Studia i Materiały Informatyki Stosowanej, Tom 2, Nr 2, 2010

PRZEGLĄD I KLASYFIKACJA ZASTOSOWAŃ, METOD ORAZ TECHNIK EKSPLORACJI DANYCH

Marcin Mirończuk

Politechnika Białostocka
Wydział Elektryczny
ul.Wiejska 45E, 15-351 Białystok
e-mail: m.marcinmichal@gmail.com

Wzrost ilości danych jak i informacji w aktualnych systemach informacyjnych wymusił powstanie nowych procesów oraz technik i metod do ich składowania, przetwarzania oraz analizowania. Do analizy dużych zbiorów danych aktualnie wykorzystuje się osiągnięcia z obszaru analizy statystycznej oraz sztucznej inteligencji (ang. artificial intelligence). Dziedziny te wykorzystane w ramach procesu analizy dużych ilości danych stanowią rdzeń eksploracji danych. Aktualnie eksploracja danych pretenduje do stania się samodzielną metodą naukową wykorzystywaną do rozwiązywania problemów analizy informacji pochodzących m.in. z systemów ich zarządzania. W niniejszym artykule dokonano przeglądu i klasyfikacji zastosowań oraz metod i technik wykorzystywanych podczas procesu eksploracji danych. Dokonano w nim także omówienia aktualnych kierunków rozwoju i elementów składających się na tą młodą stosowaną dziedzinę nauki.

Eksploracja danych, metody eksploracji danych, techniki eksploracji danych, zastosowania eksploracji danych, ED

Data mining review and use's classification, methods and techniques

The large quantity of the data and information accumulated into actual information systems and their successive extension extorted the development of new processes, techniques and methods to their storing, processing and analysing. Currently the achievement from the statistical analyses and artificial intelligence area are use to the analysis process of the large data sets. These fields make up the core of data exploration – data mining. Currently the data mining aspires to independent scientific method which one uses to solving problems from range of information analysis comes from the data bases menagments systems. In this article was described review and use's classification, methods and techniques which they are using in the process of the data exploration. In this article also was described actual development direction and described elements which require this young applied discipline of the science.

Keywords: *Data mining, data mining methods, data mining techniques, data mining classification, data minig review*

1. WSTĘP

Rozwój technologii umożliwia składowanie coraz to większych ilości danych. Zasoby gromadzonej informacji w istniejących systemach informatycznych (różnego rodzajów bazach danych) zwiększyły się na tyle, iż przestają się sprawdzać uprzednio wystarczające rozwiązania przeszukiwania, wydobywania i transformowania informacji w wiedzę lub inny rodzaj informacji. Z tego też względu w procesie tworzenia systemów informatycznych bardzo istotne staje się konstruowanie stosownych narzędzi informatycznych i metod umożliwiających odpowiednie przekształcanie danych w informację. Nowe metody wraz z odpowiednimi narzędziami powinny umożliwiać też wydobywanie informacji z baz danych, z możliwością łatwego transformowania ich do wiedzy.

Odpowiedzią na rosnące zapotrzebowanie dotyczące gromadzenia, przeszukiwania czy analizy ciągle powiększających się zbiorów informacji w platformach informatycznych stała się eksploracja danych (*ang. data mining – DM*) tzw. wydobywanie wiedzy z danych. Ze względu na odmienne podejścia środowisk naukowych i biznesowych do eksploracji danych (ED) powstały różne

definicje tego zagadnienia. Najczęściej wykorzystywanymi w publikacjach definicjami, które najlepiej ukazują aktualny pogląd na temat ED, są:

Definicja 1 Interdyscyplinarna stosowana metoda naukowa która wywodzi się z dziedziny nauki i techniki, jaką jest informatyka zajmująca się wizualizacją i wydobywaniem ukrytej „implicite" informacji z dużych zasobów informacyjnych (baz danych) [1]. Wykorzystuje w tym celu technologie przetwarzania informacji oparte o statystykę i sztuczną inteligencję: uczenie maszynowe, metody ewolucyjne, logikę rozmytą oraz zbiory przybliżone.

Definicja 2 Jeden z etapów procesu odkrywania wiedzy z baz danych, określany w skrócie jako KDD (*ang.* k*nowledge discovery in databases*). Termin ten został użyty w pierwszej pracowni KDD w 1989 roku w celu uwydatnienia tego, iż wiedza jest końcowym produktem odkrywania sterowanego danymi (*ang. data-driven discovery*). Odkrywanie wiedzy z baz danych jest wykorzystywane w takich naukach jak, sztuczna inteligencja (*ang. artificial inteligence – AI*) czy też uczenie maszynowe (*ang. machine learning*) [2, 3].

Definicja 3 Kompletna metodologia CRISP-DM (*ang.* c*ross-industry standard process for data mining*) opracowana przez trzy przedsiębiorstwa przemysłowe: SPSS (*ang.* s*tatistical package for the social science*), NCR (*ang. national cash register corporation*) oraz Daimler Chrysler. Metodologia ta dostarcza ujednolicony, elastyczny oraz kompletny model przeprowadzania procesu eksploracji danych w przedsiębiorstwach, niezależnie od ich specyfikacji [4-6].

Do definicji pierwszej należy odnosić się z dystansem ze względu na to, iż eksploracja danych jest młodą, dynamicznie rozwijającą się dyscypliną naukową. Natomiast definiowanie dyscypliny naukowej jest zawsze zadaniem kontrowersyjnym, ponieważ badacze często nie zgadzają się co do dokładnego zakresu i granic swojego obszaru badań [7]. Niemniej skala zastosowań eksploracji danych, stosowany aparat matematyczny i ilości publikacji odnoszących się do niej w opisach badań i analizy danych, pozwala na wstępne postrzeganie i definiowanie jej w ramach samodzielnej dyscypliny naukowej.

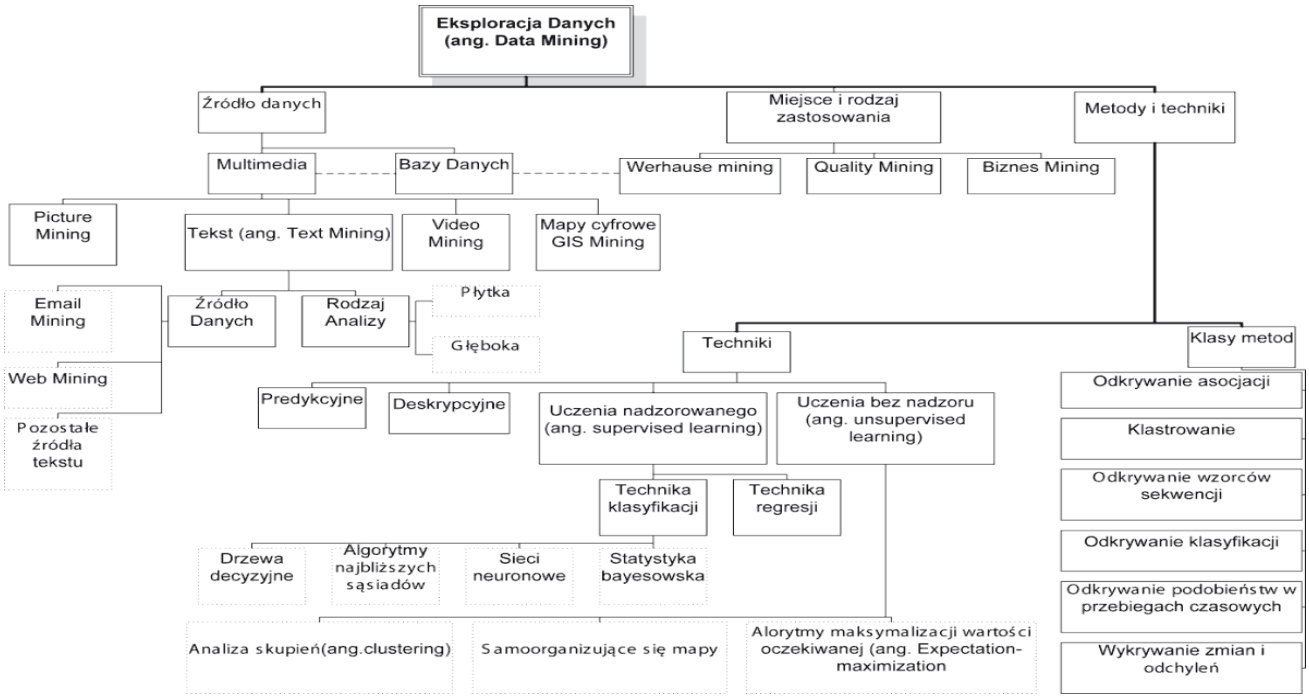
Przytoczone wyżej trzy definicje (*Def.1, Def.2, Def.3*) różnią się pod względem przyjmowanego punktu widzenia na temat eksploracji danych. Wspólną ich podstawę stanowi analiza zbiorów danych obserwowanych w celu znalezienia nieoczekiwanych związków i podsumowania danych w oryginalny sposób tak, aby były zarówno zrozumiałe, jak i przydatne w odpowiednich zastosowaniach [7]. Analiza ta zachodzi najczęściej w istniejących systemach informatycznych, w których zgromadzone informacje przekształcane są w inne rodzaje informacji przystosowanych do przeprowadzenia eksploracji danych (systemy preeksploracyjne). Analiza danych może także zachodzić w specjalnie skonstruowanych platformach informatycznych, przy konstruowaniu których odgórnie zakłada się potrzebę przeprowadzenia eksploracji danych (systemy posteksploracyjne).

W artykule skupiono się na opisie eksploracji danych jako całościowej dziedzinie nauki służącej do wydobywania wiedzy z systemów informacyjnych. Opis ED jako procesu można znaleźć w pracach [3, 8, 9]. W punkcie drugim niniejszego artykułu przedstawiono autorską klasyfikację eksploracji danych. Podział ten obejmuje trzy główne grupy: źródła danych, miejsce i rodzaj zastosowania oraz metody i techniki. W dalszych punktach rozwinięto i omówiono ww. trzy grupy. W punkcie trzecim przedstawiono odmiany eksploracji danych występujące w zależności od źródła danych jakie są do niej wykorzystywane. Punkt czwarty zawiera opis rodzaju badań jakie są wykonywane za pomocą ED oraz w jakich dziedzinach jest ona wykorzystywana. W punkcie piątym omówiono matematyczne podstawy całej dziedziny. Przedstawiono w nim podstawowe techniki i metody wykorzystywane do przeprowadzenia eksploracji danych.

PRZEGLĄD I KLASYFIKACJA ZASTOSOWAŃ, METOD ORAZ TECHNIK EKSPLORACJI DANYCH

Eksploracja danych jest dziedziną multidyscyplinarną, która skupia wokół siebie wiele dziedzin związanych z przechowywaniem, przetwarzaniem i analizowaniem informacji jak i wdrażaniem pozyskanej wiedzy w wyniku jej przeprowadzenia w zadany proces. Aktualnie termin eksploracja danych stosowany jest w różnych kontekstach, zależnych najczęściej od zastosowań czy tematu przeprowadzanych badań. Zastosowanie eksploracji danych w wielu różnych dziedzinach nauki jak i biznesu spowodowało powstanie wokół niej dużej liczby rozmaitej i nie do końca usystematyzowanej terminologii.

W punkcie tym przedstawiono autorską usystematyzowaną klasyfikację eksploracji danych i związane z nią pojęcia. Wynika ona z analizy dostępnej autorowi literatury, w której przedstawiono różne nie usystematyzowane pojęcia i zastosowania ED. Opracowaną klasyfikację prezentuje Rysunek. 1.



Rysunek. 1 Klasyfikacja eksploracji danych w zależności od rodzaju źródła danych, zastosowania oraz metod i technik [opracowanie własne]

nie przedstawia kompletnej klasyfikacji ED ze względu na to, iż jest to w dalszym ciągu młoda interdyscyplinarna dziedzina w trakcie dynamicznego kształtowania [7, 8]. Niemniej da się już wyodrębnić podstawowe nazewnictwo związane z kontekstem jej użycia oraz metod i technik używanych do jej przeprowadzenia. Autorski podział obejmuje trzy podstawowe grupy według których można rozpatrywać ED. Są to: źródła danych (rozdział 3), miejsce i rodzaj zastosowań (rozdział 4) oraz metody i techniki (rozdział 5).

ŹRÓDŁA DANYCH DO EKSPLORACJI DANYCH

Źródła danych do eksploracji danych stanowią różnego rodzaje pliki płaskie oraz bazy danych i systemy ich zarządzania (systemy zarządzania bazą danych – SZBD) [10]. Ogólnie, eksploracja danych przeprowadzana jest na pewnego rodzaju informacjach przechowywanych w SZBD, na który składa się baza danych. W szczególności baza danych jak i system zarządzający nią może być dostosowany do składowania i operowania na danych multimedialnych. Z tego też względu wyróżnione zostały dwa główne źródła ED: bazy danych i szczególny ich przypadek, który stanowią multimedialne bazy danych.

1.1. Bazy danych

Bazy danych z systemami ich zarządzania a także bazy multimedialne z systemami ich zarządzania można podzielić na dwie kategorie, wyodrębnione pod względem zastosowania czasowego i przeznaczenia ED. Pierwszą kategorię stanowią systemy *preeksploracyjne*, w których najczęściej analiza danych zachodzi w istniejących systemach informatycznych, w których zgromadzone informacje przekształcane są w inne rodzaje informacji przystosowanych do przeprowadzenia eksploracji danych [8]. Drugą kategorię stanowią natomiast platformy *poteksploracyjne*, w których analiza danych może zachodzić w specjalnie skonstruowanych platformach. Przy ich budowie odgórnie zakłada się potrzebę przeprowadzenia eksploracji danych [8].

1.2. Multimedia i multimedialne bazy danych

Multimedia stanowią szczególny przypadek bazy danych, która wykorzystuje różne formy składowania i przeszukiwania informacji w celu dostarczania odbiorcom wiedzy na ich temat. Multimedialne bazy danych i systemy ich zarządzania przystosowane zostały do pracy z takimi źródłami danych jak np.: tekst, dźwięk, grafika, wideo. Ze względu na to na jakim typie mediów dokonywana jest ED wydzielono osobne podgałęzie klasyfikacji obejmujące: eksplorację obrazów

(podrozdział 3.3), eksplorację nagrań audio (podrozdział 3.4), eksplorację danych wideo (podrozdział 3.5), eksplorację danych w geograficznym systemie informacji (podrozdział 3.6) oraz eksplorację danych tekstowych (podrozdział 3.7).

1.3. Eksploracja obrazów

Eksploracja obrazów (*ang. image mining*, *ang. picture mining*) [11, 12] dotyczy wydobywania wiedzy poprzez odkrywanie relacji między obrazami, czy też wzorów ukrytych (niejawnie) występujących w obrazach oraz pomiędzy nimi. Dziedzina ta wykorzystuje metody pochodzące z: widzenia komputerowego (*ang. computer vision*), przetwarzania obrazu, odzyskiwania obrazu, eksploracji danych, uczenia maszynowego, baz danych i sztucznej inteligencji. W eksploracyjnej analizie obrazów wyróżnia się dwa podejścia. Pierwsze polega na odkrywaniu z dużych zbiorów obrazów ich pojedynczych egzemplarzy. Drugie natomiast polega na odkrywaniu połączeń między zbiorami obrazów oraz występujących między nimi asocjacji.

1.4. Eksploracja nagrań audio

Eksploracja nagrań audio (*ang. audio mining*) [12, 13] polega na przetwarzaniu i analizowaniu danych dźwiękowych. Zajmuje się ekstrakcją, przetwarzaniem oraz wydobywaniem wiedzy z modeli muzycznych. Podstawowym jej celem jest wyszukiwanie informacji muzycznej (*ang. music information retrieval - MIR*). Wiedza ta pozwala użytkownikom poszukiwać i odnajdywać muzykę przy pomocy zawartości bazującej na tekście (*ang. content-based text*) i pytaniach audio takich jak: zapytanie-przez-ogłuszający/śpiewający/grający/wyjątki lub poprzez specyfikację z wykazem muzycznych terminów takich jak "szczęśliwy", "energiczny" itd. oraz poprzez połączenie i kombinację obydwu rodzajów wyszukiwań. Efektem takiej eksploracji może być ranking odpowiedzi oparty na oszacowaniu podobieństw odnoszących się do powiązanych plików audio.

1.5. Eksploracja danych wideo (nagrań filmowych)

Eksplorację danych wideo (*ang. video mining*) [13-15] definiujemy jako nienadzorowane odkrywanie wzorów w zawartościach baz multimedialnych przechowujących audio-wizualne dane. Za pomocą eksploracyjnej analizy danych wideo istnieje możliwość odkrycia interesujących zarejestrowanych zdarzeń, które dostępne są *a priori*. Wyróżniamy trzy typy nagrań audio-wizualnych, które są poddawane analizie:
- wyprodukowane np. filmy, reportaże, dramaty,
- nieopracowane dane filmowe np. monitoring ruchu ulicznego, wideo z nadzoru,
- nagrania medyczne np. ultra dźwiękowe wideo zawierające echokardiografie.

Na wszystkich trzech grupach dokonywana jest analiza dotycząca:
- wykrywania przyczyn wywołanych zdarzeń np. pojazdów wjeżdżających na teren chroniony, ludzi wchodzących i wychodzących z chronionych budynków,
- określania typowych i nieprawidłowych wzorów działalności,
- klasyfikacji obserwowanej działalności do wybranej kategorii np. chodzenie, jeżdżenie rowerem,
- grupowania i określania interakcji pomiędzy jednostkami (obiektami).

Eksploracyjna analiza wideo nie jest tylko procesem, który automatycznie ekstrahuje (wydobywa, przetwarza) zawartość i strukturę nagrań wideo, cech przesuwającego się obiektu, przestrzenne lub temporalne (czasowe) korelacje tych cech. Nastawiona jest ona również na odkrywanie wzorców struktury wideo, aktywności obiektów, zdarzeń etc. z olbrzymich zbiorów danych wideo bez niewielkich założeń co do ich zawartości. Przy użyciu eksploracyjnych technik analizy wideo takich jak, podsumowania, klasyfikacja, raportowanie o zdarzeniach (*ang. event alarm*) implementowane są tzw. sprytne aplikacje video. Zasadniczą różnicą pomiędzy konwencjonalną eksploracją danych a eksploracją nagrań filmowych jest fakt, iż głęboka analiza wideo operuje na mocno nieustrukturyzowanych danych. Surowe (nieopracowane) dane wideo zawierają tylko piksele, nawet przetworzone dane wideo są również złożonymi typami posiadającymi rozłączne wymiary. Dlatego też konwencjonalne algorytmy eksploracji danych nie mogą zostać bezpośrednio zastosowane w tej grupie danych.

1.6. Eksploracja danych w geograficznym systemie informacji

Eksploracja danych w geograficznym systemie informacji (*ang. geographic information system mining – GIS Mining*) stanowi analizę danych przeprowadzaną w tzw. geograficznym systemie informacyjnym GIS, który reprezentuje dane opisujące aspekty powierzchni ziemi takie jak np. drogi, domy etc. [16]. W sensie funkcjonalnym systemy GIS służą do analizy przestrzennej realizowanej poprzez [17-19]:
- predefiniowane w systemie raporty, zestawienia, wykresy wraz z wizualizacją przestrzenną,
- języki zapytań (np. *ang. standard query language – SQL*) do zintegrowanej graficznej i opisowej bazy,
- wyzwalacze (*ang. triggers*) używane w aktywnych systemach do ciągłego przetwarzania danych w wyniku ich uaktualnień.

Bazę graficzną tworzą cyfrowe mapy tematyczne (*ang. digital maps*), ortofotomapy, numeryczne modele terenu. Poszczególne obiekty bazy graficznej są

połączone z bazą opisową, której zawartość wynika ze struktury fizycznej - wynikającej głównie z rodzaju i zakresu informacji zawartych w bazie danych [20-22]. System zarządzania taką bazą określa się jako system zarządzania bazą danych przestrzennych SZBDP (a*ng. spatial database systems – SDBS*). Systemy zarządzania bazami danych przestrzennych są to systemy do zarządzania ww. danymi przestrzennymi poprzez np. wyszukiwanie, składowanie, uaktualnianie [23, 24]. Ilość jak i wielkość dostępnych przestrzennych baz danych szybko rośnie. Z tego też względu zostają ograniczone ludzkie możliwości analizy danych w nich zebranych dotyczących takich zagadnień, jak: odkrywanie ukrytych regularności, reguł lub skupień ukrytych w danych [16]. W celu poszerzenia i umożliwienia analizy tak dużej ilości danych zebranych w multimedialnych przestrzennych bazach danych stosuje się podejście określane jako: eksploracyjne odkrywanie danych przestrzennych (*ang. spatial data mining*) lub odkrywanie wiedzy w przestrzennych bazach danych (*ang. knowladge discovery in spatial databases*) [25]. Podejścia te reprezentują szczególny przypadek odkrywania, gdyż pozwalają wydobyć relacje, które istnieją między przestrzennymi i nie przestrzennymi danymi i innymi charakterystycznymi danymi, które jawnie nie są zgromadzone w przestrzennych bazach danych [26, 27]. Podstawową różnicą pomiędzy odkrywaniem wiedzy z relacyjnych a przestrzennych baz danych jest to, iż atrybuty sąsiadów pewnego obiektu, którym jesteśmy zainteresowani, mogą wpływać na sam obiekt zainteresowania [24].

Typowymi zadaniami odkrywania wiedzy w przestrzennych bazach danych są np. klasteryzacja czy charakteryzacja przez detekcje trendów. Używa się ich w celu odnalezienia ukrytych *implicte* regularności, czy reguł bądź wzorców w danych przestrzennych. Do podstawowych klas metod używanych w przestrzennej eksploracji danych można zaliczyć [16, 24, 28, 29]:

- grupowanie przestrzenne (*ang. spatial clustering*) – polega na grupowaniu obiektów bazy danych do znaczących podklas (klastrów) w taki sposób, aby poszczególne obiekty klastra były jak najbardziej podobne do siebie i jak najbardziej różne od elementów pozostałych klastrów. Zastosowanie klastrowania (grupowania) w przestrzennych bazach danych używa się np. w tworzeniu katalogu tematycznych map w geograficznych systemach informacji poprzez grupowanie wektorów cech,
- detekcja trendów przestrzennych (*ang. spatial trend detection*) – trend może zostać zdefiniowany jako czasowy wzór występujący w kilku seriach danych np. alarmy w sieci lub występowanie nawrotów chorób. W przestrzennym systemie bazy danych przestrzenny trend definiowany jest jako wzór zmiany nieprzestrzennych atrybutów w sąsiedztwie kilku obiektów bazy danych np. „kiedy następuje przeniesienie się ludzi z miasta X, siła nabywcza spada”,
- klasyfikacja przestrzenna (*ang. spatial classification*) – zadaniem klasyfikacji jest przydzielenie obiektu do klasy ze zbioru dostępnych wyselekcjonowanych klas na podstawie wartości atrybutów obiektu. Przestrzenna klasyfikacja może zostać użyta do wyjaśnienia odchyleń pomiędzy teoretycznymi a odkrytymi trendami przestrzennymi,
- charakterystyka przestrzeni (*ang. spatial characterization*) – jej zadaniem jest odnalezienie zwięzłego opisu (uogólnienia na pewien temat) dla wybranego podzbioru bazy danych.

1.7. Eksploracja danych tekstowych

Eksploracja danych tekstowych (*ang. text mining lub text data mining*) jest to analiza tekstu polegająca na wykorzystaniu inteligentnych reguł z zakresu uczenia maszynowego, lingwistyki komputerowej zajmującej się analizą języka naturalnego (*ang. natural language processing – NLP*), metod statystycznych oraz technik m.in. z zakresu przeszukiwania i grupowania danych [30]. Wykorzystywana jest do pozyskiwania informacji (wiedzy) z dużych nieustrukturyzowanych zbiorów danych tekstowych [31, 32]. Grupę tą można podzielić na dwie dodatkowe podgrupy. Pierwsza uwzględnia rodzaj przeprowadzanej analizy na tekście stanowiącym źródło danych (podrozdział 3.8). Druga natomiast została wydzielona ze względu na typ danych, rodzaj dokumentów (podrozdział 3.9).

1.8. Eksploracja danych tekstowych ze względu na typ analizy

W eksploracyjnej analizie tekstu, przeprowadzanej na tekście stanowiącym źródło danych, dostępne są dwie metody przetwarzania tekstu: płytkie i głębokie. Pierwsza metoda dotycząca płytkiej analizy tekstu (*ang. shallow text processing – STP*), określa grupę działań na tekście, których efekt jest niepełny w stosunku do głębokiej analizy tekstu. Polegają one na rozpoznawaniu struktur tekstów nierekurencyjnych lub o ograniczonym poziomie rekurencji, które mogą być rozpoznane z dużym stopniem pewności. Struktury wymagające złożonej analizy wielu możliwych rozwiązań są pomijane lub analizowane częściowo. Analiza skierowana jest głównie na rozpoznawanie nazw własnych, wyrażeń rzeczownikowych, grup czasownikowych bez rozpoznawania ich wewnętrznej struktury i funkcji w zdaniu. Analiza dotyczy też głównie dużych zbiorów dokumentów tekstowych a nie pojedynczych dokumentów a także takich zagadnień jak m.in. klasyfikacja (kategoryzacja) dokumentów (*ang. document classification lub document categorization*) ich

grupowania (*ang. dokument clustering*) i wyszukiwania z nich informacji (*ang. information retrieval – IR*) [33-35].

Druga metoda opiera się na tzw. głębokiej analizie tekstu (*ang. deep text processing – DTP*) i jest procesem komputerowej analizy lingwistycznej wszystkich możliwych interpretacji i relacji gramatycznych występujących w tekście naturalnym. Zazwyczaj jest bardzo złożona i z reguły dotyczy pojedynczego dokumentu. Pomija się wszelkie zależności statystyczne i stosuje się rozwiązania polegające na przetwarzaniu danych w oparciu o predefiniowane wzorce lub gramatyki [33, 36].

1.9. Eksploracja danych tekstowych ze względu na rodzaj dokumentów

Istnieje wiele źródeł danych nadających się do przeprowadzenia na nich eksploracyjnej analizy tekstu. Jedynym ich wymogiem jest to, aby informacja w nich była zakodowana w postaci znaków ASCI. Do źródeł danych w postaci dokumentów tekstowych, na których przeprowadzana jest tekstowa eksploracja danych, należą:
- wiadomości email (*ang. email mining*) – eksploracja tych wiadomości może być rozpatrywana jako specyfikacja badań z zakresu ogólno pojętej eksploracji tekstu nad internetowymi wiadomościami email [37-39]. Zasadniczymi cechami wyróżniającymi tę grupę od innych grup tekstowych są m.in.: wiadomości email są częściowo uporządkowane i posiadają zorganizowaną narzuconą przez standardy formę [40, 41], wiadomości tekstowe są znacznie krótsze od dokumentów, które podaje się zwykle analizie tekstu, emaile mogą zawierać treści na temat rozmaitych dyskusji na dowolne tematy. Fakt ten prowadzi do tego, że np. klasyfikacja poczty staje się bardziej trudna [38],
- dokumenty ogólnoświatowej multimedialnej sieci oprogramowania w internecie (*ang. world wide web wining – WEB Mining*) – technika wykorzystywana przy eksploracji tych danych ma na celu odkrywanie i uzyskiwanie przydatnych informacji, wiedzy i wzorców z dokumentów i usług Internetowych powszechnie określanych jako World Wide Web (WWW) [42, 43]. W obrębie tej techniki możemy wyróżnić trzy jej specjalizacje [44, 45]: eksploracja struktury dokumentów WWW (*ang. web structure mining – WSM*), eksploracja zawartości dokumentów WWW (*ang. web content mining*) i eksploracja użyteczności dokumentów WWW (*ang. web usage mining*). Eksploracja struktury dokumentów jest procesem, którego zadaniem jest wydobycie struktury informacji z sieci Web poprzez analizę hiperlinków tzn. linków wchodzących i wychodzących z dokumentu (strony, serwisu). Metoda ta wykorzystuje strukturę dokumentu, w którym strony jako węzły są połączone z innymi stronami za pomocą odnośników [46]. Algorytmami wykorzystywanymi do przeprowadzania WSM są m.in.: Hits (*ang. hyperlink-induces topic search*) i Page Rank (Google) [42]. Web Mining koncentruje się na dostarczaniu rozwiązań z zakresu [44, 47-51]: odnajdywania powiązanych informacji na podstawie np. analizy linków [52] i zawartości stron [53], tworzenia nowej wiedzy na podstawie informacji dostępnych na stronach Web, personalizowania i adaptowania stron Web, uczenia stron o zachowaniach klientów lub indywidualnych użytkowników np. na podstawie poruszania się użytkowników po portalu internetowym czy też segmentacji użytkowników danego serwisu i uzyskiwaniu informacji o ich położeniu geograficznym (przestrzenna natura Web Mining) [54]. Ponadto badania za pomocą Web Mining takich struktur WWW jak: blogi, fora, aukcje internetowe, obwieszczenia sklepowe mogą służyć do badania zmienności rynku [55] oraz wzmacniać tzw. analizę wokół klienta [56] poprzez wykrywanie odpowiednich grup społecznych [57], do których można zaadresować wybraną ofertę bądź ochrony go przed oszustwem poprzez zastosowanie np. odpowiednich algorytmów z zakresu badania reputacji uczestników aukcji *on-line* [58],
- pozostałe niesklasyfikowane dokumenty – jest to grupa uwzględniająca zbiór dokumentów niewymienionych i aktualnie nieobjętych klasyfikacją, które wraz z pojawianiem się nowych źródeł i form danych tekstowych czekają na sklasyfikowanie.

MIEJSCE I RODZAJ ZASTOSOWAŃ EKSPLORACJI DANYCH

Grupa ta obejmuje miejsca i sektory zastosowania technik eksploracji danych w przedsiębiorstwach zaprojektowanych według procesu planowania zasobów (*ang. enterprise resource planning – ERP*) i ze względu na ich zapotrzebowania. Nazwy poszczególnych podkategorii wywodzą się głownie od typu zastosowania i zapotrzebowania na ED. Można wyróżnić tutaj takie podgrupy, jak: eksploracja danych w biznesie (*ang. business mining*), która opisana została w podrozdziale 4.1; eksploracja danych na rzecz jakości (*ang. quality mining*), która opisana została w podrozdziale 4.2; eksploracja danych w hurtowniach danych (*ang. warehouse mining*), która opisana została w podrozdziale 4.3; eksploracja danych w przemyśle (*ang. industry mining*), która opisana została w podrozdziale 4.4 oraz pozostałe sektory i dziedziny zastosowań, które opisane zostały w podrozdziale 4.5.

1.10. Eksploracja danych w biznesie

Eksploracja danych w biznesie jest przeprowadzana w środowisku biznesowym, które w przeciwieństwie do

organizacji *no-profit*, nastawione jest na przynoszenie zysków. W organizacjach przynoszących, jak i nieprzynoszących zysku, eksploracja może być realizowana na wszystkich możliwych poziomach organizacji wyrażonej według np. modelu planowania zasobów przedsiębiorstwa (*ang. enterprise resource planning – ERP*) stanowiącego rozwinięcie systemu planowania zasobów produkcyjnych drugiego rzędu (*ang. manufacturing resource planning - MRP II*) [59]. Platforma ERP ma charakter komponentowy i składa się z takich modułów, jak [60-63]: system zarządzania klientami (*ang. customer relationship management – CRM*), system realizacji produkcji (*ang. manufacturing execution system – MES*), system zarządzania zasobami ludzkimi (*ang. human resource management – HRM*), system zarządzania finansami (*ang. financial resource management – FRM*). Dodatkowymi modułami mogą być: system planowania zapotrzebowania materiałowego (*ang. material requirements planning – MRP*) i zapotrzebowania na sprzęt. Przedstawiona lista nie wyczerpuje wszystkich możliwych modułów systemu ERP, daje natomiast, poprzez wykorzystywane jednoznaczne, opisowe pojęcia, wyraźny podział na wewnętrzne jednostki w przedsiębiorstwie profilowane do realizacji jego konkretnych funkcji, procesów i reakcji na jego zapotrzebowania.

W obrębie platformy ERP można wyróżnić dodatkowy abstrakcyjny komponent inteligencji biznesowej nazywany także analizą biznesową (*ang. business intelligence – BI*) [56]. Termin analiza biznesowa ma szerokie znaczenie. Najbardziej ogólnie można przedstawić ją jako proces przekształcania danych w informacje, a informacji w wiedzę, która może być wykorzystana do zwiększenia konkurencyjności przedsiębiorstwa a w przypadku przedsiębiorstwa *no-profit* może przynieść np. poprawę i ulepszenie pewnych aspektów prowadzonych przez nią działań. W obrębie jednostki analizy biznesowej można wyróżnić takie moduły, jak: system powiadamiania kierownictwa (*ang. executive information systems – EIS*), systemy wspomagania decyzji SWD (*ang. decision support systems – DSS*), system wspomagania zarządzania (*ang. management information systems – MIS*), system informacji geograficznej (*ang. geographic information systems – GIS*). Jednostka analizy biznesowej jest platformą pośredniczącą w wymianie i dostarczaniu informacji/wiedzy między poszczególnymi elementami struktury ERP. Jego najniższa warstwa (warstwa danych) reprezentowana jest przez różnego rodzaju systemy zarządzania danymi. W szczególności warstwę danych w tego rodzaju systemach stanowią hurtownie danych. Dostęp do systemów zarządzania danymi i wydobywanie z nich informacji/wiedzy realizowane jest przy wykorzystaniu interfejsu dostępu, na który składają się: aplikacje lub serwer aplikacji przeznaczony do przetwarzania analitycznego, transakcyjnego lub eksploracyjnego.

1.11. Eksploracja danych na rzecz jakości

Eksploracja danych na rzecz jakości (*ang. quality mining, ang. quality control data mining – QC DM*) jest to eksploracja danych wykonywana na rzecz kontroli jakości i polega na zgłębianiu danych w sterowaniu jakością [64]. Od zwykłej eksploracji odróżnia się m.in. tym, iż występuje tutaj konieczność reagowania na zmiany w danych na bieżąco (*on-line*). Innym wyróżnikiem eksploracji danych na rzecz jakości jest konieczność stosowania metod typowych dla sterowania jakością takich jak, karty kontrolne, analiza zdolności procesu, planowanie doświadczenia itp. Dane mają tutaj także swoją specyfikę – pochodzą z procesów technologicznych automatyki przemysłowej. Przez to zazwyczaj uzyskuje się mnóstwo parametrów, które często nie mają żadnego wpływu na wytwarzany w danej chwili produkt, ale mogą być decydujące dla innego produktu. Najczęstszym modelem jaki uzyskuje się w QC DM jest model „czarna skrzynka" [65].

1.12. Eksploracja danych w hurtowniach danych

Hurtownie danych, nazywane także magazynami danych (*ang. data warehouse*), są bardzo dużymi bazami danych, w których gromadzone są dane pochodzące z wielu heterogonicznych źródeł np. innych scentralizowanych lub rozproszonych baz relacyjnych, relacyjno-obiektowych, obiektowych, przestrzennych oraz ze źródeł innych niż bazy danych takich, jak arkusze kalkulacyjne, pliki XML, zasoby WWW [66-68]. Dane zebrane w hurtowni opisują (reprezentują) pewien określony tematycznie wycinek modelowanej rzeczywistości. Po za tym są zintegrowane, zmienne w czasie i stanowią nieulotną kolekcję. Dlatego też hurtownie wykorzystuje się najczęściej do wspierania procesu wspomagania decyzji [68] i stanowią główną bazę SWD. Hurtownia danych jest zazwyczaj wydzieloną centralną (na dany obszar) bazą danych odizolowaną od baz operacyjnych o typie przetwarzania danych transakcyjnych na bieżąco (*ang. on-line transaction processing – OLTP*) a jej struktura i użyte do jej budowy narzędzia są zoptymalizowane pod kątem przetwarzania analitycznego (*ang. on-line analytical processing – OLAP*) [69] lub hybrydowego (OLAP w połączeniu z OLTP) [68].

Hurtownie danych wyróżnione zostały z tego względu iż agregują dane historyczne, które przed załadowaniem do nich są poddawane procesowi oczyszczania, transformacji i ładowania (*ang. extraction, transformation, loading – ETL*). Proces ten przyspiesza

znacznie etap wstępnego przetwarzania danych w procesie KDD i CRISP-DM [2-5], który pochłania większość czasu przy przeprowadzaniu ED [70, 71]. Dlatego pod tym względem platformy te stanowią doskonałe środowisko do przeprowadzania zadań eksploracyjnych. Ponadto rozwija się szereg technik oraz metod (np. eksploracyjne zapytania ad-hoc rozszerzone o asocjacyjne reguły) jak i architektur systemów informatycznych (np. równoległa skalowalna infrastruktura OLAP i ED) łączących tradycyjne przetwarzanie analityczne z zawansowaną analityką w postaci eksploracji danych [72-75].

1.13. Eksploracja danych w przemyśle

Eksploracja danych w przemyśle (*ang. industry mining*) [76, 77] polega na wykorzystaniu dogłębnej analizy do rozwiązywania problemów pojawiających się w trakcie: produkcji przemysłowej jak i w etapach projektowych produkcji. Najczęściej firmy produkcyjne posiadają już systemy informatyczne dwojakiego rodzaju:
- pierwsze z nich związane są z ogólną obsługą, a więc z administracją, zarządzaniem, księgowością, zaopatrzeniem, zbytem itp. – system ERP,
- drugie związane są z obsługą procesów technologicznych, które rejestrowane są przez czujniki. Pochodząca z nich informacja gromadzona jest następnie w systemach bazodanowych np. relacyjnych, hurtowniach danych etc. wynikających z przyjętych założeń projektowych związanych z agregacją i przetwarzaniem danych.

Eksploracja danych w przemyśle skupiona jest zwłaszcza na drugim rodzaju systemów, w którym używa się jej do: projektowania i doskonalenia produktu (kontrola, poprawa jakości produktu poprzez kontrolę i poprawę procesu technologicznego), sterowania i optymalizacji procesu produkcyjnego, analiz reklamacji i niezawodności oraz bezpieczeństwa. Niemniej w pierwszym rodzaju systemu eksploracja może zostać zastosowana na etapie projektowania produktu oraz do polepszenia związków z klientami np. poprzez identyfikację ich potrzeb czy też prognozowania popytu na produkt. Cechami rozróżniającymi i wyróżniającymi eksplorację danych w przemyśle od pozostałych grup jest fakt, iż ma ona bliski związek ze statystycznym sterowaniem jakością procesów (SPC) i z eksploracją na rzecz jakości (podpunkt 4.2). Kolejnym wyróżnikiem jest to, iż procesy na linii technologicznej zachodzą w sposób ciągły (*on-line*) więc analiza eksploracyjna musi też zachodzić w sposób ciągły (*ang. on-line data mining*) [78] tak, aby można było reagować na zmiany w czasie rzeczywistym. Ponadto w przemyśle podczas analizy eksploracyjnej wykorzystuje się modele typu „czarna skrzynka” ze względu na bardzo szybkie zmiany produkcji – czas życia produktów i okres stosowania konkretnej technologii ciągle się zmienia.

1.14. Pozostałe

Spektrum aktualnie przeprowadzanych badań w różnych dziedzinach z zastosowaniem ogólnie pojętej eksploracji danych jest tak bogate i różnorodne, iż nie da się w jednym artykule przedstawić i opisać wszystkich możliwych kombinacji i zastosowań tej dziedziny. Lepszym rozwiązaniem zdaje się być skupienie na podstawach ED a mianowicie aparacie matematycznym opisanym w punkcie .

Do interesujących, bliżej niescharakteryzowanych w artykule badań, w których ogólno pojęta ED jest wykorzystywana należą takie dziedziny, jak: zabezpieczenia systemów informatycznych i wykrywanie w nich ataków sieciowych [79], medycyna [80, 81], badania naukowe np. dotyczące cząstek elementarnych w akceleratorze CERN [82], matematyczne dowodzenie twierdzeń [83], ochrona publiczna [84, 85], systemy ekspertowe i wspomagania decyzji [86, 87], energetyka i prognozowanie na zapotrzebowanie energii elektrycznej [88] będące zagadnieniem predykcyjnym [89], telekomunikacja i badanie np. fałszywych alarmów [90, 91] czy też powszechnie opisywane oraz badane zastosowania ED w business mining (przedsiębiorstwach produkcyjnych) [92] a dokładniej w handlu elektronicznym (*ang. electronic commerce, e-commerce*) wykorzystującym web mining) [70, 71, 93, 94]. Szacowany udział procentowy wykorzystania eksploracji danych w różnych sektorach działalności przedsiębiorstw można znaleźć w opracowaniu [95].

TECHNIKI I METODY EKSPLORACJI DANYCH

Grupa technik i metod eksploracji danych jest najbardziej priorytatywna ze względu na to, iż zawiera matematyczne podstawy całej dziedziny, które umożliwiają fizyczną realizację algorytmów eksploracji [7] na rzecz badań w wybranej dziedzinie poprzez implementację aplikacyjną.

1.15. Techniki eksploracji danych

Techniki eksploracji można podzielić ogólnie na cztery równoległe kategorie, w skład których wchodzą: techniki predykcyjne (podrozdział 5.2), techniki deskrypcyjne (podrozdział 5.3), techniki uczenia nadzorowanego (podrozdział 5.4) i techniki uczenia bez nadzoru (podrozdział 5.5). Przedstawione kategorie nie są ścisłe tj. technika predykcyjna może posługiwać się technikami z zakresu uczenia nadzorowanego i na odwrót. A zatem mogą istnieć pewnego rodzaju permutacje technik w celu osiągnięcia wyznaczonego celu badań.

1.16. Techniki predykcyjne

Techniki predykcyjne, inaczej nazywane technikami lub modelami przewidywania (*ang. predictive techniques*), starają się na podstawie odkrytych wzorców dokonać uogólnienia i przewidywania wartości danej zmiennej. Pozwalają na przewidywanie wartości zmiennej wynikowej na podstawie wartości pozostałych zmiennych (badawczych lub przewidujących) [7, 96, 97]. Techniki te w SWD wykorzystywane są do przewidywania i szacowania np. zasobów (sprzętu/ludzi) do rozwiązywania postawionego problemu.

1.17. Techniki deskrypcyjne

Techniki deskrypcyjne, nazywane także technikami bądź modelami opisowymi (*ang. description techniques*), służą do formułowania uogólnień na temat badanych danych w celu uchwycenia ogólnych cech opisywanych obiektów oraz ich najważniejszych aspektów [7, 97]. Techniki te w SWD stosuje się do odkrywania grup i podgrup podobnych zdarzeń lub identyfikacji zdarzeń.

1.18. Techniki uczenia nadzorowanego

Techniki uczenia nadzorowanego (*ang. supervised learning*) wykorzystują zbiory danych w których każdy obiekt posiada etykietę przypisującą go do jednej z predefiniowanych klas. Na podstawie zbioru uczącego budowany jest model, za pomocą którego można odróżnić obiekty należące do różnych klas [7, 97]. Technikami z zakresu uczenia nadzorowanego są techniki klasyfikacji stosowane od 1984 roku, do których należą drzewa decyzyjne (1984 rok) [98], algorytmy najbliższych sąsiadów (1992 rok) [99], sieci neuronowe (1991 rok) [100], statystyka bayseyowska (klasyfikacja bayseyowska 1992 rok i sieć bayseyowska 1995 rok) [101], algorytmy maszyny wektorów wspierających SVM (*ang. support vector machine*, 1995 rok) [102] oraz techniki regresji [7].

1.19. Techniki uczenia bez nadzoru

W przypadku technik uczenia bez nadzoru (*ang. unsupervised learning*) brak jest etykiet obiektów, nie ma także zbioru uczącego. Techniki te starają się sformułować model (modele) wiedzy najlepiej pasujące do obserwowanych danych [96, 97]. Technikami z zakresu uczenia bez nadzoru są: techniki analizy skupień, klastrowania (*ang. clustering*) [103], samoorganizujące się mapy (*ang. self-organization map*) [104], algorytmy aproksymacji wartości oczekiwanej (*ang. expectation-maximization*) [105] czy też zbiory przybliżone [106].

1.20. Metody eksploracji danych

Metody eksploracji danych bazują na technikach i stanowią ich uogólnienie. Realizowane są za pomocą wybranej techniki przy użyciu odpowiedniego dla niej algorytmu eksploracji danych [7]. Do metod ED zaliczamy m.in.: odkrywanie asocjacji (podrozdział 5.7), klastrowanie (podrozdział 5.8), odkrywanie wzorców sekwencji reguł (podrozdział 5.9), odkrywanie klasyfikacji (podrozdział 5.10), odkrywanie podobieństw w przebiegach czasowych (podrozdział 5.11) i wykrywanie zmian i odchyleń (podrozdział 5.12).

1.21. Metody odkrywania asocjacji

Pojęcie reguł asocjacyjnych (*ang. association rulet*) zostało po raz pierwszy wprowadzone w 1993 roku przez R. Agrawala, T. Imielinskiego, A. Swami [107]. Odkrywanie asocjacji (powiązań) polega na wykrywaniu różnego rodzaju zależności występujących między danymi w bazie danych. Precyzyjniej mówiąc zależności te określone są za pomocą korelacji reguł asocjacyjnych wiążących współwystępowanie podzbiorów elementów w dużej kolekcji zbiorów. Znalezione korelacje prezentowane są jako reguły postaci $X \Rightarrow Y$ (wsparcie, ufność), gdzie X i Y są rozłącznymi zbiorami elementów. Termin *wsparcie* oznacza częstotliwość występowania zbioru $X \cup Y$ w kolekcji zbiorów, zaś termin *ufność* określa prawdopodobieństwo warunkowe P(X|Y) [108, 109].

1.22. Metody klastrowania

Klastrowanie, nazywane także grupowaniem lub analizą skupień (*ang. clustering*), polega na znajdowaniu skończonych zbiorów klas obiektów (klastrów) w bazie danych posiadających podobne cechy. Podczas tego procesu zbiór obiektów dzielony jest na takie podzbiory aby jednocześnie maksymalizować podobieństwo między obiektami przypisanymi do tego samego podzbioru i minimalizować podobieństwo miedzy obiektami przypisanymi do różnych podzbiorów zgodnie z zadaną miarą podobieństwa między obiektami [97]. Podczas dokonywania klastrowania nie są znane docelowe podzbiory (grupy) obiektów oraz zazwyczaj nie jest znana ich liczba [108]. Z tego względu klastrowanie należy do tzw. klasyfikacji bez nadzoru i jest rozwiązywana za pomocą przeznaczonych do tego technik wymienionych w podpunkcie 5.5. Ponadto algorytmy przeznaczone do analizy skupień można podzielić na kilka podstawowych kategorii na które składają się [97, 108, 110, 111]: metody hierarchiczne (procedury aglomeracyjne i deglomeracyjne), grupy metod *k*-średnich (*ang. k-means*),

metody rozmytej analizy skupień (*ang. fuzzy clustering*) oraz metody niechierarchiczne.

1.23. Metody odkrywania wzorców sekwencji reguł

Problem odkrywania wzorców sekwencji został po raz pierwszy sformułowany w 1995 roku przez niektórych twórców metody asocjacyjnej m.in. Rakesh Agrawal oraz Ramakrishnan Srikant. Sekwencję stanowi uporządkowany ciąg zbiorów elementów, w którym każdy zbiór posiada znacznik czasowy [108]. Wzorce sekwencji stanowią rozwinięcie modelu reguł asocjacyjnych o takie elementy, jak [97, 111]: następstwa zdarzeń, ograniczenia dotyczące maksymalnych interwałów czasowych między kolejnymi wystąpieniami elementów sekwencji. Wprowadzenie interwałów czasowych umożliwiło nakładanie pewnego rodzaju okien czasowych do filtrowania sekwencji. Odkrywanie wzorców sekwencji polega ogólnie na znalezieniu w bazie danych sekwencji, podsekwencji występujących częściej niż zadany przez użytkownika próg częstości, zwany progiem minimalnego wsparcia (*ang. minsup*) w pewnym przedziale czasu.

1.24. Metody odkrywania klasyfikacji

Klasyfikacja (*ang. classification*) polega na zbudowaniu modelu przypisującego nowy, wcześniej nie znany obiekt, do jednej ze zbioru predefiniowanych klas. Przypisanie to następuje na podstawie wcześniejszego uczenia klasyfikatora (modelu umożliwiającego takie przypisanie) na zbiorze uczącym [97]. Najczęściej stosowanymi technikami do klasyfikacji są: klasyfikacja bayesowska, adaptywna sieć Bayesa, algorytmy indukcyjnych drzew decyzyjnych, algorytm *k* najbliższych sąsiadów, sieci neuronowe czy też algorytm SVM [108].

1.25. Metody odkrywania podobieństw w przebiegach czasowych

Odkrywanie podobieństw w przebiegach czasowych polega na odnalezieniu punktów wspólnych opisujących grupę wyselekcjonowanych przebiegów opisujących zadany proces trwający ciągle w czasie [109].

1.26. Metody wykrywania zmian i odchyleń

Wykrywanie zmian i odchyleń polega na znajdowaniu różnic pomiędzy aktualnymi a oczekiwanymi wartościami danych. Wykorzystywane jest podczas znajdowania anomalnych tj. niepasujących do trendu danych które od niego odstępują [109].

1.27. Metody odkrywania cech

Odkrywanie cech wykorzystywane jest najczęściej we wstępnych procesach (*ang. preprocessing*) eksploracji danych [3] w celu zmniejszenia wymiarowości rozpatrywanego problemu a więc i zwiększenia efektywności metod eksploracji danych. W celu zmniejszenia wymiarowości problemu stosuje się tzw. wybór cech (*ang. feature selection*) i odkrywanie cech (*ang. feature extraction*) czy też analizę składowych głównych (*ang. principal components analysis – PCA*). Pierwsza z metod polega na wyselekcjonowaniu z grupy tych atrybutów tylko które posiadają istotną wartość informacyjną. Dwie następne metody polegają na połączeniu atrybutów i stworzeniu ich liniowej kombinacji w celu zmniejszenia liczby wymiarów i uzyskania nowych składowych głównych [7, 108, 111, 112]. Wybór i generacja nowych atrybutów może odbywać się w sposób nadzorowany lub bez nadzoru [111].

WNIOSKI

Eksploracja danych, stanowiąca jeden z etapów procesu np. odkrywania wiedzy z baz danych czy też traktowana jako dziedzina nauki, niewątpliwie jest zagadnieniem interdyscyplinarnym. Na jej interdyscyplinarność ma wpływ nie tylko szerokie spektrum jej aktualnych, opisanych w artykule, zastosowań ale także bogaty aparat matematyczny zaczerpnięty z różnych dziedzin nauki w celu pozyskiwania wiedzy z ogromnych zbiorów danych, które zazwyczaj są tylko częściowo ustrukturyzowane bądź wcale. Niezależnie od rodzaju danych, na których przeprowadzana jest eksploracja, wymagany jest zawsze dodatkowy nakład na skonstruowanie i opisanie samego celu badania jak i na określenie metody a następnie techniki oraz procesu do jego zrealizowania.

Skonstruowana i opisana klasyfikacja pozwala w łatwy sposób odnaleźć, umiejscowić i opisać własne badania w szerszym kontekście ED oraz umożliwia w łatwy sposób odnalezienie potrzebnej metody i techniki do ich realizacji. Ponadto przedstawiona klasyfikacja dostarcza początkowego usystematyzowanego słownika pojęć związanych z eksploracją danych, który w łatwy sposób można rozszerzać poprzez uzupełnianie go (odpowiednich gałęzi klasyfikacji) o własne definicje pojęć na danym polu zastosowań i badań naukowych.

Niektóre źródła danych mogą wymagać specyficznych metod oraz technik do przeprowadzenia na nich eksploracji danych. Wszystkie one mogą zostać dodane do wybranych gałęzi klasyfikacji według danych a następnie mogą zostać powiązane z odpowiednimi metodami oraz technikami przeprowadzania na nich eksploracji danych. Użycie takiego podejścia umożliwia więc kompleksowe, systematyczne i elastyczne

klasyfikowanie nowych zastosowań i powstających w ich obrębie pojęć z zakresu eksploracji danych.

[1] Wilk-Kołodziejczyk D. Pozyskiwanie wiedzy w sieciach komputerowych z rozproszonych źródeł informacji. In: Lesław H.H., editor. *Społeczeństwo informacyjne Wizja czy rzeczywistość?* [on-line] Kraków: Uczelniane Wydawnictwa Naukowo - Dydaktyczne, 2003, 30 maja. [dostęp: 16 listopada 2007] Dostępny w Internecie: http://winntbg.bg.agh.edu.pl/skrypty2/0095/285-295.pdf.

[2] Piatetsky-Shapiro G and Frawley JW. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

[3] Fayyad U, Piatetsky-Shapiro G and Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Magazine, 1996.

[4] Chapman P, *Clinton* J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. *CRISP-DM 1.0* Step-by-step data mining guide. [on-line]. [dostęp: 1 czerwca 2008] Dostępny w Internecie: http://www.crisp-dm.org/CRISPWP-0800.pdf.

[5] *CRISP-DM*. [on-line] [dostęp: 1 czerwca 2008] Dostępny w Internecie: http://www.crisp-dm.org/.

[6] Metodologia Data Mining - model referencyjny CRISP-DM. [on-line] [dostęp: 1 czerwca 2008] Dostępny w Internecie: http://www.spss.pl/konsulting/konsulting_datamining_metodologia.html.

[7] Hand D, Mannila H and Smith P. *Eksploracja danych*. Wydanie 1. Warszawa: Wydawnictwo Naukowo-Techniczne, 2005.

[8] Mirończuk M. Eksploracja Danych w kontekscie procesu Knowledge Discovery In Databases (KDD) i metodologii Cross-Industry Standard Process for Data Mining (CRISP-DM). 2009.

[9] Fayyad UM, G Piatetsky-Shapiro and Smyth P. From Data Mining to Knowledge Discovery: An Overview. AAAI Press/MIT Press, s. 1-36.

[10] Krasuski A and Maciak T. Historia rozwoju Systemów zarządzania bazami danych. Bezpieczeństwo i Technika Pożarnicza: Wydawnictwo CNBOP Józefów, 2006. p. 213-226.

[11] Stanchev P. Using Image Mining For Image Retrieval. 2003. Dostępny w Internecie: http://paws.kettering.edu/~pstanche/mexico.pdf.

[12] Kotsiantis S, Kanellopoulos D and Pintelas P. Multimedia mining. WSEAS Transactions on Systems, No 3, 2004, s. 3263-3268.

[13] Leman M, Clarisse LP, Baets BD, Meyer HD, Lesaffre M, Martens JP, et al. Tendencies, Perspectives, and Opportunities of Musical Audio-Mining. 2002. [dostęp: 15 września 2009] Dostępny w Internecie: http://www.sea-acustica.es/Sevilla02/mus01002.pdf.

[14] Dai K, Zhang J and Li G. Video Mining: Concepts, Approaches and Applications. [Beijing]: Multi-Media Modelling Conference Proceedings, 2006 12th International, 2006.

[15] Divakaran A, Miyahara K, Peker KA, Radhakrishnan R and Xiong Z. Video Mining Using Combinations of Unsupervised and Supervised Learning Techniques. SPIE Conference on Storage and Retrieval for Multimedia Databases, 2004. p. 235-243.

[16] Ester M, Kriegel H-P and Sander J. Spatial Data Mining: A Database Approach. Springer, 1997.

[17] Woźniak J and Ferenc J. Budowa systemów geoinformacyjnych w zakładach górniczych. Prace Naukowe Instytutu Górnictwa Politechniki Wrocławskiej, No 106, 2004, s. 225-232

[18] Agrawal R and Psaila G. Active Data Mining.

[19] Wang W, Yang J and Muntz R. An Approach to Active Spatial Data Mining Based on Statistical Information.

[20] Górniak-Zimroz J, Woźniak J and Zimroz R. Możliwości metod data mining w geograficznych systemach informacyjnych zorientowanych na zarządzanie zasobami ziemi. Prace Naukowe Instytutu Górnictwa Politechniki Wrocławskiej No 113, 2005, s. 75-86.

[21] Gramacki J and Gramacki A. Dane przestrzenne w bazach relacyjnych. Wykorzystanie danych przestrzennych, systemy zarządzania danymi przestrzennymi.

[22] Gramacki J and Gramacki A. Dane przestrzenne w bazach relacyjnych. Model danych, zapytania przestrzenne.

[23] Gueting RH. An Introduction to Spatial Database Systems. Special Issue on Spatial Database Systems of the VLDB Journal, No 3, 1994.

[24] Ester M, Kriegel H-P and Sander J. Knowledge Discovery in Spatial Databases. [Bonn Germany]: Invited Paper at 23rd German Conf on Artificial Intelligence (KI '99), 1999.

[25] Santos M and Amaral L. Knowledge Discovery in Spatial Databases through Qualitative Spatial Reasoning. Portugal, 2000. [dostęp: 5 maja 2009] Dostępny w Internecie: http://repositorium.sdum.uminho.pt/bitstream/1822/5584/1/PADD2000_MS_LA.pdf.

[26] Koperski K and Han J. Discovery of Spatial Association Rules in Geographic Information Systems. [Maine]: Proc 4th International Symposium on Large Spatial Databases (SSD95), 1995.

[27] Santos M and Amaral L. Knowledge discovery in spatial databases - the padrão's qualitative approach.

[28] Kolatch E. Clustering Algorithms for Spatial Databases: A Survey. 2001. [dostęp: 10 września 2009]

Dostępny w Internecie: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.1145&rep=rep1&type=url&i=0.
[29] Ester M, Kriegel H-P and Sander J. Algorithms and Applications for Spatial Data Mining. 2001. [dostęp: 10 września 2009] Dostępny w Internecie: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.4689&rep=rep1&type=url&i=0.
[30] Lula P. Text mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych. StatSoft, 2005.
[31] Rajman M. Text Mining - Knowledge Extraction from Unstructured Textual Data. 6th Conference of International Federation of Classification Societies (IFCS-98), 1998.
[32] Hearst MA. Untangling Text data Mining. University of Maryland, 1999. p. 3-10.
[33] Sołdacki P. Wprowadzenie do eksploracji tekstu i technik płytkiej analizy tekstu.
[34] Kozłowski J and Neuman Ł. Wspomaganie wyszukiwania dokumentów mapami samoorganizującymi. [Wrocław]: III Krajowa Konferencja MISSI 2002, 19-20 września - „Multimedialne i Sieciowe Systemy Informacyjne”, 2002. [dostęp: 10 czerwca 2009] Dostępny w Internecie: http://www.zsi.pwr.wroc.pl/zsi/missi2002/pdf/s507.pdf.
[35] Borycki Ł and Sołdacki P. Automatyczna klasyfikacja tekstów. [Wrocław]: III Krajowa Konferencja MISSI 2002, 19-20 września - „Multimedialne i Sieciowe Systemy Informacyjne”, 2002. [dostęp: 10 czerwca 2009] Dostępny w Internecie: http://www.zsi.pwr.wroc.pl/zsi/missi2002/pdf/s504.pdf.
[36] Mykowiecka A. Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym. Warszawa: PJWSTK, 2007.
[37] Vel Od, Anderson A, Anderson A, Corney M and Mohay G. Mining E-mail Content for Author Identification Forensics. SIGMOD Record, No 30, 2001, s. 55-64.
[38] Katakis I, Tsoumakas G and Vlahavas I. Email Mining: Emerging Techniques for Email Management. Aristotle University of Thessaloniki, Department of Informatics, Greece, 2006. [dostęp: 19 sierpnia 2009] Dostępny w Internecie: http://lpis.csd.auth.gr/publications/katakis2006-idea.pdf.
[39] Vel Od. Mining E-mail Authorship. Salisbury Australia: Information Technology Division Defence Science and Technology Organisation, 2000. [dostęp: 10 sierpnia 2009] Dostępny w Internecie: http://www.cs.cmu.edu/~dunja/KDDpapers/DeVel_TM.pdf.
[40] RFC2822. [on-line] [dostęp: 19 sierpnia 2009] Dostępny w Internecie: http://www.faqs.org/rfcs/rfc2822.html.
[41] Email RFC. [dostęp: 15 czerwca 2009] Dostępny w Internecie: http://www.lemis.com/email/email-rfc.html.
[42] Szełemej Ł. Przegląd metod ekstrakcji wiedzy w serwisach WWW - Web Structure Mining.
[43] Bing L. Web Data Mining Exploring Hyperlinks, Contents, and Usage Data. 2007.
[44] Kosala R and Blockeel H. Web Mining Research: A Survey. SIGKDD Explorations, No 2, 2000, s. 1-15.
[45] Zaane OR, Osmar N and Zaane R. Resource And Knowledge Discovery From The Internet And Multimedia Repositories. Simon Fraser Uuniversity, 1999.
[46] Estivill-Castro V. Web Usage Mining Mining. 2003.
[47] Staś T. Wykorzystanie technik ewolucyjnych w procesie nowoczesnej personalizacji portali internetowych. Studia i materiały polskiego stowarzyszenia zarządzania wiedzą. Bydgoszcz: PSZW, 2007.
[48] Stwarz T. Teoretyczne podstawy adaptacyjnych stron, techniki odkrywania wiedzy stosowane do ich personalizacji oraz modułowa implementacja technik rozwiązań. [dostęp: 10 maja 2009] Dostępny w Internecie: http://www.klubinformatyka.pl/artykul.php?a=8&s=5.
[49] Wojciechowski M. Odkrywanie wzorców zachowań użytkowników WWW. [Poznań]: POLMAN'99, 1999. [dostęp: 5 sierpnia 2008] Dostępny w Internecie: http://www.cs.put.poznan.pl/mwojciechowski/papers/polman99a.pdf.
[50] Kita R. Analiza sposobu poruszania się użytkowników po portalu internetowym. StatSoft Polska, 2002. [dostęp: 10 sierpnia 2008] Dostępny w Internecie: http://www.statsoft.pl/czytelnia/dm/kita.pdf.
[51] Migut G. Segmentacja użytkowników serwisu WWW z użyciem metod statystycznych i sieci neuronowych. StatSoft, 2007. [dostęp: 10 sierpnia 2008] Dostępny w Internecie: http://www.statsoft.pl/czytelnia/8_2007/Migut05.pdf.
[52] Eiron N and McCurley KS. Analysis of Anchor Text for Web Search. 2003. [dostęp: 10 czerwca 2009] Dostępny w Internecie: http://www.mccurley.org/papers/anchor.pdf.
[53] Dill S, Kumar R, McCurley K, Rajagopalan S, Sivakumar D and Tomkins A. Self-Similarity in the Web. ACM TRANS INTER TECH, No 2, 2001, s. 2002.
[54] Morimoto Y, Aono M, Houle ME and McCurley KS. Extracting Spatial Knowledge from the Web. [dostęp: 10 sierpnia 2009] Dostępny w Internecie: http://www.mccurley.org/papers/SAINT03.pdf.
[55] Antweiler W and Frank MZ. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. The Journal of Finance, No 59, 2004, s. 1259-1294.
[56] Surman J. Analiza wokół klienta. Business intelligence Systemy wspomagania decyzji biznesowych.

Warszawa: Wydawnictwo Naukowe PWN, 2009. p. 95-104.
[57] M.Morzy. On Mining and Social Role Discovery in Internet Forums. [Warsaw, Poland]: 1st International Conference on Social Informatics SocInfo 2009, 2009. [dostęp: 22-24 June 2009].
[58] Morzy M. New Algorithms for Mining the Reputation of Participants of Online Auctions. Algorithmica, No 52, 2008, s. 95-112.
[59] Grześkowiak D. Historia ERP. [on-line] [dostęp: 10 kwietnia 2009] Dostępny w Internecie: http://www.erp-view.pl/erp/historia_erp.html.
[60] Planowanie Zasobów Przedsiębiorstwa (Enterprise Resource Planning - ERP) [on-line] [dostęp: 10 kwietnia 2009] Dostępny w Internecie: http://www.intuitivemfg.com/worldwide/polish/erp.htm.
[61] Systemy ERP. [on-line] [dostęp: 3 marca 2009] Dostępny w Internecie: http://www.magazynit.pl/systemy-erp-artykuly/.
[62] Planowanie zasobów przedsiębiorstwa [on-line] [dostęp: 20 sierpnia 2009] Dostępny w Internecie: http://pl.wikipedia.org/wiki/Planowanie_zasobów_przedsiebiorstwa.
[63] Górecki P. Systemy ERP - rys historyczny, wizje przyszłości. 2007.
[64] *StatSoft QC Miner*. [on-line] [dostęp: 1 maja 2009] Dostępny w Internecie: http://www.statsoft.pl/qcminer.html.
[65] Demski T. Data Mining w sterowaniu procesem (QC Data Mining). StatSoft Polska. [dostęp: 20 września 2008] Dostępny w Internecie: http://www.statsoft.pl/czytelnia/jakosc/sixqcminer.pdf.
[66] Zhang Y, Kunqing X, Xiujun M, Dan X, Cuo C and Shiwei T. Spatial Data Cube Provides Better Support for Spatial Data Mining. Geoscience and Remote Sensing Symposium, 2005 IGARSS '05 Proceedings 2005 IEEE International, No 2, 2005, s. 4.
[67] Wrembel R, Królikowski Z and Morzy M. Magazyny danych - stan obecny i kierunki rozwoju. Poznań: ProDialog, 2000. [dostęp: 10 czerwca 2009] Dostępny w Internecie: http://www.cs.put.poznan.pl/mmorzy/papers/prodialog01.pdf.
[68] Morzy M. Aktywne hurtownie danych. [Warszawa]: Najnowsze rozwiązania i praktyczne zastosowania: hurtownie danych i business intelligence, Centrum Promocji Informatyki, 2004. [dostęp: 23 March 2004].
[69] Traczyk T. Hurtownie danych – wprowadzenie. Infofestiwal'98, 1998.
[70] Ansari S, Kohavi R, Mason L and Zheng Z. Integrating E-Commerce and Data Mining: Architecture and Challenges. Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001) IEEE Computer Society, 2000. p. 27-34.
[71] Kohavi R and Provost F. Applications of Data Mining to Electronic Commerce. Data Mining and Knowledge Discovery, No 5, 2001, s. 2001.
[72] Goil S and Choudhary A. A Parallel Scalable Infrastructure for OLAP and Data Mining. IEEE Computer Society, 1999.
[73] Nestorov S. Ad-Hoc Association-Rule Mining within the Data Warehouse. In Proceedings of the 36 th Hawaii International Conference on System Sciences (HICSS 2003): IEEE Computer Society, 2003. p. 232-242.
[74] Liu Z and Guo M. A proposal of integrating data mining and on-line analyticalprocessing in data warehouse. [Beijing, China]: Info-tech and Info-net, 2001 Proceedings ICII 2001 - Beijing 2001 International Conferences on, 2001.
[75] Surman J. Zaawansowana analityka biznesowa. Business Intelligence Systemy wspomagania decyzji biznesowych. Warszawa: Wydawnictwo Naukowe PWN, 2009.
[76] Wyrozumski T. Eksploracja danych – dlaczego nie w przemyśle ? [Kościelisko]: VIII Konferencja PLOUG, 2002. [dostęp: Październik].
[77] Demski T. Data Mining w przemyśle: Projektowanie, Udoskonalanie, Wytwarzanie. StatSoft Polska, 2004. [dostęp: 14 września 2009] Dostępny w Internecie: http://www.statsoft.pl/czytelnia/jakosc/przemysl.pdf.
[78] Zakrzewicz M. On-Line Data Mining. [on-line] [Zakopane]: Konferencja PLOUG'98, 1998. [dostęp: 16 listopada 2007] Dostępny w Internecie: http://www.cs.put.poznan.pl/mzakrzewicz/presentations/ploug98.pdf.
[79] Kozłowski M. Systemy uczące się - studium problemów. Warszawa: Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych. [dostęp: 12 stycznia 2010] Dostępny w Internecie: http://home.elka.pw.edu.pl/~mkozlow3/artykuly/M.Kozlowski.pdf.
[80] Houston AL, Chen H, Hubbard SM, Schatz BR, Ng TD, Sewell RR, et al. Medical Data Mining on the Internet: Research on a Cancer Information System. Artificial Intelligence Review, No 13, 1999, s. 437-466.
[81] Mining Data to Save Children with Brain Tumors. [on-line] [dostęp: 1 czerwca 2008] Dostępny w Internecie: http://www.spss.com/success/.
[82] CERN. Dostępny w Internecie: http://cdsweb.cern.ch/.
[83] Bancerek G. Exploring MIZAR library with MML query. Zeszyty Naukowe Politechniki Białostockiej Seria Informatyka No 2, 2007, s. 5-19.
[84] Clinton B. New York University speech, Salon.com [on-line]. [dostęp: 1 czerwca 2008] Dostępny w Internecie:

http://www.salon.com/politics/feature/2002/12/06/clinton/print.html.
[85] McCue C. Data Mining and Predictive Analytics in Public Safety and Security. IT Professional, No 8, 2006, s. 12-18.
[86] Mirończuk M and Karol K. Koncepcja systemu ekspertowego do wspomagania decyzji w Państwowej Straży Pożarnej. In: Grzech A., Juszczyn K., Kwaśnicka H. and Nguyen N.T., editors. Inżynieria Wiedzy i Systemy Ekspertowe. Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2009.
[87] Mirończuk M and Maciak T. Problematyka projektowania modelu hybrydowego systemu wspomagania decyzji dla Państwowej Straży Pożarnej. Zeszyty Naukowe SGSP, 2009.
[88] Wyrozumski T. Sieci neuronowe a energetyka - prawdy i mity o prognozowaniu. [Kościelisko]: X Konferencja PLOUG, 2004.
[89] Wątroba J. Przykład rozwiązania zagadnienia predykcyjnego za pomocą technik Data Mining. StatSoft Polska, 2002. [dostęp: 11 września 2009] Dostępny w Internecie: http://www.statsoft.pl/czytelnia/prognozowanie/ZagadnieniePredykcyjne.pdf.
[90] Muraszkiewicz M. Eksploracja danych dla telekomunikacji. [Zakopane]: VI Konferencja PLOUG, 24-28 października, 2000. [dostęp: 10 września 2009] Dostępny w Internecie: http://www.ploug.org.pl/konf_00/pdf/muraszkiewicz.pdf.
[91] Kryszkiewicz M. Eksploracja danych w telekomunikacji. III edycja konferencji HURTOWNIE DANYCH I BUSINESS INTELLIGENCE - problematyka, rozwiązania, zastosowania, 2004.
[92] Kozielski S, Małysiak B and Kasprowski P. Zastosowanie metod eksploracji danych do badania sprzedaży w przedsiębiorstwie produkcyjnym. In: Mrozek D., editor. Bazy Danych: Struktury, Algorytmy, Metody. Gliwice: WKŁ, 2006.
[93] Banks DL and Said YH. Data Mining in Electronic Commerce. Statistical Science, No 21, 2006, s. 234-246.
[94] Yun C-H and Chen M-S. Mining Web Transaction Patterns in an Electronic Commerce Environment. In Proceedings of the 4th Pacific-Asia Conf on Knowledge Discovery and Data Mining, 2000. p. 216-219.
[95] Pal SK and Mitra P. Pattern Recognition Algorithms for Data Mining Scalability, Knowledge Discovery and Soft Granular Computing. London New York Washington, D.C.: CHAPMAN & HALL/CRC, 2004.
[96] Graf M, Mazurek T, Mycka A and Najdzionek E. Hurtownie danych metody eksploracji [on-line]. [dostęp: 1 czerwca 2008] Dostępny w Internecie: http://www.ioz.pwr.wroc.pl/Pracownicy/mercik/zbiory/Prezentacje%202007/publikacja%20-%20poprawiona.pdf.
[97] Morzy M. Co dalej z tą eksploracją? Rola i miejsce eksploracji danych w architekturze współczesnych systemów baz danych [Warszawa]: IX Edycja "Hurtownie danych i business intelligence", Centrum Promocji Informatyki, 2007. [dostęp: 23 października 2007].
[98] Quinlan JR. Induction of Decision Trees. Machine Learning, No 1, 1986, s. 81-106
[99] Aha DW. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. International Journal of Man-Machine Studies, No 36, 1992, s. 267-287
[100] McCord-Nelson M and Illingworth WT. A practical guide to neural nets. Addison-Wesley Longman Publishing Co., Inc., 1991.
[101] Bolstad WM. Introduction to Bayesian Statistics. Wydanie 2. Wiley-Interscience.
[102] Cristianini N and Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.
[103] Everitt BS, Landau S and Leese M. Cluster Analysis. 2001.
[104] Kohonen T. Self-Organizing Maps. In: Sciences S.S.i.I., editor. Wydanie 3. Berlin: Springer, 2001.
[105] Dempster AP, Laird NM and Rdin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, No 39, 1977, s. 1-38.
[106] Rutkowski L. Metody i techniki sztucznej inteligencji. Wydawnictwo Naukowe PWN, 2005.
[107] Agrawal R, Imielinski T and Swami A. Mining associations between sets of items in massive databases. [Washington]: In ACM SIGMOD International Conference on Management of Data, 1993.
[108] Morzy M. Eksploracja danych - przegląd dostępnych metod i dziedzin zastosowań [Warszawa]: VI edycja Hurtownia danych i business intelligence, 2006.
[109] Morzy T. Eksploracja danych: problemy i rozwiązania. [on-line] [Zakopane]: V Konferencja PLOUG 1999 - Integracja danych i systemów informatycznych, 1999, 12-16 października. [dostęp: 16 listopada 2007] Dostępny w Internecie: http://www.ploug.org.pl/konf_99/pdf/7.pdf.
[110] Gulczyński M. Techniki „odkrywania wiedzy" (data mining) oraz ich zastosowania. [Bydgoszcz]: Studia i materiały polskiego stowarzyszenia zarządzania wiedza, 2004.
[111] Morzy M. Oracle Data Mining - odkrywanie wiedzy w dużych wolumenach danych. [Zakopane]: XI Krajowa Konferencja PLOUG'2005 "Systemy informatyczne, 2005.
[112] Larose DT. Metody i modele eksploracji danych. In: PWN W.N., editor. Metody i modele eksploracji danych. Warszawa: Wydawnictwo Naukowe PWN, 2008. p. 2-35.