Eksploracja danych w kontekście procesu Knowledge Discovery In Databases (KDD) i metodologii Cross-Industry Standard Process For Data Mining (CRISP-DM)

Marcin Mirończuk[1], Tadeusz Maciak[2]

[1]*Politechnika Białostocka, Wydział Elektryczny,*
[2]*Politechnika Białostocka, Wydział Informatyki*

Abstract:

Article aims at introducing for the readers few problems connected with KDD process, Data Mining project modeling with the use of CRISP-DM. The systemized knowledge, approaches to and generic terms was presented in the article. In the first part article describes approach to Data Exploration as one of the KDD cycle, which is specialized Knowledge Discovery process. Then article takes the subject of CRISP-DM method. The context of method usage depending on scale and integration of project, which they concern – investigate of useing text mining in Inteligent Decission Support System (IDSS) develop by informatic faculty of Fire Service. At the end of the article the summary was made, which contains common features between the two looks on the exploration and extracting knowledge from data bases.

Keywords:

data exploration, data mining, Cross-Industry Standard Process for Data Mining, CRISP-DM, knowledge discovery in databases, KDD, pre-exploration systems, post-exploration systems, text mining, text analysis

1. Wstęp

Systemy informatyczne stanowiące integralną część platform informacyjnych stały się nieodłącznym elementem sprawnie działających firm, przedsiębiorstw, instytucji zarówno z sektora prywatnego jak i państwowego. Aktualnie odgrywają one także ważną rolę w środowiskach naukowych gdzie stanowią doskonałe narzędzie do przeprowadzania i realizacji badań z wybranej dziedziny. W wyniku prac badawczych zarówno środowiska naukowego jak i biznesowego, lub ich kooperacji, powstają nowe rozwiązania dla systemów informacyjnych z zakresu: sprzętu, oprogramowania, technik organizacji i zarządzania zasobami różnego rodzaju czy też budowy elementów samego systemu informacyjnego. Platformy informacyjne mają za zadanie przekształcić dane – szczątkowe nieuporządkowane sygnały, które mogą pochodzić ze źródeł pierwotnych albo wtórnych, tworzonych wewnątrz, jak i na zewnątrz organizacji, lub innych systemów generujących dane – w informację. Informacja powstaje natomiast jako rezultat integrowania i porządkowania danych, które w ten sposób nabierają znaczenia. Informacja następnie może zostać przekształcona w wiedzę – wartościową i zaakceptowaną informację, która integruje dane, fakty, informacje [1].

Systemem informacyjnym można nazwać także platformę, w której zestawienie procesów i sposobów komunikacji (zarówno na poziomie infrastruktury sprzętowo – programo-

wej jak i personalnej) zapewnia prawidłowe przekształcenie danych w informację i późniejszy jej obieg. Platforma informatyczna stanowi natomiast narzędzie, lub grupę powiązanych ze sobą narzędzi. Funkcją tych narzędzi jest przetwarzanie informacji przy użyciu technik komputerowych w inny rodzaj informacji, które mogą zostać następnie przetransformowane w wiedzę. Systemy informatyczne znacznie przyśpieszają i normalizują przetwarzaną informację stanowiącą zestaw faktów, jakie mają miejsce w otoczeniu funkcjonującego systemu. Na jego efektywność ma wpływ wiele czynników m.in.: poziom wyszkolenia pracowników, zastosowane oprogramowanie, infrastruktura sprzętowa, rozwiązania sieciowe, rozwiązania wykorzystane przy budowie oprogramowania, dostosowanie systemu do realizowanego zadania.

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 Data Mining, są:

Definicja 1 Interdyscyplinarna stosowana metoda naukowa – która wywodzi się z dziedziny nauki i techniki jaką jest informatyka. Zajmuje się ona wizualizacją i wydobywaniem ukrytej „implicite" informacji z dużych zasobów informacyjnych (baz danych) [2]. 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. *Knowledge 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*). Knowledge Discovery in Databases jest wykorzystywane w takich naukach jak Sztuczna Inteligencja (ang. *Artificial Inteligence – AI*) czy też Uczenie Maszynowe (ang. *Machine Learning*) [3, 4].

Definicja 3 Kompletna metodologia **CRISP-DM** (ang. *Cross-Industry Standard Process for Data Mining*) opracowana przez trzy przedsiębiorstwa przemysłowe: **SPSS** (ang. *Statistical 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 [5-7].

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 [8]. Z tego też względu aktualnie bezpieczniej jest (do czasu usystematyzowania nauki) odnosić się do Eksploracji Danych i osadzać ją w kontekście **KDD** [3, 9] lub **CRISP-DM** [5, 7].

Przytoczone wyżej trzy definicje (*Def.1, Def.2, Def.3*) różnią się pod względem przyjmowanego punktu widzenia na temat Eksploracji Danych. Wspólną ich podstawą jest 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 [10]. 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 poteksploracyjne).

W niniejszym artykule w punkcie drugim omówiono zastosowanie **ED** w kontekście procesów **KDD** i **CRISP-DM**. Dokonano szczegółowego opisu tych procesów, wraz z przytoczeniem ich genezy. W punkcie trzecim opisano eksperymentalne zastosowanie procesu **KDD** do eksploracyjnej analizy tekstu. Analiza ta zostanie przeprowadzona na potrzeby badań dotyczących budowy i zastosowania Inteligentnego Systemu Wspomagania Decyzji (ang. *Inteligent Decission Support System – IDSS*) [11] dla Państwowej Straży Pożarnej (**PSP**) zapoczątkowanych przez wydział informatyki Szkoły Głównej Straży Pożarnej (**SGSP**) [12].

2. Data mining w kontekście procesu KDD i CRISP-DM

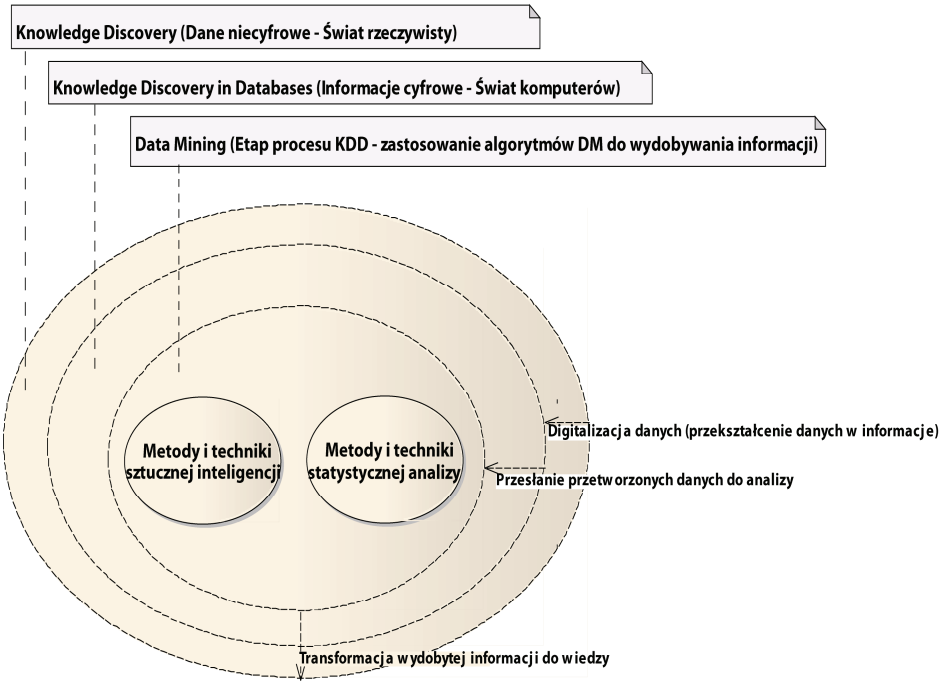
Knowledge Discovery in Databases i Cross-Industry Standard Process for Data Mining zazwyczaj dotyczą informacji składowanych w bazach danych z przyczyn innych niż analizy mające na celu wydobycie wiedzy (systemy preeksploracyjne). Aktualnie w wielu sytuacjach koniecznością stało się stosowanie do procesu analizy informacji w bazach danych metod i technik **KDD** i **CRISP-DM,** oraz związanych bezpośrednio z nimi algorytmów Eksploracji Danych. Wynika to z szybkiego przyrostu informacji w bazach danych oraz nieprzystosowania dotychczasowych metod i technik (np. *On-line Analytical Processing*, zapytań typu *ad-hoc Standard Query Language*) do wnikliwej analizy, generowania i samoistnej weryfikacji hipotez bez bezpośredniego udziału ekspertów [13, 14].

Wydobywaniem wiedzy z danych, oprócz grup naukowo-badawczych i prywatnych firm (korporacji), zainteresowane są także rządy niektórych krajów, w szczególności Stany Zjednoczone. Finansują one i wspierają badania w kierunku Data Mining, gdyż dostrzegły tkwiący w tych technikach potencjał dotyczący polepszenia ochrony zdrowia [15], wewnętrznego i zewnętrznego bezpieczeństwa narodowego [16] oraz bezpieczeństwa ekonomicznego.

2.1. Data Mining jako element procesu Knowledge Discovery in Databases

Knowledge Discovery in Databases jest specjalizacją procesu Knowledge Discovery (**KD**). Został opracowany w środowisku naukowym przez: Frawley, Piatetsky-Shapiro, Matheus, Fayyad i Smyth [3, 4, 17]. Odkrywanie wiedzy (ang. *Knowledge Discovery*) jest nie trywialnym procesem identyfikowania ważnych, nowych, potencjalnie przydatnych i ostatecznie zrozumiałych wzorów w danych [3]. Dodanie końcówki „In Databases" do Knowledge Discovery ma na celu podkreślenie, że transformacji podlegają zdigitalizowane dane, składowane w bazach danych, do postaci informacji użytecznej dla użytkownika

i algorytmu przetwarzania. Na rysunku 1 został przedstawiony konceptualny model relacji zachodzących pomiędzy pojęciami **KD**, **KDD** i Data Mining. Wynika z niego, że Knowledge Discovery jest szerszym pojęciowo znaczeniem odnoszącym się do relacji zachodzących w świecie rzeczywistym, które są badane i analizowane bez udziału komputerów.



Rysunek 1. Interpretacja graficzna pojęć KD, KDD, DM. Źródło: opracowanie własne

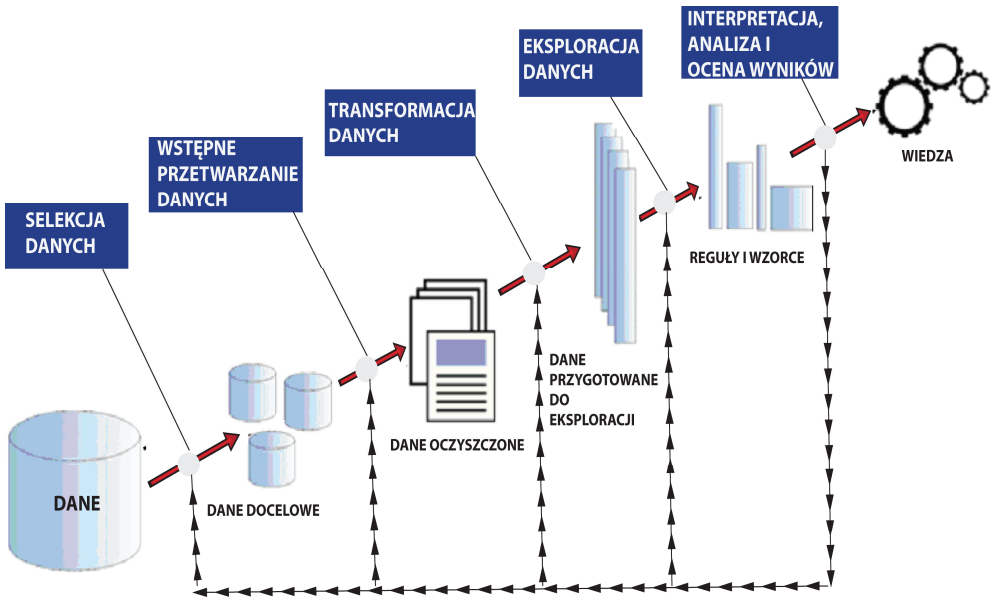
Poza graficzną interpretacją pojęć **KD**, **KDD** i **DM** na rysunku 1 przedstawiono też ogólny sposób przetwarzania danych w poszczególnych koncepcyjnych warstwach. Dane ze świata rzeczywistego poddane procesowi digitalizacji i modelowania mogą zostać analizowane za pomocą technik i metod wykorzystujących do obliczeń komputery. Odpowiednio przetworzone informacje cyfrowe, na początkowych etapach procesu **KDD**, są następnie przetwarzane przez metody i algorytmy Data Mining oparte o Sztuczną Inteligencję (**SI**) i statystykę w celu wydobycia informacji i utworzenia modelu procesu z analizowanych danych.

Proces **KDD** jest iteracyjny, zawsze istnieje możliwość cofnięcia się w celu ewentualnej korekty parametrów czy weryfikacji założeń. Ponadto na każdym z jego etapów niezbędny jest udział człowieka w postaci eksperta lub grupy ekspertów z danej dziedziny do której dostosowuje się **KDD**. Na rysunku 2 zostały przedstawione podstawowe etapy **KDD,** na które składają się następujące czynniki [3, 18, 19]:

— zapoznanie się z wiedzą dziedzinową aplikacji i zidentyfikowanie celów procesu odkrywania wiedzy;
— selekcja danych ze skonsolidowanego zbioru danych utworzonego najczęściej ze zbioru danych pochodzących z heterogenicznych i rozproszonych źródeł, ważnych dla badanego problemu;
— wstępne przetwarzanie danych (czyszczenie, integracja, selekcja, uzupełnianie) – etap, na którym następuje usuwanie niespójnych, niepełnych informacji nieistotnych z punktu widzenia dokonywanej analizy. Dokonywane są także zabiegi uzupełniania danymi

zmiennych wybranych do analizy, które nie posiadają przypisanej wartości za pomocą np.: uśredniania, wpisania stałej wartości, wygenerowania losowej wartości z obserwowanego rozkładu zmiennej;

— odnajdywanie i redukcja zmiennych – odkrywanie przydatnych cech (zmiennych) do modelowania postawionego celu zadania. Redukowanie wymiarowości zagadnienia (redukcja zmiennych) lub transformacja zmiennych poprzez odnajdywanie niezmiennych zależności (bliskiej korelacji) między nimi. Odkrywanie punktów oddalonych zmiennych (ang. *outliers*);



Rysunek 2. Proces KDD. Źródło: [3]

— dostosowanie procesu **KDD** do realizacji celu ustalonego w punkcie pierwszym – wybór odpowiedniej metody lub algorytmu Eksploracji Danych z zakresu: opisu (podsumowania), szacowania (estymacji), przewidywania (predykcji), klasyfikacji, grupowania, odkrywania reguł w badanym zbiorze danych;

— transformacja danych – procesowi temu podlegają dane oczyszczone i ustrukturyzowane, które można przekształcić do postaci wymaganej przez algorytm eksploracji danych. Przekształcanie danych dokonywane jest m.in. za pomoc normalizacji, jak: min-max (ang. *min-max normalization*) lub standaryzacji np. za pomocą *Z-score standardization;*

— eksploracja danych – etap na którym zostaje utworzony model (modele) bądź wydobyty wzorzec (wzorce) za pomocą wcześniej ustalonych metod i algorytmów Data Mining odpowiednich do rozwiązywanego problemu;

— interpretacja, analiza i ocena wyników – identyfikacja interesujących wydobytych zależności ze wszystkich otrzymanych modeli, wzorców. Oszacowanie ich i wybranie rozwiązania które najbardziej pasuje do realizacji zakładanych na samym początku celów. Etap ten powinien prowadzić do uzyskania przydatnego modelu badanego procesu w kolejnych iteracjach poprzez wprowadzanie odpowiednich modyfikacji na przestrzeni całego procesu odkrywania wiedzy, a w końcu do uzyskania użytecznej wiedzy;

— prezentacja wyników – graficzna reprezentacja wydobytych danych w postaci różnego rodzaju wykresów, czy diagramów. Krok przeprowadzany równolegle z interpretacją, analizą i oceną wyników.

2.2. Data Mining jako model CRISP-DM w zarządzania projektami

Cross-Industry Standard Process for Data Mining jest alternatywnym podejściem do zagadnienia eksploracji danych. **CRISP-DM** został opracowany w środowisku biznesowym tj. przez trzy koncerny: **SPSS** (ang. *Statistical Package for the Social Science*), **NCR**(ang. *National Cash Register Corporation*) oraz Daimler Chrysler [5, 20]. Prace zostały zainicjowane w 1996 roku, niedługo po utworzeniu modelu **KDD** i pierwszych publikacjach na jego temat. Twórcami **CRISP-DM** oraz jego specyfikacji byli: Chapman (NCR), Clinton (SPSS), Kerber (NCR), Khabaza (SPSS), Reinartz (Daimler Chrysler), Shearer (SPSS) i Wirth (Daimler Chrysler). W sierpniu 2000 roku opublikowana została specyfikacja wersji 1.0 [7]. W przeciwieństwie do **KDD**, które jest chętnie adoptowane w środowisku naukowo-badawczym, **CRISP-DM** powstał do celów komercyjnych (handlowych, przemysłowych) i jest częściej adoptowane w środowisku biznesowym np. do budowy elementów systemu **ERP** (ang. *Enterprise Resource Planning*) takich jak **CRM** (ang. *Customer Relationship Management*) i platformy **BI** (ang. *Business Intelligence*) [21, 22].

W **CRISP-DM** zostały zebrane najlepsze praktyki tworzących go przedsiębiorstw. Oferuje kompletny, uniwersalny model (Rysunek 3), który może być stosowany do przeprowadzania wielokrotnego procesu eksploracji danych niezależnie od rodzaju przemysłu, narzędzi i oprogramowania w bardziej dogodny, efektywny, niezawodny i mniej kosztowny sposób [5, 7].

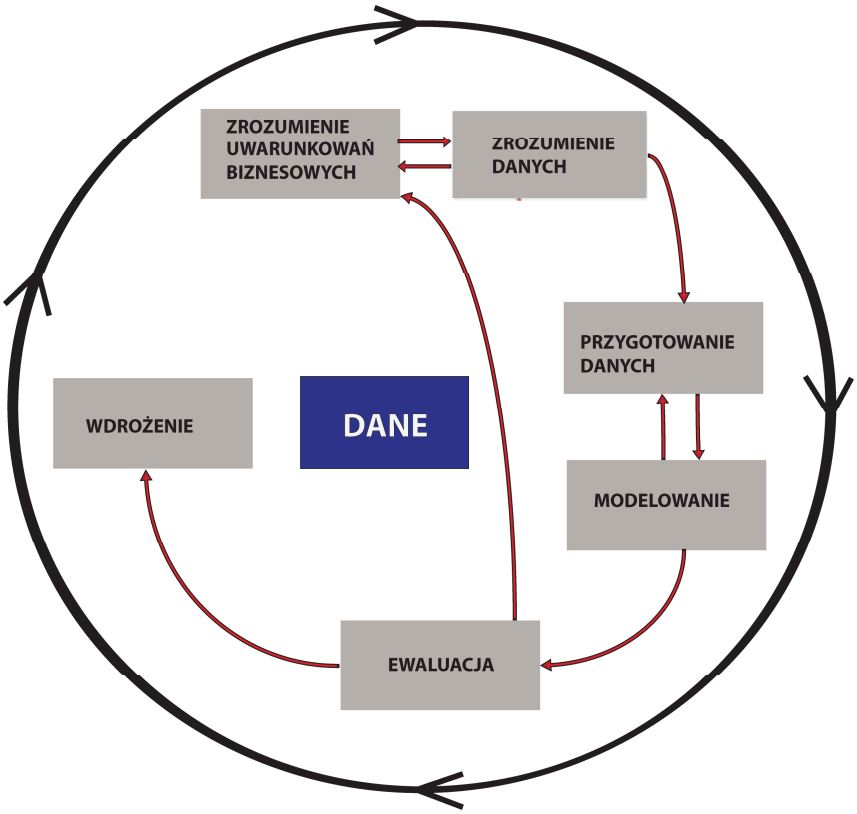
Zgodnie z metodologią **CRISP-DM** cykl procesu eksploracji danych, który został przedstawiony na rysunku 3, podzielony jest na sześć głównych etapów [5, 7]:

— zrozumienie uwarunkowań biznesowych (ang. *Business Understanding*) – pierwsza faza projektu, która może być także traktowana jako zrozumienie uwarunkowań badawczych. Skupia się na zrozumieniu celów i wymagań biznesowych i przekształceniu tej wiedzy w definicje problemu Data Mining. Następuje ustalenie wstępnego planu realizacji projektu systemu, który posłuży do osiągnięcia wyznaczonych celów;

— zrozumienie danych (ang. *Data Understanding*) – początkowa faza dotycząca zrozumienia danych. Zgromadzenie potrzebnych danych w jedno skonsolidowane źródło. Zapoznanie się z danymi, zidentyfikowanie problemu jakości danych, odkrywanie pierwszych zależności w danych lub wykrywanie interesujących grup (podzbiorów), w celu utworzenia wstępnych hipotez;

— przygotowanie danych (ang. *Data Preparation*) – faza zajmująca największą ilość czasu. Pracochłonny etap przekształcania wstępnych nieobrobionych danych, którego finałem są zestawy danych wykorzystywane we wszystkich pozostałych fazach (dane te są wykorzystane w modelowaniu przez programy narzędziowe). Zadanie przygotowania danych jest czasami wielokrotnie powtarzanym etapem, który nie posiada określonego porządku wykonywania czynności z nim związanych. Odpowiednio wyselekcjonowane, przetransformowane i oczyszczone tabele, rekordy i atrybuty są gotowe do wykorzystania przez programy narzędziowe;

— modelowanie (ang. *Modeling*) – w fazie tej stosuje się wyselekcjonowane i optymalnie skalibrowane pod względem parametrów, różne techniki i metody modelujące. Zwykle stosuje się kilka technik dla tego samego problemu eksploracji danych w celu otrzymania grupy modeli, z pośród których można wybrać optymalny dla osiągnięcia postawionego celu. Niektóre techniki modelowania mają określone warunki narzucone na formę przetwarzanych danych. Z tego też względu następuje częste cofanie się do fazy przygotowania danych;



Rysunek 3. Model referencyjny (iteracyjny i adaptacyjny) CRISP-DM. Źródło: [23]

— ewaluacja (ang. *Evaluation*) – na tym etapie projektu zostaje otrzymany model (lub grupa modeli). Przed przejściem do końcowego zastosowania wybranego modelu dokonywana jest gruntowna ocena otrzymanego modelu. Sprawdzane jest czy wybrane rozwiązanie spełnia wszystkie założenia ustalone w pierwszym etapie. Na końcu tej fazy zostaje podjęta decyzja co do wykorzystania wyników eksploracji danych;
— wdrożenie (ang. *Deployment*) – utworzenie modelu zazwyczaj nie stanowi końca projektu. Nawet jeśli celem modelu jest zwiększanie wiedzy o posiadanych danych, zyskana wiedza musi zostać zorganizowana i zaprezentowana w sposób użyteczny dla klienta. Zależąca od wymagań faza wdrożenia może przybierać bardzo prosty charakter i wymagać tylko wygenerowania sprawozdania. Okazać się też może bardziej złożona np. będzie wymagać wdrożenia w innym dziale. W wielu przypadkach decyzję o procesie wdrożenia podejmuje klient, nie analityk danych. Nawet jeśli analityk nie przeprowadzi wdrożenia jest ważne aby klient rozumiał i mógł właściwie wykorzystać utworzone, wdrażane modele.

Przedstawiony model na rysunku 3 opisuje kolejne fazy (ang. *phase*) procesu eksploracji danych, które realizują określone zadania (ang. *task*) i stanowią najwyższy poziom abstrakcji modelowania. Obejmuje on poszczególne stadia projektu **DM** z odpowiadającymi im zadaniami generycznymi (ang. *generic task*), którym z kolei odpowiadają zadania specjalizowane (ang. *specialized task*) wraz z powiązaniami między tymi zadaniami [7]. Przejścia pomiędzy kolejnymi etapami nie są ścisłe, jednakże pewna hierarchia przejść kształtowana jest w zależności od otrzymanych wyników z poprzedniego etapu. Wymusza to wstępną, odpowiednio dopasowaną kolejność faz [20].

Strzałki na rysunku 3 wskazują najczęstsze przejścia między etapami. Zewnętrzny okrąg na wykresie symbolizuje natomiast iteracyjną (wielokrotną) naturę procesu eksploracji da-

nych [5]. Każda pełna iteracja projektu stanowi instancję procesu (ang. *process instance*), czyli zapis przedsięwziętych akcji, podjętych decyzji i otrzymanych rezultatów. Pojedyncza instancja opisuje wyniki faktycznego działania (działań).

Na początkowym poziomie opisu projektu nie jest możliwe zidentyfikowanie wszelkich możliwych etapów. Poszczególne etapy mogą występować pomiędzy wszystkimi zadaniami wymaganymi przez Data Mining w zależności od celu, kontekstu, zamierzeń użytkownika i przede wszystkim od danych . Często rozwiązanie określonego problemu biznesowego lub badawczego prowadzi do dalszych interesujących kwestii, które można rozwiązać za pomocą tego samego ogólnego planu co poprzednio [20] dlatego też elastyczna budowa modelu **CRISP-DM** doskonale się do tego nadaje.

3. Wykorzystanie KDD do eksploracyjnej analizy tekstu

Opis spektrum zastosowań **ED** przeprowadzanej m.in. za pomocą wyżej opisanych metod i nie tylko, znacznie wykracza poza ramy niniejszego artykułu. Przedstawić można jednak kilka wyróżniających się dziedzin wykorzystujących w badaniach metody, techniki i algorytmy z zakresu **ED**. Należą do nich:

— bazy danych – w szczególności hurtownie danych i bazy multimedialne. Bazy multimedialne stanowią szczególny przypadek bazy danych, która wykorzystuje różne formy składowania i przeszukiwania informacji w celu dostarczania odbiorcom wiedzy na ich temat. Bazy multimedialne przystosowane zostały do pracy z takimi źródłami danych ja np. tekst, dźwięk, grafika, animacja, wideo. Ze względu na to na jakim typie mediów dokonywana jest **ED** wydzielone zostały osobne gałęzie tej dziedziny do których należą m.in. [24-28]: eksploracja obrazów (ang. *Picture Mining*), eksploracja tekstów (ang. *Text Mining, TM*), eksploracja wideo (ang. *Video Mining*) czy też eksploracja map cyfrowych (ang. *Spatial Mining*). Uwzględnione w tej grupie, obok multimedialnych baz danych, hurtownie danych wyróżnione zostały z tego względu iż agregują dane historyczne, które są oczyszczone w procesie **ETL** (ang. *extraction, transformation, loading*). Proces ten przyspiesza znacznie etap czyszczenia danych występujący w procesie **KDD** i **CRISP-DM** a który pochłania większość czasu przy przeprowadzaniu **ED** [29, 30]. Dlatego m.in. pod tym względem platformy te stanowią doskonałe środowisko do przeprowadzania zadań eksploracyjnych.

— Internet i związana z nim tzw. eksploracja stron internetowych (ang. *Web Mining*) – technika **ED**, mająca na celu odkrywanie i uzyskiwanie przydatnych informacji, wiedzy i wzorów z dokumentów i usług Internetowych powszechnie określanych jako *World Wide Web* (**WWW**) [31]. W obrębie tej techniki możemy wyróżnić trzy jej specjalizacje [32, 33]: eksploracja struktury stron internetowych (ang. *Web Structure Mining)*, eksploracja zawartości stron internetowych (ang. *Web Content Mining*) i eksploracja użyteczności stron (ang. *Web Usage Mining)*.

— przemysł i związany z nim tzw. *Ang. Quality Mining* - a ściślej *Ang. Quality Control Data Mining* czyli **ED** wykonywana na rzecz kontroli jakości. Polega na zgłębianiu danych w sterowaniu jakością [34]. Od zwykłej eksploracji odróżnia się m.in. tym iż występuje tutaj konieczność reagowania na zmiany w danych na bieżąco (*on-line*).

— środowisko biznesowe (*profit i no-profit*) - integruje, wykorzystuje i generuje nowe zastosowania oraz wyzwania badawcze na polu **ED** np. analiza danych on-line w systemach e-commerce, czy systemy tłumaczeń on-line w platformach e-learning wykorzystujące pośrednio techniki **TM** [35].

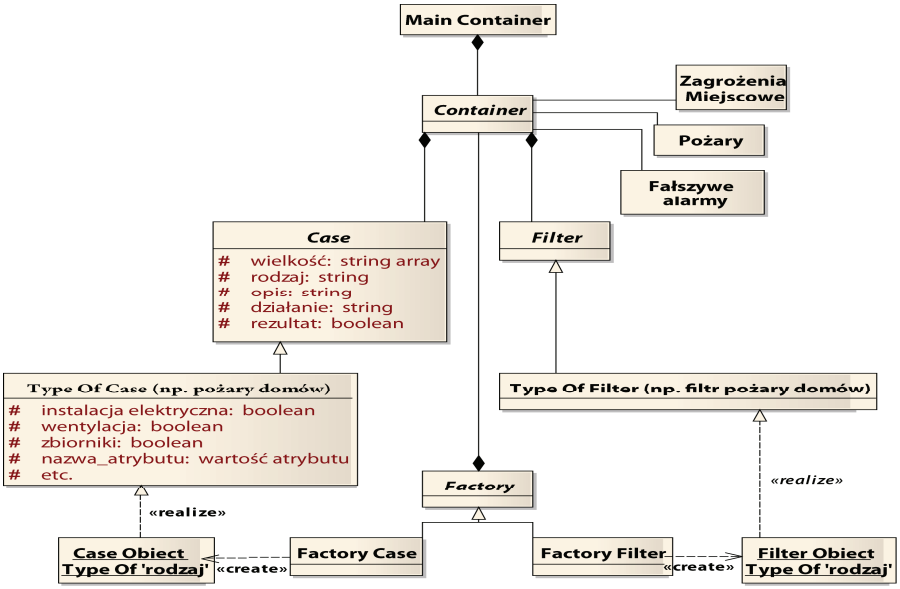
Systemy informacyjne (wiedzy) budowane w oparciu o **ED** i towarzyszące jej metody przeprowadzania nie opierają się mocno na technologii lecz na charakterystyce danych po-

chodzących z wycinka modelowanej (badanej) rzeczywistości. W obszarze zainteresowań i badań autorów leży wykorzystanie metody **KDD** do przeprowadzenia eksploracyjnej analizy tekstu (ang. *Text Mining*) stanowiącej podgrupę **ED** wydzieloną ze względu na charakterystyczny rodzaj danych tj. tekst. Analiza ta polega na wykorzystaniu inteligentnych reguł z zakresu Lingwistyki Komputerowej zajmującej się analizą języka naturalnego **NLP** (ang. *Natural Language Processing*), metod statystycznych oraz technik m.in. z zakresu przeszukiwania i grupowania danych. Wykorzystywana jest do pozyskiwania informacji (wiedzy) z dużych nieustrukturyzowanych zbiorów danych tekstowych [36]. Aktualnie w Szkole Głównej Straży Pożarnej (**SGSP**) prowadzone są badania i eksperymenty z zakresu zastosowania rozproszonych baz danych [37-39] i ontologii do opisu zdarzeń zachodzących w obrębie Państwowej Straży Pożarnej (**PSP**) [12, 40]. W ramach prowadzonych badań powstał problem dotyczący możliwości wykorzystania już istniejących danych tekstowych z bazy danych systemu ewidencji zdarzeń **EWID** [41, 42] i przekształcenia ich w wiedzę. Przeprowadzane badania kierują się w stronę utworzenia Grupowego Inteligentnego Systemu Wspomagania Decyzji (ang. *Group Intelligent Decision Support System – GIDSS*) [12]. Podstawę jego miałby stanowić tekstowy system wnioskowania na podstawie przypadków (ang. *Case Based Resoning – CBR*) [38, 43] lub System Ekspertowy (**SE**) (ang. *Expert System – ES*) oparty o reguły wydobyte za pomocą eksploracyjnej analizy tekstu. Bufor danych (wiedzy) dla wybranego systemu stanowiłaby rozproszona lub zcentralizowana, w zależności od wyników badań, multimedialna tekstowa baza danych. Dokonywać się w niej będzie zapis i odczyt np. przypadków zdarzeń zorganizowanych w postaci ontologii. Dodatkowo techniki z zakresu analizy tekstu wykorzystane zostaną do efektywnego przeszukiwania bufora wiedzy i wydobywania z niego potrzebnych w danym momencie określonych opisów przypadków zdarzeń.

Metodę **KDD** do przeprowadzenia eksploracyjnej analizy tekstu wybrano ze względu na mało skomplikowany pod względem ilości założeń i kroków, przejrzysty metodycznie model (rozdział 2.1). Umożliwia on w badaniu o małej skali i rozproszoności szybkie nakreślenie zarysu obszaru badań wraz z utworzeniem jego planu realizacji. Ponadto jest elastyczny i daje się w łatwy sposób modyfikować i dostosować do realizowanego zadania badawczego. Nawet jeśli w jego trakcie zachodzą potrzeby zmian np. pewnych założeń projektowych. Pamiętać należy przy tym, że im wyżej (głębiej) znajdujemy się w procesie **KDD** lub **CRISP-DM** to koszt powrotu do punktów poprzednich znacznie rośnie. Z tego względu ważne jest dopracowanie początkowych założeń etapów powyższych metod. Przykładowa, opisana poniżej, realizacja eksploracyjnej analizy tekstu w oparciu o **KDD** prowadzi do powstania szkieletu systemu do eksploracji tekstu (ang. *framework of system to text mining*). Podstawowa specyfikacja systemu realizowanego na potrzeby prowadzonych badań przez **SGSP** opisana jest za pomocą procesu **KDD** i realizowany jest następująco:

a) wiedza dziedzinowa i cele procesu odkrywania wiedzy – wiedzę dziedzinową stanowią m.in.: rozporządzenie Ministra Spraw Wewnętrznych i Administracji [44], system ewidencji zdarzeń **EWID**, analizy zdarzeń sporządzane po niektórych akcjach i informacje zebrane z wywiadów ze strażakami Jednostki Ratowniczo-Gaśniczej **SGSP**. Celem procesu odkrywania wiedzy jest po pierwsze dostarczenie dotąd nieznanych lub nieuwzględnionych podczas innych analiz atrybutów i ich wartości do tworzenia w zależności od wyboru: trójek Obiekt-Atrybut-Wartość (ang. *Object-Attribute-Value – OAV*), sieci semantycznych lub ontologii. Wybrana warstwa reprezentacji danych (wiedzy) ma ułatwić strukturyzację dokumentów oraz budować bazę klas systemu wyrażoną poprzez przypadki użycia w systemie **CBR** [38] (Rysunek 4) lub bazę wiedzy **SE** [12] wyrażoną poprzez inteligentne reguły. W zależności od wyboru platformy – tekstowy **CBR** lub **SE** – jeden z nich ma stanowić główną podstawę **GIDSS** w **PSP** [43]. Drugim celem reali-

zowanym po ewaluacji rozwiązania i w drugim cyklu projektowym jest dokonanie minimalizacji wyznaczonego zbioru atrybutów potrzebnych do wyszukania i określenia odpowiedniego działania. Minimalizacja ma służyć m.in. polepszeniu użyteczności (ang. *Usability*) graficznego interfejsu użytkownika (ang. *Graphic Interface User – GUI*), poprzez który użytkownik komunikuje się z systemem. Na tym etapie należy także dokonać implementacji mechanizmu przeszukiwania/wyszukiwania odpowiedniego przypadku użycia lub reguły – w zależności od zaimplementowanego rozwiązania. W razie potrzeby należy dokonać optymalizacji tego rozwiązania. W początkowych etapach analizy dotyczących pozyskiwania przypadków użycia z ww. źródeł wiedzy wykluczono takie techniki jak kategoryzacja tekstu (ang. *text categorization*) i grupowanie tekstu (ang. *text clustering*) [38]. Niemniej zaznaczyć trzeba iż techniki te mogą w drugim etapie badań stanowić pomocne narzędzia do przeprowadzania odpowiedniej optymalizacji tworzonej platformy i należałoby rozpatrzyć ich użycie.



Rysunek 4. Szkic specyfikacji obiektowej – schemat modelu klas reprezentujący podział interwencji w PSP. Źródło: [opracowanie własne]

Na Rysunku 4 przedstawiono schemat modelu klas wyrażony w języku modelowania obiektowego UML (ang. *Unified Modeling Language)* [45]. Stanowi on szkic systemu **CBR**, po zamianie wyrażenia *„Case”* na *„Reguła”* można go rozpatrywać w kontekście **SE**. Schemat stanowi formalny opis obiektowy realizowanego przedsięwzięcia i rozszerza pomysł przedstawiony w pracy [38] o pola „*działanie*” i „*rezultat*”. Znajdować się mają w nich opisy podjętych działań w celu rozwiązania zaistniałego zdarzenia oraz rezultat (wniosek, konkluzja) tych działań wrażony w postaci logicznej. Zastosowano tutaj także pomysł bliższy rozwiązaniom programowym (obiektowym) poprzez zastosowanie Fabryki Abstrakcji (ang. *Factory Abstrac*) [46] w celu generowania z wybranego Kontenera (ang. *Container*) odpowiedniego przypadku użycia lub reguły i związanego z nim filtru (ang. *Filter*) [38]. Kontener jest abstrakcją i zawiera się w Kontenerze Głównym (ang. *Main Container)*, który stanowi zbiór rodzajów interwencji **PSP**. Na Kontenery składają się: zagrożenia miejscowe, pożary i fałszywe alarmy. W każdym kontenerze przechowywane są odpowiednie dla niego zbiory klas przypadków zdarzeń lub reguł i powiązanych z nimi kolekcje filtrów;

b) selekcja danych – dokonana zostaje ze źródeł wymienionych w poprzednim kroku. W pierwszym etapie do analizy wybrane zostaną: dane tekstowe z sekcji opisowej zdarzeń znajdujących się w **EWID** oraz dane z analiz pochodzących z Komend Wojewódzkich. W drugim etapie nastąpi selekcja danych w postaci atrybutów i ich wartości opisujących zdarzenia. Selekcja będzie dokonywana już z utworzonej bazy wiedzy **SE** lub **CBR**;
c) wstępne przetwarzanie danych – podczas realizacji pierwszego opisywanego cyklu do selekcji, czyszczenia i uzupełniania danych tekstowych proponuje się kilka zabiegów z zakresu płytkiej analizy tekstu [47] m.in.: odfiltrowanie zbędnych, niemających znaczenia wyrażeń np. różnego rodzaju przyimków i zaimków, oraz określanie częstotliwości występowania form wyrazowych. Filtr zbudowany w oparciu o płytką analizę tekstu może być niewystarczający i jakość uzyskanych danych niedostateczna. Z tego też względu rozpatruje się zastosowanie niskopoziomowej (głębokiej) analizy tekstu rozszerzającej funkcjonalność filtru [48]. Do dodatkowej funkcjonalności należeć będzie m.in. wychwytywanie synonimów, antonimów, homofonów, homogramów, hiponimów, hiperonimów, geronimów, homonimów, skrótów, apostrofów, myślników, zakończeń zdań oraz możliwość przeprowadzania lematyzacji (ang. *stemming*) [48]. Wszystkie te zabiegi mają na celu utworzenie korpusu językowego, na który składałby się m.in. słownik używanych słów do opisu zdarzeń zachodzących w obrębie **PSP** oraz worek słów (ang. *bag-of-words*) [28]. Na podstawie korpusu będzie można zaimplementować odpowiednią reprezentację wiedzy np. w postaci ontologii w wybranym systemie.

   Drugi etap realizacji eksploracji danych w zakładanej bazie wiedzy nie posiada aktualnie żadnych założeń co do wstępnego przetwarzania danych. Brak założeń wynika z tego iż aktualnie nie jest dostępna pełna wiedza dotycząca atrybutów, które mogą się znaleźć w bazie wiedzy;
d) odnajdywanie i redukcja zmiennych – w przypadku analizy tekstu punkt ten jest realizowany w etapie poprzednim przy budowie filtru. Poprzez redukcję i transformację zmiennych można rozumieć np. proces lematyzacji. W drugim etapie projektowym odnajdywanie i redukcja zmiennych bezpośrednio wiąże się z celem eksploracji tj. optymalizacji zbioru przypadków/reguł opisujących zdarzenie poprzez odnajdywanie bliskiej korelacji między nimi;
e) dostosowanie procesu **KDD** do realizacji celu ustalonego w punkcie pierwszym – do analizy tekstu przewiduje się wybranie metody z zakresu streszczenia tekstu (ang. *document summarization*) [49]. Służy ona do wytwarzania streszczeń z obszernego dokumentu lub grupy dokumentów. Przykładowy algorytm bada powiązania między wyrażeniami. Jeżeli kilka wyrażeń odwołuje się do danego wyrażenia, wówczas zwiększa się jego ranking. Jako podsumowanie analizy wyświetlane jest „*n*" zdań o najwyższym rankingu, tworząc streszczenie. Druga część eksperymentu związana z optymalizacją zbioru atrybutów opisujących przypadki/reguły zakłada użycie metody ukrytego indeksowania semantycznego (ang. *Latent Semantic Indexing – LSI*) [8];
f) transformacja danych – w zależności od formy uzyskanych danych z filtru będzie zależeć to czy dane te należy przekształcić. Filtr może zostać zaimplementowany w taki sposób aby dane te automatycznie transformował do wymaganej postaci przez wyselekcjonowany algorytm;
g) eksploracja danych – fizyczna implementacja i realizacja, w wybranej technologii, zaplanowanych w poprzednich etapach zadań. Wykonywana jest ona za pomocą wyselekcjonowanych algorytmów i metod z etapu „dostosowanie procesu KDD";
h) interpretacja, analiza i prezentacja wraz z oceną wyników – krok w którym zostanie dokonane podsumowanie badań i w zależności od wyników dokonana zostanie odpo-

wiednia ewaluacja rozwiązania. Na tym etapie aktualnie autorzy nie są w stanie przedstawić wyników gdyż eksperyment jest w trakcie planowania i przygotowań do implementacji oraz realizacji. Jeśli chodzi o prezentację i interpretację wyników to rozpatrywane są takie metody jak: sieci semantyczne, histogramy czy też modyfikacja grafu strony (ang. *Websites as graphs)* za pomocą którego można by było reprezentować dokumenty i powiązane z nim wyselekcjonowane frazy;

4. Podsumowanie

W wyniku przeprowadzenia eksploracji danych za pomocą metody **CRISP-DM** jak i procesu **KDD** otrzymywane są zazwyczaj nowe interesujące informacje na temat badanego podmiotu (badanych danych). Z procesów tych często wynikają nowe, bardziej sprecyzowane pytania, które prowadzą do kolejnych iteracji eksploracji danych co umożliwia weryfikację poprzednich wyników (modeli, wzorców) [20].

Proces **KDD** można traktować jako jedną z wielu możliwych odmian modelu **CRISP-DM**. Wynika to z faktu iż **CRISP-DM** w porównaniu z procesem **KDD** jest bardziej abstrakcyjną i szerszą projektowo koncepcją w podejściu do **ED**, która posiada formalną, sprecyzowaną formę opisu modelu. Nie traktuje on jej jako jednego ze swoich kroków, lecz problem Eksploracji Danych określa jako cały cykl działań podjętych na rzecz wydobycia potencjalnie interesujących dla nas informacji. Używając języka z dziedziny programowania można stwierdzić iż **KDD** dziedziczy pewne cechy i metody **CRISP-DM**. Mimo pojęciowych różnić i podejścia do problemu, obie koncepcje są strategiami przetwarzania informacji i wiedzy w stylu - z dołu do góry (ang. *bottom-up*) i z góry do dołu (ang. *top-down*). Oba podejścia są zsyntezowane tj. składają się z prostych elementów które podlegają dalszej analizie - rozkładowi na poszczególne zadania wraz z ich szczegółowym opisem. Niezależnie od tego czy **KDD** uznamy za potomka (w koncepcji projektowania obiektowego) **CRISP-DM** czy też za oddzielny niezależny proces, to w obu podejściach kluczową rolę odgrywają metody i algorytmy Data Mining.

Bez względu na to czy do modelowania stosowany jest **CRISP-DM** czy **KDD**, oba podejścia wymuszają wstępne przetwarzanie danych w celu otrzymania poprawnych modeli przyjętych hipotez. Oba procesy, Knowledge Discovery In Databases i Cross-Industry Standard Process for Data Mining, mogą dotyczyć wydobywania informacji z danych zebranych w przeszłości, co było ich pierwotnym zadaniem. Mogą również zostać uwzględnione i zaadaptowane przy tworzeniu i projektowaniu nowych systemów (np. systemów ekspertowych, baz danych, hurtowni danych w połączeniu z *On-line Analytical Processing* [14, 50]).

Analizy zdarzeń zachodzących w obrębie **PSP** realizowane są przez różne osoby. W wyniku tego często język i forma w jakiej są one sporządzane i opisywane nie daje się bezpośrednio zastosować do przetwarzania komputerowego m.in. wnioskowania. Dlatego do pozyskiwania informacji z tego typów dokumentów słusznym zdaje się zastosowanie zaawansowanych technik oraz metod z zakresu analizy tekstu jak i metody do opisu realizacji przedsięwzięcia wyrażonego za pomocą np. **KDD.** W opisanym przypadku w wyniku zastosowania **KDD,** pośrednio dostajemy częściową specyfikację systemu wyrażoną w postaci formalnego opisu obiektowego. Służy ona do utworzenia eksploracyjnego tekstowego szkieletu aplikacji (ang. *framework application*) dającego się potencjalnie w łatwy sposób sprząc z systemem **CBR** lub **SE** w zależności od dokonanego wyboru.

Dalsze badania autorów kierować się będą w stronę budowania i implementacji w realizowanej platformie „inteligentnego" filtra danych. Filtr ten oparty o analizę tekstu umożliwi budowę korpusu językowego oraz dostarczy informacji do reprezentacji wiedzy w systemie.

Bibliografia

[1] Brilman J. *Nowoczesne koncepcje i metody zarządzania*. Wydanie 1. Polskie Wydawnictwo Ekonomiczne, 2002.
[2] Wilk-Kołodziejczyk D. *Pozyskiwanie wiedzy w sieciach komputerowych z rozproszonych źródeł informacji*. In: Lesław H.H. (red.). *Społeczeństwo informacyjne Wizja czy rzeczywistość?* [on-line] Kraków: Uczelniane Wydawnictwa Naukowo - Dydaktyczne, 2003, 30 maja. [dostęp: 16.11.2007] http://winntbg.bg.agh.edu.pl/skrypty2/0095/285-295.pdf.
[3] Fayyad U., Piatetsky-Shapiro G., Smyth P. *From Data Mining to Knowledge Discovery in Databases*. AI Magazine, 1996.
[4] Piatetsky-Shapiro G., Frawley J. W. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.
[5] *CRISP-DM*. [on-line] [dostęp: 1.06.2008] http://www.crisp-dm.org/.
[6] *Metodologia Data Mining – model referencyjny CRISP-DM*. [on-line] [dostęp: 01.06.2008] http://www.spss.pl/konsulting/konsulting_datamining_metodologia.html.
[7] 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: 01.06.2008] http://www.crisp-dm.org/CRISPWP-0800.pdf.
[8] Hand D., Mannila H., Smith P. *Eksploracja danych*. Wydanie 1. Warszawa: Wydawnictwo Naukowo-Techniczne, 2005.
[9] Fayyad U. M., Piatetsky-Shapiro G., Smyth P. *From Data Mining to Knowledge Discovery: An Overview*. AAAI Press/MIT Press, s. 1-36.
[10] Hand D., Maninila H., Smyth P. *Principles of Data Mining*. Cambridge: MIT Press, 2001.
[11] Zhou F., Yang B., Li L., Chen Z. *Overview of the New Types of Intelligent Decision Support System*. Innovative Computing Information and Control, No 1(10), 2008, s. 267-267.
[12] Mirończuk M., Karol K. *Koncepcja systemu ekspertowego do wspomagania decyzji w Panstwowej Strazy Pozarnej*. W: Grzech A., Juszczyn K., Kwaśnicka H., Nguyen N.T., (red.). *Inżynieria Wiedzy i Systemy Ekspertowe*. Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2009.
[13] Morzy T. *Eksploracja danych: problemy i rozwiązania*. [on-line] [Zakopane]: V Konferencja PLOUG 1999 – Integracja danych i systemów informatycznych, 1999, 12-16 października. [dostęp: 16.11.2007] http://www.ploug.org.pl/konf_99/pdf/7.pdf.
[14] Zakrzewicz M. *Data Mining i odkrywanie wiedzy w bazach danych*. [on-line] [Zakopane]: Konferencja PLOUG'97 - Sieciowy system informacji o projektach badawczych, 1997. [dostęp: 16.11.2007] http://www.cs.put.poznan.pl/mzakrzewicz/presentations/ploug97.pdf.
[15] Houston A. L., Chen H., Hubbard S. M., Schatz B. R., Ng T. D., Sewell R. R., et al. *Medical Data Mining on the Internet: Research on a Cancer Information System*. Artificial Intelligence Review, No 13, 1999, s. 437-466.
[16] Clinton B. *New York University speech, Salon.com* [on-line]. [dostęp: 01.06.2008] http://www.salon.com/politics/feature/2002/12/06/clinton/print.html.
[17] Frawley J. W., Piatetsky-Shapiro G., Matheus C. *Knowledge Discovery in Databases: An Overview*. AI Magazine, No 13(3), 1992, s. 57-70.
[18] Janicki M., Ślęzak D. *Data mining w praktyce. Wprowadzenie w tematykę* [on-line]. [dostęp: 01.06.2008] http://www.qed.pl/WstepDM.html.

[19] Kobos M. *Data Mining. Przegląd Eksploracji danych* [on-line]. [dostęp: 01.06.2008] http://www.mini.pw.edu.pl/~mandziuk/23-11-05.pdf.
[20] Larose T. D. *Odkrywanie wiedzy z danych*. Wydanie 1. Warszawa: Wydawnictwo Naukowe PWN, 2006.
[21] Berson A., Smith S., Thearling K. *Building Data Miting Applications for CRM*.
[22] Żytniewski M. *Budowa modeli dla analitycznych systemów CRM* [on-line]. [dostęp: 01.06.2008] http://www.mini.pw.edu.pl/~mandziuk/23-11-05.pdf.
[23] *Crisp*. [on-line] [dostęp: 01.06.2008] http://www.spss.pl/dodatki/obrazki/crisp.gif.
[24] Jia J. W., Wang J. Z., Li J., Lin S. C. *Evaluation Strategies For Automatic Linguistic Indexing Of Pictures*. Proc IEEE Int Conf Image Processing, 2003.
[25] Kotsiantis S., Kanellopoulos D., Pintelas P. *Multimedia mining*. WSEAS Transactions on Systems, No 3, 2004, s. 3263-3268.
[26] Quack T., Ferrari V., Gool L. V. Gool. *Video mining with frequent itemset configurations*. In Proc CIVR: Springer, 2006. p. 360-369.
[27] Santos M., Amaral L. *Knowledge Discovery in Spatial Databases through Qualitative Spatial Reasoning*. Portugal, 2000. [dostęp: 05.05.2009] http://repositorium.sdum.uminho.pt/bitstream/1822/5584/1/PADD2000_MS_LA.pdf.
[28] Solka J. L. *Text Data Mining: Theory and Methods*. Statistic Survey.
[29] Ansari S., Kohavi R., Mason L., Zheng Z. *Integrating e-commerce and Data Mining: Architecture and Challenges.*
[30] Kohavi R., Provost F. *Applications of Data Mining to Electronic Commerce* Kluwer Academic, 2001.
[31] Szełemej Ł. *Przegląd metod ekstrakcji wiedzy w serwisach WWW – Web Structure Mining*.
[32] Kosala R., Blockeel H. *Web Mining Research: A Survey*. SIGKDD Explorations, No 2, 2000, s. 1-15.
[33] Staś T. *Wykorzystanie technik ewolucyjnych w procesie nowoczesnej personalizacji portali internetowych. Studia i materiały polskiego stowarzyszenia zarządzania wiedzą*. Bydgoszcz: PSZW, 2007.
[34] Demski T. *Data Mining w sterowaniu procesem (QC Data Mining)*. StatSoft Polska. [dostęp: 20.09.2008] http://www.statsoft.pl/czytelnia/jakosc/sixqcminer.pdf.
[35] Naoki Y. S., Morikawa M. H., Fellow T. I., Shi Y. *Open Smart Classroom: Extensible and Scalable Learning System in Smart Space Using Web Service* Technology. IEEE Transactions on Knowledge and Data Engineering, No 6(21), 2009.
[36] Rajman M. *Text Mining – Knowledge Extraction from Unstructured Textual Data*. 6th Conference of International Federation of Classification Societies (IFCS-98), 1998.
[37] Krasuski A. *Rozproszona baza danych − możliwości wykorzystania w PSP*. Przegląd Pożarniczy, No 5, 2006, s. 30−33.
[38] Krasuski A., Maciak T. *Wykorzystanie rozproszonej bazy danych oraz wnioskowania na podstawie przypadków w procesach decyzyjnych państwowej straży pożarnej*.
[39] Krasuski A., Maciak T. *Rozproszone bazy danych w Państwowej Straży Pożarnej − model systemu*. In: Kozielski T., editor. Bazy danych, technologie, narzędzia. Warszawa WKŁ, 2005. p. 135−142.
[40] Krenski K., Maciak T., Krasuski A. *An overview of markup languages and appropriateness of XML for description of fire and rescue analyses*. *Zeszyty Naukowe SGSP*, No 37, 2008, s. 27-39.
[41] *Abakus: System EWID99*. [on-line] [dostęp: 01.05.2009] http://www.ewid.pl/?set=rozw_ewid&gr=roz.

[42] *Abakus: System EWIDSTAT*. [on-line] [dostęp: 01.05.2009] http://www.ewid.pl/?set=ewidstat&gr=prod.
[43] Mirończuk M., Maciak T. *Projekt Grupowego Inteligentnego Systemu Wspomagania Decyzji dla Państwowej Straży Pożarnej*. Zeszyty Naukowe SGSP, preprint (2009).
[44] Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 29 grudnia 1999 r. w sprawie szczegółowych zasad organizacji krajowego systemu ratowniczo-gaśniczego. Dz.U.99.111.1311 § 34 pkt. 5 i 6.
[45] Fowler M. *UML Distilled: A Brief Guide To The Standard Object Modeling Language*. Wydanie 3. Addison-Wesley Professional 2004.
[46] Gamma E. *Wzorce Projektowe*.
[47] Sołdacki P. *Wprowadzenie do eksploracji tekstu i technik płytkiej analizy tekstu*. [on-line] [dostęp: 01.03.2009] http://www.icie.com.pl/ZISI/Soldacki_Text_Mining.ppt.
[48] Makowiecka A. *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*. Warszawa: PJWSTK, 2007.
[49] Stergos A., K. Vangelis, Panagiotis S. *Summarization from medical documents*. Artificial Intelligence in Medicine, No 2, 2005, s. 157−177.
[50] Goil S., Choudhary A. *A Parallel Scalable Infrastructure for OLAP and Data Mining*. IEEE Computer Society, 1999.