadai CINDI (Concordia INdexing and Discovery System) rendszer működésébe szakértői rendszer technológiáját integrálták, amely kétféle feladatkört lát el: egyrészt „feldolgozó könyvtárosként" támogatja a HTML dokumentum szerzőjét a megfelelő tárgyszavak kiválasztásában és meghatározásában. Másrészt „tájékoztató könyvtárosként" segíti a felhasználót az elképzeléséhez legközelebb álló tárgyszavak felkutatásában. A HTML dokumentumról készül egy szemantikai fejléc, amely a dokumentum legfontosabb metaadatait tartalmazza, és leírja annak szemantikai tartalmát és struktúráját. A rendszerben lefuttatott lekérdezés elsősorban a szemantikai fejlécekre irányul, amelyek egy különálló osztott adatbázisba kerülnek a regisztráció után. Amennyiben a keresés konkrét találatokat eredményez, a rendszer azonnal megkeresi a szemantikai fejlécekhez kapcsolódó dokumentumokat is.[6]

Adattárházak és a „Web farming" technika

Az üzleti vállalatok adattárházainak kialakítását olyan igény hívta életre, amely fontosnak tekintette az üzleti növekedéshez szükséges információk kinyerését a vállalati szinten felhalmozott működési adatokból. Az adattárházak a vállalati tranzakciós adatok rendszerezését és integrálását foglalják magukba, valamint a különböző adatformák egységesítését. Testre szabott üzleti szolgáltatások kialakítását is megkövetelik. Óriási méretű adatbázisokról van itt szó, ezért a hasznos információk kinyerését ezekből az adattárházakból a mesterséges intelligencia technikák segítségével oldják meg.

A „Web farming" technika értelmében az üzleti szervezeten kívüli információk ugyanolyan értékesek lehetnek a stratégiai döntések meghozatalában, mint az azon belüliek. Tehát a „Web farming" technika megnöveli a korábban kialakított adattárházat és értékes információkat szolgáltat az adattárházon kívül eső területeken is.[7] *Hackathorn* szerint a „Web farming" technika céljai a következők:

- az üzleti élettel kapcsolatos hálózati információk feltárása és begyűjtése;
- az összegyűjtött információk átalakítása egy olyan formátummá, amely kompatibilis az adattárházzal;
- ezeknek az információknak a megfelelő célcsoportokhoz történő eljuttatása, amely közvetlenül és pozitívan befolyásolja az üzleti folyamatokat;

- az előző lépesek megfelelő sorrendben történő végrehajtása;[8]
- a „Web farming" technika mindig az adattárház adatkörnyezetén belül használatos. Célja egy olyan rendszer létrehozása, amely a weblekérdezések automatikus és szisztematikus feldolgozására alkalmas.

A „Web farming" technika bevezetése egy több lépésből álló folyamatot igényel, amelynek négy fő szintjét különíthetjük el:

1. szint: jól képzett üzleti szakértőt követel meg, aki értékeli és rangsorolja a külső információs forrásokat aszerint, hogy azok mennyire relevánsak az üzleti élet szempontjából.

2. szint: komoly elkötelezettséget jelent információkezelésnél, miközben a folyamatot egy biztonságos szerver környezetben indítják el. Ezen a szinten a szűrés és a frissítés automatizálható.

3. szint: az adatbázist egy olyan teljes feladatkörrel rendelkező intranetes webhellyé alakítják át, amely információs központként működik az üzleti társaság számára.

4. szint: az adattárházat metaadatokkal látják el és azok attribútumait az adattárházban található attribútumokkal kapcsolják össze.

Általában a szakkönyvtáros az a személy, aki felelős a folyamat sikeres lebonyolításáért. Munka közben lehetősége nyílik arra is, hogy másokkal megismertesse a gyűjtemény jelentősebb információs forrásait és szolgáltatásait.[7]

Az adatbányászat megvalósításával kapcsolatos technikai kérdések

Fontos adatmigrációs és adatfrissítési kérdéseket kell megoldani egy könyvtárban, amikor egy régi rendszerről egy új, kifinomultabb rendszerre térnek át. Nem csupán az adatbázisban található adatmezőket kell figyelembe venni, hanem olyan programokat kell használni, amelyek alkalmasak a metaadatok kinyerésére. A metaadatok tájékoztatnak minket az adatbázis mezőiről, az adatok begyűjtéséről, definiálásáról, attribútumairól, stb. Az adatkonzisztencia rendkívül fontos követelmény a rendszerben.[1] A MARC formátum elterjedt használata nagymértékben csökkentette az adatmigrációs problémákat.[3]

A jövőben fokozott igény mutatkozik egy olyan széles körben elfogadott adatcsere formátum iránt, amely támogatja az adatgyűjtést és az elemzést. Az XML szabvány tűnik a legmegfelelőbb megoldásnak, mivel biztosítja a különböző adattárak rugalmas kezelését, valamint a strukturált és a strukturálatlan adatok összekapcsolását. Hosszabb távon az igazi megoldást az jelentené, hogyha a könyvtárak, a különböző érdekcsoportok, a tartalomszolgáltatók és a kiadók egy konzorcium keretében megállapodnának egy közös adatcsere formátum használatában. Elfogadható megoldás lenne az is, ha az XML (eXtensible Markup Language) szabványnak egy olyan speciális változatát használnák, amely a MARC formátumot integrálja egy speciális könyvtári környezetre szabott ún. dokumentumtípus definícióval (Document Type Definition).

Könyvtári környezetben az adatok többnyire változatos, egymással nem kompatibilis rendszerekből származnak. Az adatok sokfélesége problémát jelent egy egységes adatbázis struktúra kialakításánál. A könyvtáraknak alkalmazásokhoz nem kötődő adatokat kell létrehozniuk, ami kétféle stratégiával valósítható meg:

1. Bizonyos időközönként a különböző alapkörnyezetekből (platform) adatokat nyernek ki, amelyeket egy relációs adatbáziskezelő rendszerbe töltenek be. Az adatok először ASCII állományformátumba kerülnek, ezt követően azok egy konverziós eszköz segítségével általános mező csoporttá alakulnak át. A konverziós folyamat eredményeként bizonyos adatmezők eltűnhetnek, vagy azok egyszerűen elveszíthetik relevanciájukat.

2. Teljesen kiküszöbölik az adatok különböző alapkörnyezetekből történő kinyerését és azonnal közös adatszerkezetet használnak. Ez tulajdonképpen azt jelenti, hogy kompatibilis alapkörnyezeteket választanak és konzisztens relációs adatbáziskezelő rendszerrel dolgoznak. Tehát adatintegritásra törekszenek a könyvtári rendszer egészében.[9]

Az adatbányászat jövője

Az adatbányászati technikák és az általuk nyújtott perspektívák rendkívül izgalmasak. Ezek tulajdonképpen forradalmasítják a döntéshozatal jövőbeli megközelítési módját. Annak ellenére, hogy az adatbányászati technológiák teljes mértékben a rendelkezésünkre állnak, azok még, sajnos, gyerekcipőben járnak. Tehát kifinomultabb és költséghatékonyabb megoldásokra van szükség ezen a területen.[1]

Az adatbányászat nagy sikereket ért el eddig az üzleti világban, kizárólag statisztikai jellegű feladatok megoldásában, azonban ez a siker a könyvtári bibliográfiai adatok esetében még várat magára. A könyvtáraknak továbbra is más alternatívákat kell keresniük a hagyományos katalóguson kívül a dokumentumok hatékony elérésének biztosítására. Az adatbányászati technológiák használata mellett szól az az érv is, hogy a nyomtatott dokumentumokra alkalmazott hagyományos visszakeresési és indexelési mechanizmusok nem nyújtanak kielégítő hozzáférést a digitális környezet változatos adattípusaihoz és struktúráihoz. Ráadásul egy könyvtárban túl sok információt dolgoznak fel manuálisan, ezért az adatbányászati eszközök ígéretes megoldásnak tűnnek a könyvtárosok számára.[3]

Irodalomjegyzék

1. SCHULMAN, Sandy: Data mining: life after report generators: libraries use this decision-support technique to chart a future course. In: Information Today, vol. 15. no. 3. March, 1998. p. 52.
2. ADRIAANS, Pieter – ZANTINGE, Dolf: Data mining. Longman, 1996. p. 5.
3. BANERJEE, Kyle: Is data mining right for your library? In: Computers in Libraries vol. 18. no. 10. November/December 1998. p. 28-31.
4. SU, Siew-Phek T. – NEEDAMANGALA, Ashwin: Harvesting information from a library data warehouse In: Information Technology and Libraries vol. 19. no. 1. March, 2000. p. 17-27.
5. COLE, Karen – SOMERS, Michael – EMERY, Jill: Data warehousing: developing a support system prototype. In: Serials Librarian vol. 40. no.3/4. 2001. p. 349-353.
6. DESAI, Bipin C. – SHINGHAL, Rayan – SHAYAN, Nader R. et al: CINDI: A virtual library indexing and discovery system. In: Library Trends vol. 48. no. 1. Summer 1999. p. 209-233.
7. FYE, Eleanor C.: Old MacDonald didn't have IT: Web farming in the info age. In: Information Outlook vol. 2. no. 12. December 1998. p. 42-43.
8. HACKATHORN, Richard D.: Reaping the web for your data warehouse. In: DBMS August 1998. (http://www.dbmsmag.com/9808d14.html)
9. GUENTHER, Kim: Applying data mining principles to library data collection. In: Computers in Libraries vol. 20. no. 4. April 2000. p. 60-63.