https://doi.org/10.18778/7525-920-9.03

dr Grzegorz Czapnik
Katedra Bibliotekoznawstwa
i Informacji Naukowej
Uniwersytet Łódzki

Bibliomining – o zastosowaniu eksploracji danych w badaniach bibliotek

Eksploracja danych

Pojęcie *bibliomining* zaproponowali w 2003 roku dwaj badacze z amerykańskiego Syracuse University School of Information Studies – Scott Nicholson i Jeffrey Stanton, definiując je jako: „zastosowanie technik eksploracji danych do sondowania ogromnych zasobów danych generowanych przez typowe biblioteki zautomatyzowane”[1]. Termin ten miał zastąpić złożone określenie *data mining for libraries* [tłum. dosł.: eksploracja danych dotycząca bibliotek, lub: eksploracja danych bibliotecznych], które sprawiało kłopoty przy poszukiwaniach literatury przedmiotu[2]. Skrócona nazwa wywiedziona została z połączenia pojęć *bibliometrics* [bibliometria] i *data mining* [eksploracja danych][3]. Ostatnie z wymienionych pojęć oznacza „proces odkrywania istotnych zależności (korelacji), wzorców i tendencji poprzez przesiewanie dużych ilości danych przechowywanych w repozytoriach za pomocą technik

[1] Nicholson S., Stanton J.: *Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries.* [W:] *Organizational data mining: Leveraging enterprise data resources for optimal performance*. [dokument elektroniczny]. Idea Group Inc. (2003), Updated on 2/16/04. Dostępny w Internecie: http://www.bibliomining.com/nicholson/odm-com.html [data dostępu: 15.04.2012].

[2] Ze względu na powszechne wykorzystanie w literaturze informatycznej zestawu pojęć: „data mining” oraz „libraries” w odniesieniu do podprogramów (czyli „bibliotek” w znaczeniu informatycznym) stosowanych w procesie eksploracji danych, użycie tych fraz nie jest skuteczne podczas wyszukiwania informacji na temat stosowania eksploracji danych do badań książnic.

[3] Nicholson S.: *The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making*. „Information Technology and Libraries”. [dokument elektroniczny]. (2003) vol. 22(4). Dostępny w Internecie: http://www.ala.org/ala/lita/litapublications/ital/2204nicholson.htm [data dostępu: 15.04.2012].

rozpoznawania wzorców oraz technik statystycznych i matematycznych”[4]. *Data mining* jest czasem traktowany jako synonim metody „odkrywania wiedzy z [baz] danych” [ang. *Knowledge Discovery from Databases*, w skrócie KDD], lub też jedynie jako jej etap, obejmujący przede wszystkim wybór i wykorzystanie odpowiednich algorytmów oraz aplikacji służących do wydobycia z baz reguł, zależności i schematów. Cała procedura KDD jest sekwencją następujących kolejno operacji:

- rozpoznanie dziedziny badanego problemu,
- określenie celów procesu,
- wybór danych z baz danych,
- oczyszczenie danych i ich integracja w hurtowni danych,
- przygotowanie danych, redukcja i transformacja, wybór istotnych atrybutów, redukcja wymiarów i liczby zmiennych,
- wybór funkcji eksploracji, wybór algorytmów (metod),
- eksploracja danych,
- weryfikacja i walidacja wykrytych zależności,
- prezentacja odkrytej wiedzy (wizualizacja uzyskanych wyników),
- wykorzystanie odkrytej wiedzy[5].

Metody eksploracji danych zaczęły się rozwijać w latach 80. XX wieku wraz z powstawaniem pierwszych hurtowni danych[6]. Opierają się przede wszystkim na osiągnięciach statystyki wspartych mocą technologii informatycznej a także narzędziach i technikach, powstałych w efekcie badań nad bazami danych, sztuczną inteligencją i przetwarzaniem informacji. Ich wachlarz obejmuje obecnie kilkadziesiąt algorytmów oraz narzędzi i stale się

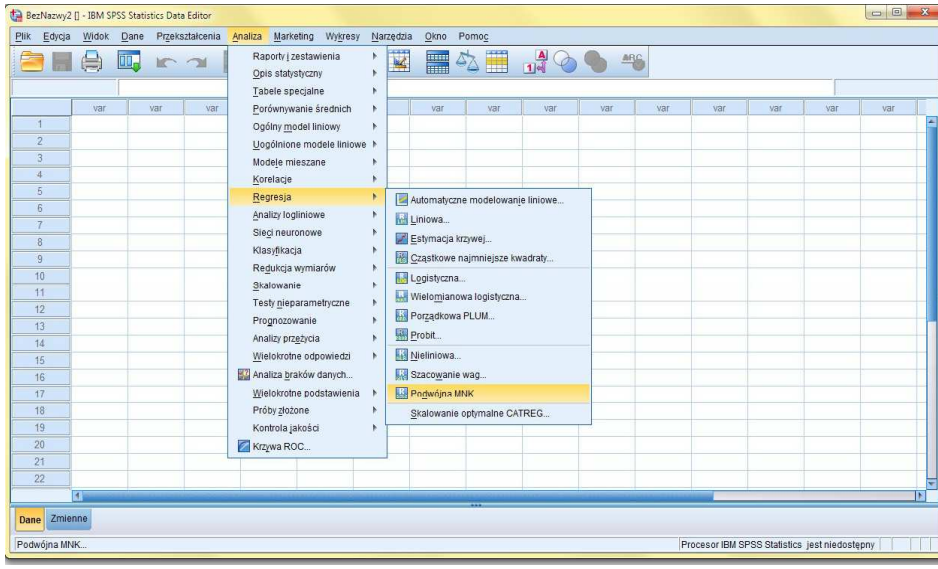
[4] Gurycz J.: *Wprowadzenie do Data Mining*. [dokument elektroniczny], StatSoft Polska. Dostępny w Internecie: http://www.statsoft.pl/czytelnia/dm/wprowdm.html [data dostępu: 11.05.2012]. Definicja ta została zaproponowana przez amerykańskiego potentata przemysłu informacyjnego, firmę Gartner Group Inc.

[5] Wróblewski J. [et al.]: *W[ykład] 8: Wspomaganie zadań eksploracji danych*. [w:] *PJWSTK - Hurtownie danych* [dokument elektroniczny]. 2006, modyfikacja: 2008. Dostępny w Internecie: http://edu.pjwstk.edu.pl/wyklady/hur/scb/index.htm [Data dostępu: 15.04.2012].

[6] Hurtownia danych (ang. data warehouse) – rodzaj bazy danych, która jest zorganizowana i zoptymalizowana pod kątem pewnego wycinka rzeczywistości. Hurtownia danych jest wyższym szczeblem abstrakcji niż zwykła relacyjna baza danych (choć do jej tworzenia używane są także podobne technologie). W skład hurtowni wchodzą zbiory danych zorientowanych tematycznie (np. hurtownia danych klientów). Dane te często pochodzą z wielu źródeł, są one zintegrowane i przeznaczone wyłącznie do odczytu. Źródło: *Hurtownia danych*. [w:] *Wikipedia* [dokument elektroniczny], Data modyfikacji: 04.04.2012. Dostępny w Internecie: http://pl.wikipedia.org/wiki/Hurtownia_danych [data dostępu: 21.04.2012].

powiększa (Ilustracja 1.), nie istnieje przy tym jeden, ustalony kanon ich doboru czy metodyka ich wykorzystania.

Il. 1. Menu algorytmów analitycznych w programie IBM SPSS



Źródło: Oprac. własne.

Jak proponują Hand, Mannila i Smyth metody eksploracji danych można poddać typologizacji, przyjmując funkcjonalne kryterium „rodzaju zadań odpowiadających różnym celom osób analizujących dane”[7]. Zgodnie z tym kryterium wyróżniają:

- eksploracyjną analizę danych – (ang EDA) – eksploracja bez założeń na temat tego czego szukamy. Często stanowi wstęp do dokładniejszej analizy. Wykorzystuje różnorodne techniki wizualizacyjne, które umożliwiają efektywny ogląd zbioru danych oraz pozwalają zauważyć występujące w nim prawidłowości, które potencjalnie mogą być źródłem nowej wiedzy,
- modelowanie opisowe – służy opisaniu zbioru jako całości. Najczęściej obejmuje takie zadania jak: segmentację zbioru, analizę skupień i grupowanie,

[7] Hand D., Mannila H., Smyth P.: *Eksploracja danych*. Warszawa: Wydawnictwa Naukowo-Techniczne, 2005, s. 46.

- modelowanie predykcyjne – na podstawie znanych wartości zmiennych przewidujemy kategorię (klasyfikacja) lub wartość (regresja) nowej zmiennej,
- odkrywanie wzorców i reguł – np. fałszerstw, plagiatów ale również rozgrywek sportowych (koszykówka NBA od 1997 roku),
- wyszukiwanie według zawartości – to przede wszystkim mechanizmy służące wyszukiwaniu treści w cyfrowych zasobach Internetu (tekst – Text Mining, ale też obrazy, filmy, muzyka itp.).

Eksploracja danych jest obecnie narzędziem szeroko stosowanym w biznesie i marketingu między innymi do automatycznego grupowania klientów na podstawie ich zakupów i/lub zachowań czy też do tzw. analizy koszykowej, czyli oceny współwystępowania produktów lub usług, nabywanych najczęściej przez klientów/użytkowników. Korzystają z niej również specjaliści w zakresie ekonomii (np. przy przewidywaniu kursów akcji lub rozpoznawaniu nielegalnych transakcji), medycyny (automatyczna diagnoza, dobór leków w oparciu o dane dotyczące pacjenta i obrazu choroby), sportu (prognozowanie wyników zawodów na podstawie statystyk drużyn) czy lingwistyki (analiza korpusów tekstów, wykrywanie plagiatów).

Bibliomining w literaturze

Techniki eksploracji danych były także stosowane do wydobywania informacji z zasobów danych generowanych przez biblioteki. We współczesnych, zautomatyzowanych książnicach – a takich jest obecnie większość – wszystkie realizowane procesy pozostawiają w systemach informatycznych ślady w postaci zapisanych danych. W trakcie każdej wizyty użytkownika w bibliotece „zanotowany" może zostać fakt odwiedzin, zapamiętane jego zapytania wyszukiwawcze, odnotowane wypożyczenia, zwroty, logowania do baz danych oraz dziesiątki innych operacji wraz z ich sygnaturami czasowymi, o ile tylko używa on zaimplementowanych w książnicy systemów. Rosnące zasoby danych, przechowywane w systemowych bazach stają się wraz z upływem czasu odwzorowaniem zachowań i przyzwyczajeń czytelników. Już na początku lat 90. XX wieku, kiedy koncepcja *data mining* dopiero się kształtowała, w czasopismach związanych z problematyką zarządzania bibliotekami pojawiły się opinie, według których odkrywanie zapisanych w danych wzorców oraz analiza „przypadków użycia" mogłyby w pewnych warunkach stać się kluczem do zrozumienia potrzeb użytkowników, korzystających z usług biblioteki. W 1996 roku ukazał się specjalny zeszyt czasopisma „Library Administration & Management", w całości poświęcony eksploracji danych bibliotecznych. Omówiono w nim między projekty wykorzystujące *data mining*, realizowane w bibliotekach amerykańskich – Management Informa-

tion System zaimplementowany w bibliotece w Dekalb County w stanie Georgia, którego zadaniem było wspieranie decyzji o podziale rozbudowywanego wolnego dostępu na potrzeby poszczególnych wydziałów, w oparciu o automatyczną analizę danych z systemu udostępniania[8] oraz dwa projekty automatyzacji selekcjonowania dokumentów i zarządzania rozbudową kolekcji realizowane w bibliotece Sterling C. Evans Library w Texas A&M University[9]. W tym samym tomie „LA&M" znajduje się również opracowana przez Patricię Cross krótka, obejmująca 25 pozycji bibliografia prac dotyczących zastosowania eksploracji danych i systemów zarządzania informacją biblioteczną[10]. Wśród wymienionych w niej publikacji znalazły się zarówno instrukcje opisujące „krok po kroku" budowę hurtowni i analizę danych, jak też dyskusje o przydatności różnych rodzajów informacji do świadomego podejmowania decyzji. Dwa lata później, w czasopiśmie „Computer In Libraries" Kyle Banerjee, odpowiadając na postawione w tytule pytanie: „Czy eksploracja danych jest dobra dla twojej biblioteki?", przedstawił przebieg procesu eksploracji oraz wskazał na możliwości zastosowania go do zwiększenia dostępności do zbiorów (w wyniku bieżącego indeksowania automatycznego)[11]. Zauważył również, że ze względu na specyfikę tej techniki lepszych wyników można się spodziewać przy eksploracji pełnotekstowych zbiorów dokumentów, niż jedynie katalogów bibliotecznych.

Pomimo że wiele zagadnień związanych z eksploracją danych bibliotecznych było omawianych we wcześniejszych publikacjach, dopiero cykl publikacji Scotta Nicholsona (i jego współpracowników) z lat 2003-2006 przyniósł teoretycznie pogłębioną analizę eksploracji danych dotyczących bibliotek[12]. Przedstawił on między innymi definicję i koncepcję *bibliominingu*, zakreślił obszar jego wykorzystania jako narzędzia wspomagającego zarządzanie procesami bibliotecznymi oraz jako technikę badania tych procesów a także sformułował praktyczne zalecenia dotyczące między innymi doboru danych, algorytmów i narzędzi, przygotowania danych do eksploracji, budowy „hurtowni danych" oraz analizy otrzymanych wyników. Prace Nicholsona są kanonem, do którego odwołuje się większość późniejszej literatury dotyczącej omawianego zagadnienia.

---

[8] Mancini D. D.: *Mining your automated system for systemwide decision making*. „Library Administration & Management", 1996, vol. 10 (1), p. 11-15.

[9] Atkins S.: *Mining automated systems for collection management*. „Library Administration & Management", 1996, 10 (1), p. 16-19.

[10] Cross P.: *Mining your automated system for better management. A brief bibliography*. „Library Administration & Management", 1996, vol. 10 (1), p. 26-27.

[11] Banerjee K.: *Is data mining right for your library?* „Computers in Libraries", 1998, vol. 18(10), p. 28-30.

[12] Wykaz najważniejszych prac S. Nicholsona, poświęconych bibliominingowi, zamieszczono w bibliografii.

Idee zawarte w wymienionych publikacjach są obecnie teoretycznie i praktycznie rozwijane również poza Stanami Zjednoczonymi, szczególnie w Indiach i Japonii a bibliografia przedmiotu, którą udało się zgromadzić autorowi tego referatu obejmuje już ponad 50 tytułów.

Bibliomining jako proces

Według Nicholsona eksploracja danych dotycząca bibliotek to proces obejmujący 6 etapów:

- określenie obszaru (dziedziny) analizy,
- identyfikację wewnętrznych i zewnętrznych źródeł danych,
- zgromadzenie, oczyszczenie i anonimizację danych oraz przygotowanie hurtowni danych,
- wybór odpowiednich narzędzi i technik eksploracyjnych,
- odkrywanie wzorców w danych za pomocą DM oraz przygotowanie wyników do analizy (wizualizacja, raportowanie),
- analiza wyników i implementacja rezultatów[13].

Proces *bibliominingu* jest często – podobnie jak procedura KDD – postępowaniem iteracyjnym i adaptacyjnym[14]. Wyniki analizy mogą wskazywać kierunek dalszych poszukiwań, a implementacja rezultatów nierzadko oznacza ponowne przeprowadzenie procesu eksploracji dla ściślej wyznaczonego obszaru analizy lub z uwzględnieniem lepiej dobranych danych i algorytmów ich przetwarzania.

Wyznaczenie obszaru i celu eksploracji jest kluczowym etapem bibliominingu, gdyż decyduje o trybie postępowania w kolejnych fazach procesu. Możliwe są dwa podejścia do tego zagadnienia – pierwsze, określane jako eksploracja ukierunkowana, w której celem analizy jest poszukiwanie wiedzy, pozwalającej na rozwiązanie konkretnego, określonego problemu i drugie – nieukierunkowane, w którym poszukiwane są wszelkie prawidłowości ukryte w danych. W tym ostatnim przypadku celem jest najczęściej pogłębienie wiedzy o funkcjonowaniu biblioteki lub sformułowanie problemu (problemów), który w kolejnych iteracjach będzie podstawą eksploracji ukierunkowanej. Warto podkreślić, że zarówno Nicholson, jak również autorzy podręczników o eksploracji danych polecają stosowanie podejścia ukierunkowanego, ponieważ jest ono obarczone mniejszym ryzykiem popełnienia błędów oraz pozwala na łatwiejsze wykrycie istotnych zależności i wzorców w analizowanym zasobie.

[13] Nicholson S.: *The Bibliomining Process…*, s. 1.
[14] Por. Larose D.: *Odkrywanie wiedzy…*, s. 5.

Drugi etap bibliominingu polega na identyfikacji dostępnych danych, które mogą potencjalnie wnieść wiedzę o analizowanym zagadnieniu. Dane mogą pochodzić ze źródeł wewnętrznych, czyli systemów informatycznych wykorzystywanych w bibliotece, jak zintegrowany system biblioteczny, bazy danych, biblioteki cyfrowe, repozytoria lub też ze źródeł zewnętrznych – na przykład systemów demograficznych (PESEL, NIP) czy systemów kodów pocztowych. Jeżeli wcześniej zdecydowaliśmy się na eksplorację ukierunkowaną, natura analizowanego problemu dyktuje wybór zasobów danych (poszukując modelu optymalnego wykorzystania miejsc w czytelni bibliotecznej prawdopodobnie nie będziemy sięgali po dane adresowe wydawnictw i hurtowni, w których biblioteka się zaopatruje).

Wyodrębnione zasoby danych w swej oryginalnej postaci i lokalizacji nie nadają się do eksploracji, są natomiast źródłem z którego wyprowadza się dane transakcyjne do hurtowni danych. „Ładowanie" hurtowni wiąże się z praco- i czasochłonnym oczyszczeniem danych, czyli wstępną identyfikacją i ewentualną korektą błędnych wartości, uzupełnieniem lub usunięciem niepełnych rekordów, podzieleniem lub połączeniem wybranych kategorii a także – ze względu na ochronę danych wrażliwych – depersonalizacją i anonimizacją danych przygotowanych do przetwarzania. Obowiązek anonimizacji danych wynikający z przepisów dotyczących ochrony danych osobowych[15] może być czynnikiem utrudniającym lub nawet uniemożliwiającym przeprowadzenie badań eksploracyjnych. *Bibliomining* jest metodą bazującą na ogromnych zbiorach danych transakcyjnych, dlatego spełnienie niektórych wymagań zawartych w tych przepisach, jak na przykład ustawowego warunku bezpośredniego informowania użytkowników o fakcie przetwarzania ich danych osobowych jest po prostu nierealne. Polski Ustawodawca zezwala na pominięcie tego obowiązku jeżeli: „dane te są niezbędne do badań naukowych, dydaktycznych, historycznych, statystycznych lub badania opinii publicznej, ich przetwarzanie nie narusza praw lub wolności osoby, której dane dotyczą, a spełnienie wymagań określonych w ust. 1 wymagałoby nadmiernych nakładów lub zagrażałoby realizacji celu badania"[16], jeżeli jednak biblioteka usunie dane transakcyjne niezwłocznie po ich wykorzystaniu[17], to badacz nie będzie miał materiału do eksploracji.

Ilustracją trudności związanych z przygotowaniem „hurtowni danych" może być niewielki eksperyment badawczy przeprowadzony przez autora niniejszego artykułu. Sprawdzając działanie wybranych technik i narzędzi eksploracyjnych zastosowano kilka wybranych algorytmów do ekstrakcji in-

[15] Podstawę prawną stanowi *Ustawa z dnia 29 sierpnia 1997 r. o ochronie danych osobowych*. (Dz.U. 1997 nr 133 poz. 883 z późn. zmianami).

[16] *Ustawa z dnia 29 sierpnia 1997 r. o ochronie danych osobowych*. (Dz.U. 1997 nr 133 poz. 883 z późn. zmianami), Art. 25., ust. 2., pkt. 3.

[17] *Tamże*, Art. 2., ust. 3

formacji z danych zaczerpniętych z baz systemu Horizon Biblioteki Uniwersyteckiej w Łodzi[18]. Eksploracja miała odkryć, czy dane dotyczące czasu wypożyczenia książek mają odniesienie do standardowej typologii użytkowników stosowanej w bibliotece akademickiej. Dane zostały pobrane z systemu bibliotecznego w kwietniu 2012 roku i obejmowały zbiór ponad 65 tys. rekordów zarejestrowanych operacji udostępniania, których data zwrotu mieściła się między 1 października 2011 a 31 marca 2012 roku. Fragment jednego z arkuszy danych w postaci źródłowej przedstawiono na Ilustracji 2. Wśród pobranych rekordów zidentyfikowano ponad 1500 błędów w datach wypożyczenia lub zwrotu (najczęściej data zwrotu poprzedzała datę wypożyczenia), dwie spośród siedmiu kategorii danych wymagały rozłożenia na „atomowe" podkategorie a jedna nie mogła zostać wykorzystana ze względu na niekonsekwentny format stosowanych danych. Sprawdzanie i oczyszczanie pobranego zasobu było najbardziej pracochłonnym etapem eksperymentu, pomimo że nie zrealizowano wszystkich zadań standardowo wykonywanych w tej fazie *bibliominingu* – ponieważ dane otrzymane z Biblioteki od razu były pozbawione atrybutów pozwalających na identyfikację użytkowników, nie była potrzebna ich anonimizacja (Ilustracja 3). Nie przeprowadzono

Il. 2. Dane otrzymane z Biblioteki UŁ w postaci źródłowej (fragment)

A4 | Studenci WES

	B	C	D	E	F
1	wydział	tytul	sygnatura	data wypoż	data zwrotu
2	Wydz. Ekonomiczno-Socjologiczny	Matematyka i jej zastosowanie w naukach ekonomicznych Janusz Piszczala Akademia Ekonomiczna w Poznaniu		07-10-11	08-12-11
3	Wydz. Ekonomiczno-Socjologiczny	Analiza matematyczna w zadaniach Cz 2 Wlodzimierz Krysicki Lech Wlodarski	ZBWES W.52809/2	07-10-11	08-12-11
4	Wydz. Ekonomiczno-Socjologiczny	Socjologia dla ekonomistow Danuta WalczakDuraj	ZBWES W.60576	07-10-11	08-12-11
5	Wydz. Ekonomiczno-Socjologiczny	Podstawy finansow Dorota Korenik Stanislaw Korenik	ZBWES W.59164	07-10-11	08-12-11
6	Wydz. Ekonomiczno-Socjologiczny	Analiza matematyczna w zadaniach Cz 1 Wlodzimierz Krysicki Lech Wlodarski	ZBWES W.50095/1	07-10-11	08-12-11
7	Wydz. Ekonomiczno-Socjologiczny	Rynek opcji pomocnicze materialy dydaktyczne Stefan Mynarski	820304	04-03-11	01-10-11
8	Wydz. Ekonomiczno-Socjologiczny	Rynek opcji walutowych w Polsce spekulacja walutowa zarzadzanie portfelem opcji	872749	04-03-11	01-10-11
9	Wydz. Ekonomiczno-Socjologiczny	Warranty subskrypcyjne Wiktor CzeszejkoSochacki Cezary Nowosad Adam Wisnie	1104815	04-03-11	01-10-11
10	Wydz. Ekonomiczno-Socjologiczny	Wybrane zagadnienia teorii opcji podrecznik dla studentow wyzszych szkol technicz	1115657	04-03-11	01-10-11
11	Wydz. Ekonomiczno-Socjologiczny	Opcje na polskim rynku finansowym wycena strategie Grzegorz Golec	891377	09-03-11	01-10-11
12	Wydz. Ekonomiczno-Socjologiczny	Modele kontraktow opcyjnych Ewa Dziawgo Uniwersytet Mikolaja Kopernika	896234	09-03-11	01-10-11
13	Wydz. Ekonomiczno-Socjologiczny	Opcje na akcje Andrzej Fierla	954311	09-03-11	01-10-11
14	Wydz. Ekonomiczno-Socjologiczny	Rynek opcji pomocnicze materialy dydaktyczne Stefan Mynarski	812334	28-10-11	26-01-12
15	Wydz. Ekonomiczno-Socjologiczny	Opcje na polskim rynku finansowym wycena strategie Grzegorz Golec	891377	28-10-11	26-01-12
16	Wydz. Ekonomiczno-Socjologiczny	Modele kontraktow opcyjnych Ewa Dziawgo Uniwersytet Mikolaja Kopernika	932281	28-10-11	27-12-11
17	Wydz. Ekonomiczno-Socjologiczny	Opcje na akcje Andrzej Fierla	938950	28-10-11	26-01-12
18	Wydz. Ekonomiczno-Socjologiczny	Wybrane zagadnienia teorii opcji podrecznik dla studentow wyzszych szkol technicz	1115657	28-10-11	26-01-12
19	Wydz. Zarzdzania	Bankowa analiza przedsiebiorstwa na potrzeby oceny ryzyka kredytowego Stanislav	HG3751.5.R96 2002E1	14-01-10	10-03-10
20	Wydz. Zarzdzania	Wspolpraca przedsiebiorstwa z bankiem Jacek Grzywacz	HG1607.P59G79 2003E2	14-01-10	10-03-10
21	Wydz. Zarzdzania	Podstawy zarzadzania finansami banku komercyjnego pod red Ewy BogackiejKisiel	812635	15-01-10	10-03-10
22	Wydz. Zarzdzania	Bankowosc podrecznik akademicki praca pod red Wladyslawa L Jaworskiego i Zofii	HG1607.P59B377 2002E2	28-01-10	10-03-10
23	Wydz. Zarzdzania	Bankowosc Zbigniew Dobosiewicz	HG3137.D624 2003E2	28-01-10	10-03-10

Źródło: Biblioteka Uniwersytetu Łódzkiego, oprac. własne.

[18] Szerszy opis eksperymentu jest przedmiotem innego opracowania.

również standaryzacji danych, co w efekcie spowodowało zafałszowanie wyników eksploracji. Pomimo tych niedociągnięć eksperyment wykazał, że *bibliomining* może być źródłem nowej wiedzy o bibliotece oraz że jakość tej wiedzy zależy w bardzo dużym stopniu od starannego przygotowania danych do przetwarzania.

Il. 3. Dane oczyszczone i przygotowane do eksploracji (fragment)

A2 fx Pracownicy

	A	B	C	D	E
1	Typ czytelnik	Uczelnia	Jednostka	Dodatkowe kryterium	Czas wypożyczeni
410	Pracownicy	UŁ	Brak danych	brak dod. kryteriów	322
411	Pracownicy	UŁ	Brak danych	brak dod. kryteriów	322
412	Pracownicy	UŁ	Brak danych	brak dod. kryteriów	322
413	Pracownicy	UŁ	BUŁ i BZ	brak dod. kryteriów	322
414	Pracownicy	UŁ	BUŁ i BZ	brak dod. kryteriów	322
415	Pracownicy	UŁ	BUŁ i BZ	brak dod. kryteriów	321
416	Pracownicy	UŁ	BUŁ i BZ	brak dod. kryteriów	321
417	Pracownicy	UŁ	BUŁ i BZ	emeryci	321
418	Pracownicy	UŁ	BUŁ i BZ	emeryci	321
419	Pracownicy	UŁ	BUŁ i BZ	brak dod. kryteriów	320
420	Pracownicy	UŁ	BUŁ i BZ	emeryci	320
421	Pracownicy	UŁ	Brak danych	brak dod. kryteriów	319
422	Pracownicy	UŁ	Brak danych	brak dod. kryteriów	319
423	Pracownicy	UŁ	BUŁ i BZ	brak dod. kryteriów	319
424	Pracownicy	UŁ	BUŁ i BZ	brak dod. kryteriów	318
425	Pracownicy	UŁ	Brak danych	brak dod. kryteriów	318
426	Pracownicy	UŁ	WNoW	brak dod. kryteriów	317
427	Pracownicy	UŁ	WZ	brak dod. kryteriów	315
428	Pracownicy	UŁ	BUŁ i BZ	brak dod. kryteriów	314
429	Pracownicy	UŁ	Brak danych	brak dod. kryteriów	313
430	Pracownicy	UŁ	WES	brak dod. kryteriów	311
431	Pracownicy	UŁ	WNoW	brak dod. kryteriów	309
432	Pracownicy	UŁ	WNoW	brak dod. kryteriów	309

Źródło: opracowanie własne.

Odpowiednie przygotowanie hurtowni danych umożliwia realizację kolejnych etapów eksploracji. Biorąc pod uwagę typ zgromadzonych danych badacz musi wybrać odpowiednie oprogramowanie oraz algorytmy eksploracyjne, których działanie pozwoli na osiągnięcie zakładanego celu. Najczęściej stosowane są systemy statystyczne wyposażone w moduły przeznaczone do DM, jak komercyjne: SAS, SPSS czy Statistica lub otwarty system WEKA. Zaletą programów płatnych jest bogate wyposażenie w algorytmy eksploracyjne i narzędzia wizualizacyjne oraz (w niektórych systemach) polskojęzyczny interfejs, wadą – niestety – bardzo wysoka cena. Dobór metodyki obejmuje między innymi różne rodzaje regresji, algorytmy klasyfikacyjne (np. algorytm

k-najbliższych sąsiadów) i grupowania, drzewa decyzyjne, sieci neuronowe, sieci Kohonena, reguły asocjacyjne i wiele innych[19].

W badaniach realizowanych według opisu Nicholsona wykonanie wybranego algorytmu (algorytmów) jest procedurą zautomatyzowaną wymagającą od badacza najczęściej jedynie ustawienia odpowiednich parametrów. Ze względu na złożoność obliczeniową niektórych metod i ogromne zbiory danych do przetworzenia etap ten może wymagać zastosowania komputerów o dużej mocy obliczeniowej, nie jest to jednak regułą. W „łódzkim" eksperymencie testowano na danych działanie algorytmów: analizy statystycznej, eksploracyjnej analizy danych, automatycznej kategoryzacji i dwustopniowego grupowania danych. Obliczenia, wykonywane za pomocą „biurkowego" komputera osobistego, wyposażonego w oprogramowanie IBM SPSS trwały od kilku do kilkudziesięciu minut, zaznaczyć jednak trzeba, że przeprowadzony eksperyment dotyczył zaledwie kilku wybranych cech badanego zagadnienia. Badacze, poszukujący wiedzy w zasobach danych często muszą uwzględniać znacznie większą liczbę analizowanych atrybutów (wymiarów) danych, czego konsekwencją jest znaczne zwiększenie złożoności obliczeniowej a czasem także konieczność realizacji tzw. redukcji wymiarów danych.

Efektem przetwarzania eksploracyjnego powinien być model lub wzorzec badanego zjawiska[20], wygenerowany przez oprogramowanie, przedstawiony w formie dogodnej do analizowania przez specjalistę w danej dziedzinie. Konieczność oceny przez człowieka jest podkreślana zarówno przez autorów piszących ogólnie o eksploracji danych, jak też w pracach Nicholsona. Larose wprost zaznacza, że przekonanie o całkowitym automatyzmie procedury DM oraz o samodzielnym rozwiązywaniu problemów przez komputer jest jednym z często spotykanych mitów[21]. Udział badacza jest konieczny i istotny na każdym etapie eksploracji a szczególnie podczas analizy, ewaluacji i implementacji rezultatów badawczych. Z definicji eksploracji danych wynika, że wygenerowany model powinien przynieść nową i nietrywialną wiedzę o badanym zjawisku a nikt nie jest w stanie lepiej ocenić jej jakości od specjalisty w odpowiedniej dziedzinie, czyli – w odniesieniu do bibliominingu – bibliologa.

Zagadnienia omawiane w artykule stanowią *novum* w polskiej literaturze. Termin *bibliomining* został wspomniany dotychczas jedynie w kilku krajowych publikacjach[22], a przydatność technik eksploracji danych do analizy

[19] Metodyki DM są omawiane w podręcznikach z zakresu eksploracji danych (np. w wymienianych wcześniej pracach D. Handa, H. Mannili i P. Smytha lub D. T, Larosea), dlatego autor nie przedstawia ich szczegółowo w tym miejscu.

[20] Por. Hand D., Mannila H., Smyth P.: *Eksploracja*…, s. 44-46.

[21] Larose D.: *Odkrywanie wiedzy*…, s. XX.

[22] Veslava Osińska użyła słowa kluczowego „Bibliomining" w artykule dotyczącym kierunków rozwoju współczesnych sposobów klasyfikacji piśmiennictwa z za-

bibliotek nie została również dogłębnie oceniona w literaturze światowej. Zastosowania DM w innych dziedzinach i doniesienia o „bibliotekarskich" eksperymentach realizowanych w innych krajach wskazują, że *bibliomining* może być znaczącym zagadnieniem badawczym i zarazem obiecującym narzędziem pozyskiwania nowej wiedzy o bibliotekach również w Polsce.

Wybrana literatura

1. Atkins S.: *Mining automated systems for collection management*. „Library Administration & Management", 1996, vol. 10(1), p. 16-19
2. Banerjee K.: *Is data mining right for your library?* „Computer in Libraries", 1998, vol. 18(10), p. 28-31
3. Cross P.: *Mining your automated system for better management. A brief bibliography*. „Library Administration & Management, 1996, vol. 10(1), p. 26-27
4. Gurycz J.: *Wprowadzenie do Data Mining*. [dokument elektroniczny] StatSoft Polska. Dostępny w Internecie: http://www.statsoft.pl/czytelnia/dm/wprowdm.html [Data dostępu: 11.05.2012]
5. Hand D., Mannila H., Smyth P.: *Eksploracja danych*. Warszawa: Wydawnictwa Naukowo-Techniczne, 2005. – ISBN 83-204-3053-4
6. Larose D. T.: *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*. Warszawa: Wydawnictwo Naukowe PWN, 2006. - ISBN 978-83-01-14836-2
7. Mancini D. D.: *Mining your automated system for systemwide decision making*. „Library Administration & Management", 1996, vol. 10 (1), p. 11-15
8. Morzy T., Morzy M., Leśniewska A.: *Wykład 1. Wprowadzenie*. [W:] *Eksploracja danych – studia informatyczne* [dokument elektroniczny]. Data modyfikacji: 10.09.2006. Dostępny w Internecie: http://wazniak.mimuw.edu.pl/index.php?title=Eksploracja_danych [Data dostępu: 15.04.2012]
9. Nicholson S., Stanton J.: *Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries*. [W:] *Organizational data mining: Leveraging enterprise data resources for optimal performance*. [dokument elektroniczny]. Idea Group Inc. (2003), Updated on 2/16/04, p. 247-262. Dostępny w Internecie: http://www.bibliomining.com/nicholson/odmcom.html [Data dostępu: 15.04.2012]
10. Nicholson S.: *The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making*. „Information Technology and Libraries" [dokument elektroniczny] 2003, vol. 22(4), p. 1-10. Dostępny w Internecie: http://www.ala.org/lita/ital/22/4/nicholson [Data dostępu: 15.04.2012]
11. Nicholson S.: *Bibliomining for automated collection development in a digital library setting: Using data mining to discover Web-based scholarly research*

kresu informatyki, a Piotr Ziembicki opisał krótko tą technologię, omawiając perspektywy rozwoju infrastruktury informatycznej bibliotek.

works. „Journal of the American Society for Information Science and Technology", 2003, vol. 54(12), p. 1081-1090
12. Nicholson S.: *Digital Library Archaeology: A Conceptual Framework for Understanding Library Use through Artifact-Based Evaluation*. „The Library Quarterly", 2005, vol. 75(4), p. 496-520
13. Nicholson S.: *A framework for Internet archeology: Discovering use patterns in digital library and Web–based information resources*. „First Monday", 2005, vol. 10(2), p. 1-7
14. Nicholson S.: *Approaching librarianship from the data: Using Bibliomining for evidence-based librarianship*. „Library Hi-Tech", 2006, vol. 24(3), p. 369-375.
15. Nicholson S. *The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services*. „Information Processing and Management", 2006, vol. 42(3), p. 785-804; Proof is in the Pattern. „Library Journal netConnect, Supplement to Library Journal", Winter 2006, p. 2-6
16. Osińska V.: *Współczesna problematyka sposobów klasyfikacji informatyki. Kierunki rozwoju*. [w:] *Opracowanie przedmiotowe dokumentów z zakresu nauk ścisłych: matematyczno-przyrodniczych i technicznych. Język haseł przedmiotowych KABA: teoria, praktyka, przyszłość*. Kazimierz Dolny, 20-22 września 2006 roku. [dokument elektroniczny], [Warszawa], 2006. (EBIB Materiały konferencyjne nr 15). Dostępny w Internecie: http://www.e-bib.info/publikacje/matkonf/kaba/osinska.php [Data dostępu 11.05.2012]
17. *Ustawa z dnia 29 sierpnia 1997 r. o ochronie danych osobowych*. (Dz.U. 1997 nr 133 poz. 883 z późn. zmianami)
18. Wróblewski J. [et al.]: *P[olsko]-J[apońska] W[wyższa] S[zkoła] T[echnik] K[omputerowych – Hurtownie danych [wykłady]* [dokument elektroniczny]. 2006, modyfikacja: 2008. Dostępny w Internecie: http://edu.pjwstk.edu.pl/wyklady/hur/ scb/index.htm [Data dostępu: 15.04.2012]
19. Ziembicki P.: *Perspektywy rozwoju infrastruktury informatycznej bibliotek oraz nowa koncepcja roli bibliotekarza*. „Bibliotekarz" 2007 nr 6 s. 2-8