Proces i metody eksploracji danych tekstowych do przetwarzania raportów z akcji ratowniczo-gaśniczych

Marcin Mirończuk[1], Tadeusz Maciak[2]

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

Abstract:

This paper describes the process for processing reports from rescue and firefighting. To reports processing methods and techniques used in the field of textual data mining (text mining). This paper also presents the classification and analysis methods section of text which is considered a potential use in the proposed process.

Keywords:

text processing, text analysis, text mining, information extraction, sentence classification

1. Wstęp

W Państwowej Straży Pożarnej PSP forma powstających raportów sporządzanych po każdej akcji ratowniczo-gaśniczej jest regulowana przez Rozporządzenia Ministra Spraw Wewnętrznych [1]. Forma raportów papierowych w postaci karty *Informacji ze zdarzenia* jest częściowo ustrukturyzowana. Częściowa strukturyzacja polega na tym, że istnieje możliwość wprowadzenia i sprawdzenia informacji o takich elementach akcji ratowniczo-gaśniczej jak np.: czas zdarzenia, czas działań ratowniczych, rodzaj prowadzonych działań i użytego sprzętu, miejsca prowadzonych działań, dane o budynku lub pomieszczeniu, w którym powstało zdarzenie etc. Większość tych elementów jest ograniczona do zamkniętej listy odpowiedzi, przy czym dla większości pytań są tylko dwie odpowiedzi i tylko w jednym punkcie (rodzaj prowadzonych działań) dostępnych możliwości jest ponad czterdzieści. Z pośród nich można dokonywać wyboru np. uwalnianie ludzi, ewakuacja ludzi, ewakuacja zwierząt etc. W przypadku czasu zdarzenia jest to data w postaci dzień-miesiąc-rok oraz godziny-minuty. Większość użytecznej informacji zawarta jest jednak w sekcji zatytułowanej – *Dane opisowe do informacji ze zdarzenia*. Sekcja ta podzielona jest na sześć podpunktów: opis przebiegu działań ratowniczych (zagrożenia i utrudnienia, zużyty i uszkodzony sprzęt), opis jednostek przybyłych na miejsce zdarzenia, opis tego co uległo zniszczeniu lub spaleniu, warunki atmosferyczne, wnioski i uwagi wynikające z przebiegu działań ratowniczych oraz inne uwagi dotyczące danych wypełnianych w formularzu odnośnie zdarzenia. Ze względu na to, że zawartość poszczególnych podpunktów tej sekcji jest wyrażona za pomocą języka naturalnego w postaci zdań, na które składają się wyrażenia oraz frazy, została ona nazwana *częścią półustrukturyzowaną*. W sekcji tej ukryta jest większość informacji i wiedzy w postaci procesu, procedur i danych na temat tego jak np. skutecznie zwalczać powstałe zagrożenie lub na co należy uważać i zwrócić uwagę przy likwidacji danego rodzaju zagrożenia. W raportach pozostaje ukryta informacja (wiedza) o zdarzeniach zawarta w wykonywanych przez Kierujących Działaniami Ratowniczymi (KDR) opisach przebiegu akcji ratowniczo-gaśniczej nieuwzględniona przez ww. akt normatywny.

Na bazie poszczególnych raportów w Komendach Wojewódzkich PSP wykonywane są wybrane analizy zdarzeń i składowane w postaci papierowej. W Komendzie Głównej PSP

raporty są analizowane przez analityków pod kątem określonych strategicznych zapytań. Raporty przechowywane są również w wersji elektronicznej w informacyjnym systemie ewidencji zdarzeń EWID [2-4]. Mają one także częściowo ustrukturyzowany charakter ze względu na to, iż sekcje oraz pola z *Karty informacji ze zdarzenia* są mapowane i przedstawiane w postaci relacji i odpowiednich typów danych. Jednak w dalszym ciągu sekcja *Dane opisowe do informacji ze zdarzenia* jest reprezentowana za pomocą tekstu opisanego językiem naturalnym. Możliwości analizy tej części pogarsza fakt, że sześć wcześniej wymienionych podpunktów składających się na tą sekcję w wersji papierowej, w systemie informacyjnym ewidencji zdarzeń reprezentowanych jest jako pojedynczy rekord danych bez zachowania należytego podziału. Z tego względu ta cyfrowa sekcja stanowi *część nieustrukturyzowaną.* Ponadto w wyniku tego, że opisy zdarzeń wykonywane są przez różnych KDR powstaje problem semantyczny, każdy z KDR definiuje i opisuje zdarzenie według własnego postrzegania i słownictwa. Powoduje to iż do określenia tych samych zdarzeń stosowane są różne nazwy. Badania wykazują, że przy opisywaniu jednego zagadnienia jedynie 20% badanych posługuje się tym samym słownictwem [5]. Obserwacja ta nie zmienia się znacząco w zależności czy badanymi są eksperci w danej dziedzinie czy też mniej doświadczone osoby.

Ewentualne pozyskanie ze zbioru takich rekordów informacji i przekształcenia ich zawartości do *użytecznych przypadków zdarzeń* systemu wnioskowania na podstawie zdarzeń (*ang. case based reasoning – CBR*) [6], mogącego stanowić podstawę hybrydowego systemu wspomagania decyzji HSWD [7], wymaga więc zastosowania wielu zabiegów semantycznych. Dzięki uzyskanym *użytecznym przypadkom zdarzeń*, poprzez zabiegi semantyczne wyrażone za pomocą skonstruowanego procesu do przetwarzania sekcji *Dane opisowe do informacji ze zdarzenia*, HSWD będzie mógł generować wiedzę dla KDR w postaci np. wskazówek, opisu ewentualnych problemów zdarzających się podczas podobnej akcji, lokalizacji punktów czerpania wody (środka gaśniczego) etc. W oparciu o nią KDR będzie mógł następnie sprawnie pokierować akcją ratowniczo-gaśniczą. HSWD postrzegany jest jako element definiujący współdziałanie oraz wykorzystanie sił i środków poprzez KDR na podstawie uzyskanej wiedzy z systemu wiedzy. Do składowania, wydobywania, zaadaptowania i ponownego wykorzystania wiedzy przez KDR w HSWD ma służyć wcześniej wspomniany podsystem CBR. Realizuje on technikę rozwiązywania aktualnie pojawiających się problemów na podstawie doświadczenia i wiedzy z przeszłości opisanych za pomocą *użytecznych przypadków zdarzeń*, które stanowią wskazówki lub rozwiązania zagrożeń z zakresu np. pożarów lasów, zagrożeń miejscowych.

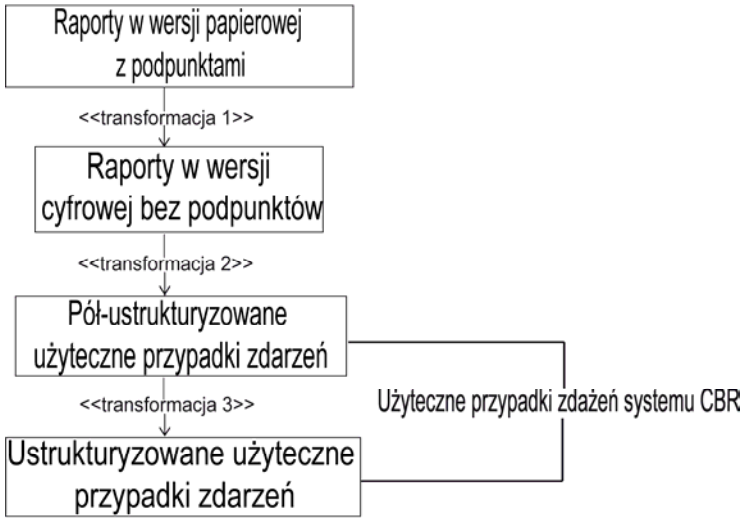
W PSP brak jest standardowego, ujednoliconego, szerokiego słownika zawierającego pojęcia z zakresu ratownictwa, który także definiowałby zachodzące między nimi relacje oraz stanowiłby ontologie dla służb ratowniczych. Słownik taki posłużyłby do utworzenia precyzyjniejszej, homogenicznej komunikacji i wymiany wiedzy na temat zdarzeń z zakresu ratownictwa w obrębie PSP. Ponadto taka różnorodna interpretacja i opis podobnych wypadków powoduje iż pozyskiwanie informacji z tego typu dokumentów tj. sekcji opisowej systemu EWID i transformacja ich bezpośrednio do ustandaryzowanego opisu *użytecznych przypadków zdarzeń* systemu CBR nie jest do końca możliwa i wymaga zastosowania technik z zakresu komputerowej analizy tekstu.

Proponowane rozwiązanie do przetwarzania sekcji *Dane opisowe do informacji ze zdarzenia* (*część nieustrukturyzowana*) bazuje na analizie części składowych tej sekcji, które stanowią zdania. Ze względu na to, że sekcja ta a więc i zdania są wyrażone za pomocą języka naturalnego, problem analizy można uogólnić na problem eksploracyjnej analizy dokumentów tekstowych. Tym samym można zastosować znane już metody i techniki z tego zakresu.

W punkcie 2 niniejszego artykułu przedstawiono zarys procesu eksploracji danych tekstowych do przetwarzania raportów z akcji ratowniczo-gaśniczej (nieustrukturyzowanej sekcji *Dane opisowe do informacji ze zdarzenia*). W punkcie 3 opracowania został opisany preferowany w badaniach model wektorowy reprezentacji dokumentów tekstowych. W punkcie tym przedstawiono także utworzoną taksonomię wybranych metod z zakresu eksploracyjnej analizy tekstu do przetwarzania raportów. Podsumowanie całości wraz z opisem kierunku rozwoju i badań zostało przedstawione w punkcie 4.

2. Szkielet eksploracyjnego procesu do przetwarzania raportów ze zdarzeń ratowniczo-gaśniczych

W pracach [6, 8] zostało zasugerowane, że metody eksploracyjnej analizy tekstu mogą służyć do budowania *użytecznych przypadków zdarzeń* systemu CBR. W niniejszym punkcie autorzy zaprezentowali możliwości wykorzystania metod z zakresu eksploracji danych tekstowych, opisanych w punkcie 3, do rozwiązywania problemów związanych ze strukturalizowaniem sekcji *Dane opisowe do informacji ze zdarzenia* oraz budowaniem *użytecznych przypadków zdarzeń* a więc też z doborem odpowiedniego słownictwa w celu jednoznacznego opisu zdarzeń oraz związanych z nim elementów. Elementami tymi mogą być np. punkty czerpania wody (hydranty). Ideowy schemat proponowanego przez autorów procesu budowy *użytecznych przypadków zdarzeń* systemu CBR, będącego częścią składową HSWD, przedstawia rysunek 1.



Rysunek 1. Proces strukturalizacji przypadków zdarzeń systemu CBR stanowiącego podsystem w HSWD. Źródło: [opracowanie własne]

Rysunek 1 przedstawia proces strukturalizacji dostępnych raportów w celu przekształcenia ich w *użyteczne przypadki zdarzeń* systemu CBR stanowiącego podsystem w HSWD. Jak wspomniano na początku artykułu, sekcja *Dane opisowe do informacji ze zdarzenia* w wersji papierowej posiada użyteczne podpunkty, które mogą zostać wykorzystane we wnioskowaniu na temat procesu likwidacji powstałego zagrożenia lub budowania na ich podstawie nowych ustrukturalizowanych rejestrów np. punktów czerpania wody. Niestety podpunkty z tej sekcji przenoszone są do systemu informacyjnego EWID często bez zachowania kolejności oraz obowiązującej sześcio punkowej struktury. Kłopotliwe i nieużyteczne jest ich wykorzystanie np. trudno jest otrzymać, na podstawie opisu samego zdarzenia, oddzielnych informacje tylko i wyłącznie na temat np.: zagrożeń i utrudnień, podjętych dzia-

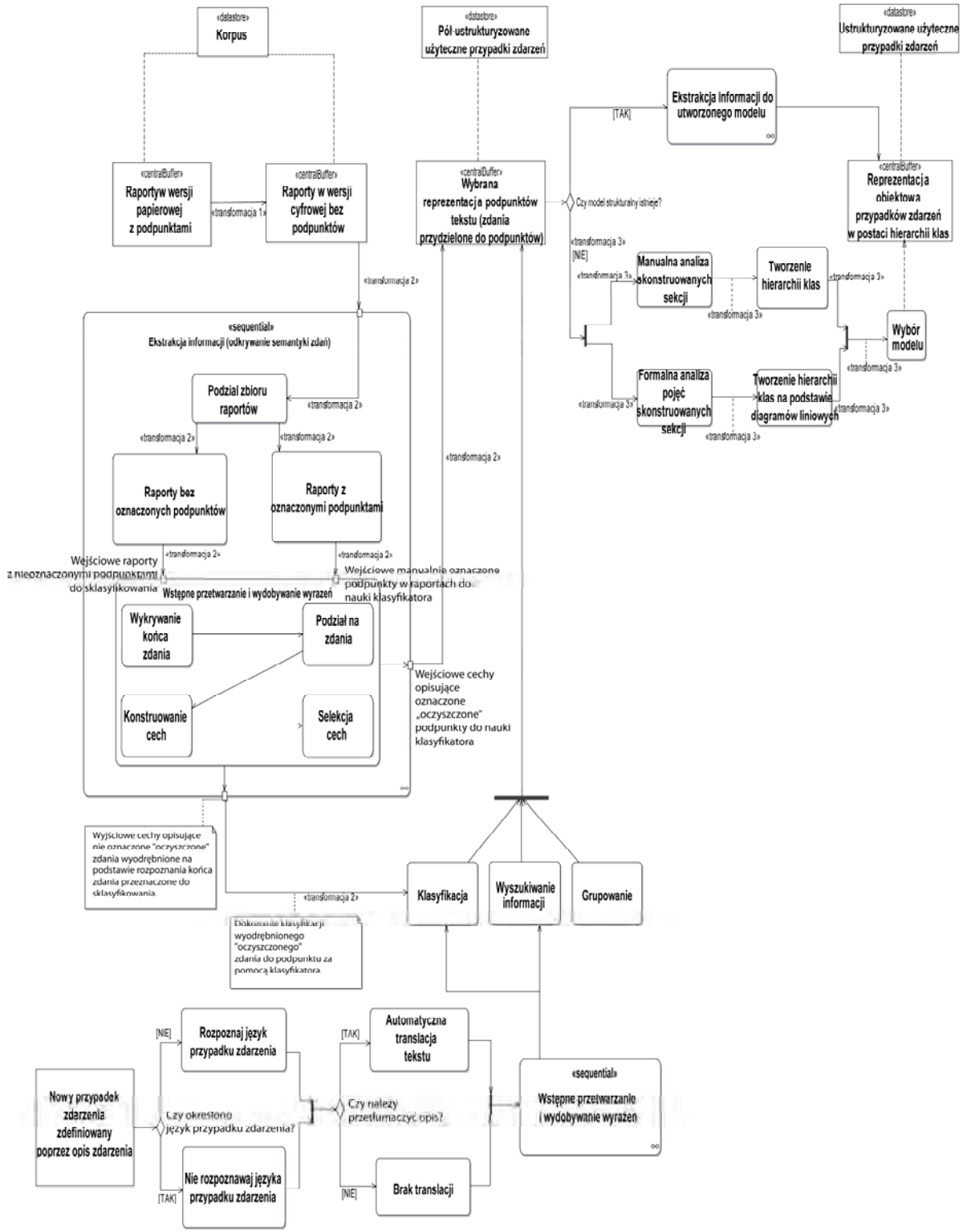
łań, dostępnych punktów czerpania wody, użytego sprzętu, przyczyn zagrożenia czy też wskazówek. Aktualnie, w wyniku przeszukiwania cyfrowej sekcji *Dane opisowe do informacji ze zdarzenia* systemu EWID, KDR otrzymałby cały raport (w ogólności ich grupę), który musiałby przeanalizować a następnie wybrać z niego interesujące go treści. Taka sytuacja nie może mieć miejsca, ze względu na ograniczony czas, który decyduje o powodzeniu całej akcji ratowniczo-gaśniczej. Z tego też względu autorzy proponują nowe podejście do analizy i przetwarzania raportów z akcji ratowniczo-gaśniczych. Podejście to polega na badaniu oraz wykrywaniu znaczenia i funkcji zdania w sekcji *Dane opisowe do informacji ze zdarzenia* na podstawie występujących wyrażeń w zdaniu. Autorzy ten sposób analizy określają mianem *odkrywania semantyki zdań*. Analiza ta ma na celu częściowe odtworzenie struktury tej sekcji z wersji papierowej lub w ogólności odtworzenie zadanej struktury. W ten sposób mają zostać utworzone *pół-ustrukturyzowane użyteczne przypadki zdarzeń*, które likwidują aktualne, wyżej wymienione problemy z użytecznością. W dalszej kolejności *pół-ustrukturyzowane użyteczne przypadki zdarzeń* mogą być transformowane do *ustrukturyzowanych użytecznych przypadków zdarzeń*. Za ich pomocą będzie można dokładnie parametryzować zdarzenia lub ich elementy i opisywać je za pomocą wyselekcjonowanych atrybutów i klas a nie wyłącznie poprzez język naturalny (często mało precyzyjny i dwuznaczny). Schemat procesu eksploracji danych tekstowych do przetwarzania raportów z akcji ratowniczo-gaśniczej przedstawia rysunek 2.

Proces przedstawiony na rysunku 2 składa się z trzech głównych etapów przetwarzania. Pierwszy etap przetwarzania raportów zakłada ich strukturalizację poprzez odkrywanie semantyki zdań (podpunkt 2.1). Drugi etap zakłada budowę modeli wybranych podpunktów wraz z ekstrakcją do nich danych (podpunkt 2.2). Trzeci etap aktualnie związany jest z wykorzystaniem *pół-ustrukturyzowanych użytecznych przypadków zdarzeń* do wyszukiwania opisów podobnych akcji ratowniczo-gaśniczych i sposobu neutralizacji powstałego zagrożenia (podpunkt 2.3).

2.1. Odkrywanie semantyki zdań

Rysunek 2 przedstawiający proponowany szkielet aplikacji jest znacznym rozszerzeniem szkieletu systemu do analizy biznesowej kierowanej tekstem (*ang. text-driven business intelligence*), którego idea został przedstawiona w pracy [9]. Autorzy niniejszego opracowania zaadaptowali i rozszerzyli tą strukturę na potrzeby analizy raportów z akcji ratowniczo-gaśniczych. Za pomocą analizy raportów w pierwszym kroku mają zostać otrzymane *pół-ustrukturyzowane użyteczne przypadki zdarzeń* wyrażone za pomocą wektorowej reprezentacji podpunktów tekstu. W drugim kroku ma zostać utworzona reprezentacja obiektowa *użytecznych przypadków zdarzeń* wyrażona w postaci utworzonej hierarchii klas. Do otrzymania danych pół-ustrukturyzowanych proponowana jest metoda klasyfikacji. Raporty z dostępnego cyfrowego korpusu EWID dzielone są na dwie grupy. Pierwsza grupa zawiera raporty z oznaczonymi ręcznie *a priori* nazwami podpunktów, do których należą zdania z raportu. Nazwy podpunktów stanowią zarazem etykiety klas, do których wybrany klasyfikator będzie przydzielał zdania. Podczas badań i manualnej analizy dostępnych raportów wydzielono cztery takie klasy: operacje, sprzęt, szkody, meteo, ogólna. Druga grupa zawiera resztę nieoznakowanych (niezaetykietowanych) raportów, z których należy wydobyć podpunkty poprzez przydzielenie zdań do odpowiedniej zdefiniowanej powyżej klasy. Następnie zachodzi proces wstępnego przetwarzania i wydobywania wyrażeń z oznakowanego zbioru raportów, po którym następuje proces nauki i testowania klasyfikatora. Przed procesem klasyfikacji nieoznakowanych zdań z grupy raportów zawierających nieoznakowane podpunkty, należy także dokonać procesu wstępnego przetwarzania i wydobywania wyrażeń. W szczególności należy

wykryć koniec zdań w raporcie a następnie podzielić go na zdania. Dalszy proces wydobywania wyrażeń przebiega podobnie jak podczas nauki klasyfikatora, a więc wykonywany jest krok konstruowania cech i selekcji cech. Po tym etapie dokonywany jest proces klasyfikacji zdania do jednej z wyżej zdefiniowanych klas za pomocą zbudowanego klasyfikatora.



Rysunek 2. Schemat procesu eksploracji danych tekstowych do przetwarzania raportów z akcji ratowniczo-gaśniczej. Źródło: [opracowanie własne na podstawie [9]]

Dla zilustrowania działania etapu odkrywania semantyki zdania przy założeniu, że dostępny jest klasyfikator posłużono się następującym przykładem – niech będzie dostępny nieustrukturyzowany raport w postaci:

Nieustrukturyzowany raport 1

Po dojechaniu na miejsce zdarzenia stwierdzono pożar instalacji elektrycznej w skrzynce z bezpiecznikami na klatce schodowej. Działania psp polegały na oddymieniu i przewietrzeniu klatki schodowej na parterze. Na miejsce zdarzenia przybyło pogotowie energetyczne z ul. chrzanowskiego celem zabezpieczenia instalacji. Sprawny hydrant nr 34922 ul. Szaserów 99.

Nieustrukturyzowany raport 1 poddany zostaje wstępnemu przetwarzaniu i wydobywaniu wyrażeń. Podzielony zostaje na cztery zdania z których usuwane są wyrażenia znajdujące się na stop liście. Tak przetworzone zdania poddaje się następnie procesowi lematyzacji oraz konstruowania i selekcji cech. Ostatni etap kończy się reprezentacją wektorową każdego z czterech zdań, które klasyfikowane są do utworzonych klas. W wyniku przeprowadzenia klasyfikacji otrzymywany jest *pół-ustrukturyzowany użyteczny przypadek zdarzenia*. Rezultat opisanych operacji prezentuje tabela 1.

Tabela 1 Pół-ustrukturyzowany użyteczny przypadek zdarzenia. Źródło: [opracowanie własne]

L.p.	Zdanie oryginalne	Reprezentacja wektorowa	Klasa
1	*Po dojechaniu na miejsce zdarzenia stwierdzono pożar instalacji elektrycznej w skrzynce z bezpiecznikami na klatce schodowej*	[dojechać, miejsce zdarzenia, stwierdzić, pożar, instalacja elektryczna, skrzynka, bezpieczniki, klatka schodowa]	Opisowa
2	*Działania psp polegały na oddymieniu i przewietrzeniu klatki schodowej na parterze*	[działania psp, polegać, oddymienie, przewietrzyć, klatka schodowa, parter]	Operacji
3	*Na miejsce zdarzenia przybyło pogotowie energetyczne z ul. chrzanowskiego celem zabezpieczenia instalacji*	[miejsce zdarzenie, przybywać, pogotowie energetyczne, ul. chrzanowskiego, cel, zabezpieczać, instalacja]	Opisowa
4	*Sprawny hydrant nr 34922 ul. Szaserów 99*	[sprawny, hydrant, nr. 34922, ul. szaserów 99]	Sprzętu

Tabela 1 prezentuje wynik opisanego procesu przetwarzania oraz klasyfikacji zdań z nieustrukturyzowanego meldunku 1, którego efektem jest *pół-ustrukturyzowany użyteczny przypadek zdarzenia*. Przypadek ten składa się z oryginalnych zdań, ich reprezentacji wektorowej oraz klasy, do której zostały przydzielone. Można zauważyć, że podczas etapu konstruowania i selekcji cech oprócz pojedynczych wyrażeń zostały utworzone złożone wyrażenia tzw. n-gramy składające się z dwóch lub większej ilości wyrażeń [10]. Związane jest to z tym, że niektóre formy wyrażeniowe występują częściej razem niż inne w meldunkach np. *miejsce zdarzenia*, *działania psp* etc. W opisach występują także bardziej złożone wyrażenia określające np. położenie – *ul. Szaserów 99*.

W omawianym przykładzie, w wyniku procesu klasyfikacji, zostały przywrócone cztery klasy semantyczne dla wyodrębnianych zdań, które określają ich kontekst a tym samym redukują niejednoznaczność, która może zajść podczas etapu wyszukiwania [11]. Podczas procesu wyszukiwania np. sprawnych hydrantów w celu uzupełnienia środka gaśniczego podczas akcji ratowniczo-gaśniczej znajdujących się przy ulicy *Szaserów* system może zwrócić nieoczekiwane rezultaty. W przypadku gdy brak jest ww. podziału na zdania, raporty analizowane są całościowo a więc rezultaty wyszukiwania mogą zawierać nie tylko hydranty znajdujące się przy tej ulicy ale też np. opisy inny działań, które tam się odbywały. Strukturalizacja raportów poprzez nadanie klas semantycznych dla poszczególnych składo-

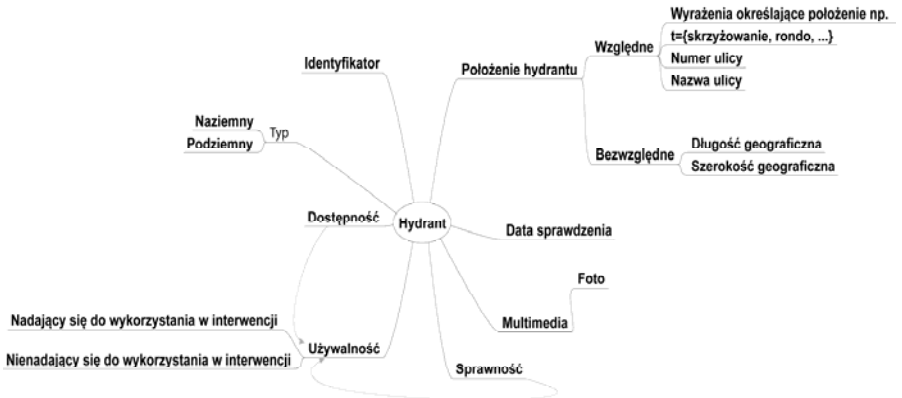
wych w postaci zdań ogranicza kontekst wyszukiwania. Posiadając do dyspozycji klasę sprzęt i wyszukując w niej sprawnych hydrantów przy podanej ulicy, system wyszukiwania zwróci nam tylko i wyłącznie znajdujące się przy niej hydranty.

2.2. Budowa wybranych modeli i ekstrakcja informacji

W przypadku gdy zostanie zbudowana baza *pół-ustrukturyzowanych użytecznych przypadków*, przy wykorzystaniu wyżej opisanego procesu, istnieje możliwość analizy zebranych tam informacji. Analiza ta prowadzi do dalszej ich strukturyzacji w kierunku opisu obiektowego, reprezentowanego za pomocą hierarchii klas. Autorzy do tego celu proponują dwie metody. Pierwsza metoda mniej formalna pod względem matematycznym opiera się o mapy myśli (*ang. mind mapping*) i transformacji ich do modelu obiektowego. Za pomocą tej metody autorzy dokonali modelowania wyrażeń znajdujących się w zdaniach zaklasyfikowanych do klasy *sprzęt*. Druga metoda opiera się o formalną analizę pojęć (*ang. formal concept analysis – FCA*), która stanowi pewien rodzaj metody grupowania oraz kraty pojęć (*ang. concept lattice)*. Druga metoda została wykorzystana do modelowania wyrażeń pochodzących z klasy *opis* i transformacji utworzonych diagramów liniowych do modelu obiektowego. W punkcie tym do analizy tekstu wykorzystano więc elementy z zakresu inżynierii wiedzy oraz oprogramowania.

2.2.1. Manualne tworzenie modeli

Na podstawie przeprowadzonej manualnej analizy zdań pochodzących z wybranej klasy istnieje możliwość utworzenia mapy myśli [12]. Można je zaliczyć do technik wizualizacyjnych. Metodę tą autorzy wykorzystali do utworzenia mapy myśli związanej z pojęciem *Hydrant* w celu zbudowania modelu rejestru punktów czerpania wody. Następnie tak utworzoną mapę poddano transformacji do opisu obiektowego tj. wyrażono ją za pomocą hierarchii klas. W celu zamodelowania ww. rejestru analizowane były wyrażenia opisujące pojęcie *Hydrant* znajdujące się w zdaniach pochodzących z klasy *sprzęt*. Na podstawie tak przeprowadzonej analiz utworzona została mapa pojęć, którą prezentuje rysunek 3.



Rysunek 3. Mapa pojęć związana z konceptem Hydrant pochodzącym ze zdań zaklasyfikowanych do sekcji sprzęt. Źródło: [opracowanie własne na podstawie [13]]

Rysunek 3 przedstawia mapę pojęć związanych z głównym konceptem *Hydrant*. Z konceptem tym związane są wyrażenia występujące w zdaniach z klasy *sprzęt*. Informacje te można zaprezentować w bardziej ustrukturyzowany sposób przydatny do przetwarzania komputerowego np. za pomocą modelu obiektowego. Możliwą transformację mapy pojęć z zakresu inżynierii wiedzy do modelu obiektowego z inżynierii oprogramowania przedstawia rysunek 4.

Rysunek 4. Hierarchia klas w notacji obiektowej. Źródło: [opracowanie własne]

Rysunek 4 przedstawia jedną z wielu możliwych hierarchii klas z zakresu projektowania obiektowego, która stanowi realizację mapy pojęć. Podstawowy koncept *Hydrant* został zamodelowany za pomocą klasy głównej *Hydrant,* po której większość atrybutów jest dziedziczona przez dwie klasy potomne *Naziemny* oraz *Podziemny*. Dodatkowo wydzielono klasę *Położenie* agregowaną przez klasę *Hydrant* w celu możliwości jej wykorzystania do opisu położenia ewentualnych innych obiektów a nie tylko hydrantów. Instancje pochodzące z tak utworzonej hierarchii klas można utrwalać bezpośrednio w rejestrze (systemie) obiektowym [14] lub posługując się odpowiednimi narzędziami do mapowania w systemie katalogowym lub relacyjnych [11, 15, 16].

2.2.2. Pół-automatyczne tworzenie modeli

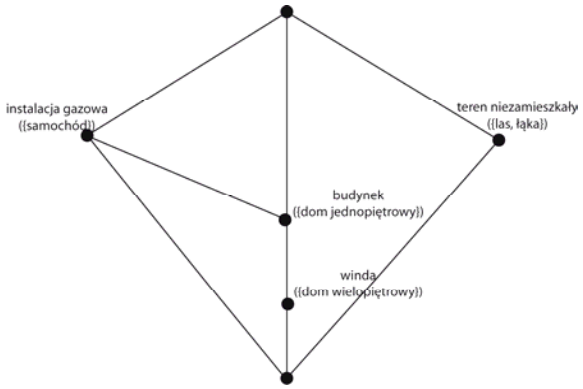
Formalna analiza pojęć wprowadzona została przez Rudolfa Willea w 1984 roku. Jej koncepcja zbudowana została na teorii sieci i częściowego porządku, które to zostały rozwinięte przez Birkhoff i innych w latach 30 XIX wieku [17-19]. FCA służy do matematyzacji określenia *Pojęcie* (określane także jako *Koncept*) oraz daje formalne narzędzie stosowane do analizy danych i reprezentacji wiedzy. Do wizualizacji zachodzących relacji pomiędzy wykrytymi pojęciami służy w FCA krata pojęć. Krata pojęć graficznie może być zaprezentowana za pomocą diagramu liniowego (*ang. line diagram*) nazywanego także diagramem Hassego (*ang. Hasse diagram*) [20, 21]. Diagram ten służy do konstruowania hierarchii pojęć. Składa się on z węzłów (wierzchołków) oraz krawędzi. Każdy wierzchołek reprezentuje pojęcie natomiast krawędzie służą do połączenia wierzchołków w określony sposób [20]. Aktualnie FCA stosowana jest w dziedzinach z zakresu m.in. [17]: psychologii, socjologii, antropologii, medycynie, biologii, lingwistyki, matematyki czy też informatyki.

Formalna analiza pojęć jest jedną z wielu metod wykorzystywanych w inżynierii wiedzy do odkrywania i budowania ontologii specyficznej dla rozważanej dziedziny z danych tekstowych [22-24]. Autorzy zaproponowali wykorzystanie tej analizy do pół-automatycznego odkrywania i budowania ontologii ze zdań zaklasyfikowanych do wybranej klasy. Proponowana analiza jest procesem dwustopniowym. W pierwszej kolejności wydobywana jest krata pojęć. W drugim kroku analizy, utworzona krata pojęć transformowana jest do opisu obiektowego. W celu omówienia i zademonstrowania pół-automatycznego tworzenia modelu wybranego podpunktu założono, że do dyspozycji dane jest pięć raportów, z których wyodrębniono zdania z klasy opisów zdarzeń (ekstensje). Dotyczą one np. pożaru: lasu, domu jednopiętrowego, samochodu, łąk oraz domu wielopiętrowego. Przykładowo też wstępne przetwarzanie i wydobywanie wyrażeń mogło wykazać, że najlepiej opisującymi te przypadki są następujące wyrażenia (intensje): teren niezamieszkały, budynek, instalacja gazowa, winda. Związek pomiędzy poszczególnymi ekstensjami i intensjami prezentuje tabela 2.

Tabela 1 opisuje jakie atrybuty posiada określony rodzaj pożaru. Jeśli pożar posiada dany atrybut wówczas ten fakt odnotowywany jest za pomocą symbolu „X”, w przeciwnym razie pole pozostaje puste. Na podstawie tak zdefiniowanych relacji można utworzyć kratę pojęć. Utworzona krata dla danych z tabeli 2 prezentuje rysunek 5.

Tabela 2. Tabela opisujące dane składające się na formalny kontekst „Pożar”.
Źródło: [opracowanie własne]

Pożar	teren niezamieszkały	budynek	instalacja gazowa	winda
Lasu	X			
Domu jednopiętrowego		X	X	
Samochodu			X	
Łąk	X			
Domu wielopiętrowego		X	X	X



Rysunek 5. Krata pojęć formalnego kontekstu Pożar.
Źródło: [opracowanie własne na podstawie [25]]

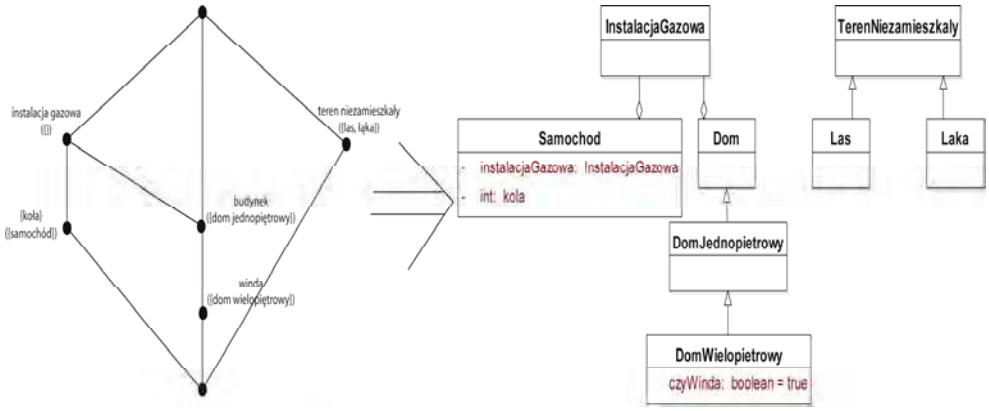
Krata pojęć zawiera węzły, linie oraz nazwy wszystkich obiektów wraz z atrybutami dla zdefiniowanego formalnego kontekstu (formalny kontekst stanowią *Pożary*). Węzły diagramu reprezentują koncepcje. Informacja o kontekście może być bezpośrednio odczytana z diagramu za pomocą prostej reguły: obiekt *o* posiada atrybut *a* tylko i tylko wtedy, kiedy istnieje ścieżka do góry np. dom wielopiętrowy posiada atrybut: winda, budynek oraz instalacja gazowa. Oznacza to, że dom wielopiętrowy stanowi rozszerzenie domu jednopiętrowego. Kontynuując te wnioskowanie dochodzi się do punktu, w którym dom wielopiętrowy jest także rozszerzeniem samochodu. Jest to wniosek w rzeczywistości błędny. W celu skorygowania tego należy dodać np. atrybut koła, który rozgraniczy samochody od domów.

Jak widać na powyższym przykładzie dodatkowym atutem diagramów liniowych jest to, iż wiedza w łatwy sposób może zostać skorygowana i dodana przez eksperta z danej dziedziny np. kierującego akcją ratowniczą lub wyspecjalizowanego inżyniera wiedzy. W rozpatrywanym przykładzie aktualizacja wiedzy polegała na dodaniu rozgraniczającego atrybutu „koła”, który w rozważanej hipotetycznej sytuacji nie został odkryty podczas etapu selekcji cech.

Od skonstruowanej i przedstawionej kraty pojęć blisko jest do hierarchii klas z notacji obiektowej. Wykorzystując wiedzę z zakresu inżynierii oprogramowania, inżynier na podstawie kraty pojęć może przekształcić go w diagram hierarchii klas obrazujący model analizowanej dziedziny w notacji obiektowej. Wynik takiego przekształcenia, przy dodatkowym założeniu, że inżynier wiedzy wprowadził dodatkowy atrybut rozdzielający *koła*, przedstawia rysunek 6.

Rysunek 6 przedstawia przekształcenie, które nie jest jednoznaczne tzn. każdy inżynier w zależności od znajomości dziedziny, doświadczenia i wiedzy dotyczącej projektowania i inżynierii oprogramowania może utworzyć różne hierarchie klas. Niemniej takie podejście

do analizy daje mu możliwość zaznajomieniem się ze słownictwem (terminologią, konceptami) oraz z podstawowymi obiektami i atrybutami występującymi w modelowanej dziedzinie, które może wyrazić za pomocą hierarchii klas.



Rysunek 6. Przekształcanie kraty pojęć w hierarchię klas w notacji obiektowej.
Źródło: [opracowanie własne]

2.2.3. Ekstrakcja informacji

Zgodnie z rysunkiem 2 przewidywana jest dwustopniowa ekstrakcja informacji. Pierwszy stopień związany jest z ekstrakcją segmentów i określaniem ich semantyki. Określanie semantyki zdań odbywa się poprzez ich przydział do wyznaczonych klas za pomocą metod klasyfikacji dostępnym w eksploracyjnej analizie tekstu. Stopień ten został omówiony w podpunkcie 2.1. Drugi poziom stanowi ekstrakcja informacji ze zdań o wydzielonym kontekście do utworzonego modelu. Ekstrakcja informacji na poziomie raportów polega na rozpoznaniu scenariusza i powiązaniu go z abstrakcyjnym modelem, który odzwierciedla złożoność zdarzeń w rzeczywistości. Odkryte klasy semantyczne zdań oraz kompleksowy *pół-strukturalny użyteczny przypadek zdarzenia* omówiony w podpunkcie 2.1, w kontekście ekstrakcji informacji nazywa się skryptem lub scenariuszem [26]. W przypadku analizowanej dziedziny abstrakcyjnym modelem jest utworzony rejestr punktów czerpania wody bądź ontologia *Pożary* wyrażona w notacji obiektowej. W jednym jak i drugim przypadku modelu zadanie ekstrakcji informacji będzie polegać na rozpoznawaniu nazw encji.

Ze względu na to, że w obu przypadkach rozpoznawanie nazw encji przebiega podobnie, zademonstrowany zostanie przykład dla modelu punktów czerpania wody. Przyjmując, że dostępny jest wyżej wspomniany model utworzony w podpunkcie 2.2.1 można do niego wyekstrahować dane ze zdań zaklasyfikowanych do klasy *sprzęt*. W tym celu należy przeprowadzić proces rozpoznawania nazw encji. Do zilustrowania tego można posłużyć się prostym przykładem. Niech będzie dostępne zdanie pochodzące z klasy sprzęt opisujące obiekt w postaci hydrantu w następującej postaci:

Sprawny hydrant nr 34922 ul. Szaserów 99.

Z wyżej przytoczonego zdania można wyekstrahować do rozpatrywanego modelu następujące dane (atrybut, wartość, opis):

- atrybut używalność przyjmuje wartość prawdy logicznej (*ang. true*), wyrażenie *sprawny* w danym zdaniu sugeruje, że hydrant działa,
- atrybut numer identyfikacyjny przyjmuje wartość 34922, które występuje w danym zdaniu,

- atrybut położenie względne może przyjąć wartość strukturalną złożoną np. z dwóch dodatkowych atrybutów w postaci nazwy ulicy o wartości *Szaserów* i jej numeru *99*. Obie wartości zostały także pozyskane z prezentowanego zdania.

Jak zademonstrowano na przykładzie, ekstrakcja encji polega na rozpoznawaniu i klasyfikowaniu wykrytych wyrażeń z tekstu takich jak: nazwy ulic, identyfikatory, stan obiektów etc. do utworzonego modelu.

2.3. Przetwarzanie pół-ustrukturyzowanych użytecznych przypadków zdarzeń

Dodatkowymi ostatnimi elementami, które prezentuje rysunek 2, są składniki które w szczególnym przypadku mogą służyć do rozpoznawania i translacji przypadków zdarzeń. Elementy te mogą zostać użyte w sytuacji działań transgranicznych np. symulacji działań ratowniczo-gaśniczych i kooperacji jednostek z różnych krajów. W takiej sytuacji system po rozpoznaniu tego, że opis zdarzenia nie jest sporządzony w aktualnie używanym języku, tłumaczy go, po czym przeszukuje bazę przypadków w celu odnalezienia najlepszego rozwiązania. Wyniki wyszukiwania, jeśli zachodzi potrzeba, tłumaczone są ponownie i zwracane odbiorcy. W sytuacji, gdy nie zachodzi potrzeba identyfikacji języka oraz translacji przypadku zdarzenia, aplikacja działa jako wyszukiwarka opisu podobnych zdarzeń. Wyszukiwanie w proponowanych modelach może być dwojakiego rodzaju. W przypadku poszukiwania rozwiązań wyszukiwanie może zostać oparte na zapytaniach pełno tekstowych do wydzielonych klas lub mogą zostać podane instancje klas zdefiniowane w ramach ontologii, z którymi związane są odpowiednie rozwiązania umieszczone w sekcji *operacje*. W przypadku odnajdywania punktów czerpania wody, wyszukiwanie powinno odbywać się w oparciu o zamodelowany rejestr i instancje klasy *Hydrant*.

Wykorzystany element grupujący, znajdujący się na rysunku 2, związany jest z formalną analizą pojęć. Niemniej rozpatrzone mogą zostać inne algorytmy do grupowania *pół-ustrukturyzowanych przypadków zdarzeń*. Ich przydatność może być badana pod kątem analizy jednorodności utworzonych klas, ewentualnie wyznaczania dodatkowych podklas.

3. Metody eksploracji danych tekstowych

Dziedzina techniki zajmująca się przetwarzaniem komputerowym nieustrukturyzowanych danych w postaci dokumentów tekstowych i wyciągania z nich informacji wysokiej jakości nazywa się eksploracją tekstu (*ang. text mining*) [27, 28]. W obrębie tej dziedziny powstało wiele nie do końca usystematyzowanych metod, technik oraz pojęć, które dla celów podjętych badań w niniejszym artykule zostały odpowiednio pogrupowane i szczegółowo omówione. W ramach badań utworzono więc autorską taksonomię metod eksploracji danych tekstowych. Omówienie jej zaczęto od przedstawienia wykorzystywanej w badaniach reprezentacji dokumentów tekstowych (podpunkt 3.1). Następnie opisano wykorzystywane metody analizy tekstu (podpunkt 3.2) oraz techniki wizualizacji (podpunkt 3.3).

3.1. Reprezentacja dokumentów tekstowych

Aktualnie rozwinięte i wykorzystywane praktycznie są dwie reprezentacje dokumentów tekstowych: reprezentacja wektorowa oraz reprezentacja grafowa. Ze względu na to, że w badaniach oraz do reprezentacji *pół-ustrukturyzowanych użytecznych przypadków zdarzeń* wykorzystuje się reprezentację wektorową została ona dokładnie omówiona w podpunkcie 3.1.1. Informacje na temat reprezentacji grafowej można znaleźć w publikacjach [29-32].

3.1.1. Model wektorowy reprezentacji dokumentów tekstowych

Model wektorowy reprezentacji dokumentów tekstowych polega na przedstawieniu ich w postaci przestrzenno-wektorowego opisu (modelu wektorowego, *ang. vector space model – VSM*). Dokumenty i występujące w nich wyrażenia są reprezentowane w postaci macierzy. Powszechnie za wyrażenie w reprezentacji przestrzenno-wektorowej uważane jest jedno wyrażenie np. *pożar* lub para wyrażeń np. *mocne zadymienie*. Zazwyczaj nie są to wszystkie możliwe wyrażenia – zwykle w etapie wstępnego przetwarzania (*ang. preprocessing*) dokonuje się ich selekcji (za pomocą metod opisanych w podpunkcie 3.2.1.1 i 3.2.2.1) oraz oceny ich istotności dla modelowanej dziedziny.

Rysunek 7 przedstawia macierzową postać zapisu dokumentów i związanych z nimi wyrażeń. Dokumenty reprezentowane są poprzez wiersze (m), natomiast wyrażenia znajdują się w kolumnach (n) macierzy A zwanej macierzą dokumentów-wyrażeń (*ang. term-document matrix*). Bardziej ogólnym pojęciem stosowanym w lingwistyce komputerowej jest *korpus* określający dużą kolekcję dokumentów, opisanych i sprowadzonych np. w szczególnym przypadku do opisywanej postaci macierzowej, którą przedstawia rysunek 7.

$$A = \begin{bmatrix} w_{11} & . & . & . & w_{1j} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ w_{i1} & . & . & . & w_{ij} \end{bmatrix}, A \in R^{m \times n}$$

$$Gdzie:$$

$$1 \le i \le m$$

$$1 \le j \le n$$

Rysunek 7. Struktura reprezentacji przestrzenno-wektorowej dokumentów [33].
Źródło: [opracowanie własne]

W rozwiązaniach praktycznych ilość wierszy macierzy A jest znacznie większa od ilości wyrażeń (m >> n). Do poprawy przetwarzania, wydajniejszego składowania takiej struktury w systemach informatycznych i analizy stosuje się konwencję odwróconą tj. w wierszach zapisywane są wyrażenia natomiast w kolumnach dokumenty. Wówczas taki zapis nosi nazwę *pliku odwróconego* a jego sposób indeksowania wyrażony jest poprzez *indeks odwrotny* [34, 35]. Element macierzy w_{ij} oznacza wagę, a tym samym znaczenie j-tego wyrażenia w i-tym dokumencie (rysunek 7 reprezentuje taki zapis). W zależności od sposobu kodowania informacji zawartej w elemencie w_{ij} czyli w wadze wyrażenia lub bardziej precyzyjnie w wartościach składowych wektora wyrażeń, istnieje możliwość otrzymania różnych odmian reprezentacji przestrzenno-wektorowej tekstu. Do popularnych, stosowanych w praktyce odmian, zaliczamy m.in. reprezentacje boolowską (binarną), częstotliwościową występowania wyrażeń (*ang. term frequency – TF*), odwrotną częstość dokumentu (*ang. inverse-document-frequency – IDF*), mieszaną TF-IDF, logarytmiczną, ważoną logarytmiczną, okapi BM25 oraz probabilistyczną [10, 33, 36-38]. Reprezentacja:

a) boolowska (binarna) – występuje wówczas, kiedy zostanie odnotowany fakt zaistnienia j-tego wyrażenia w i-tym dokumencie, natomiast nie precyzuje ona liczby wystąpień. Element w_{ij} macierzy A przyjmuje wartość 1 (j-te wyrażenie znajduje się w i-tym dokumencie) lub 0 (j-te wyrażenie nie znajduje się w i-tym dokumencie),

b) częstotliwościowa występowania wyrażeń (*ang. term frequency – TF*) – występuje wówczas, kiedy oprócz odnotowania faktu zaistnienia j-tego wyrażenia w i-tym dokumencie zostanie określona także jego częstość, czyli liczba jego wystąpień w zadanym dokumencie,

c) odwrotnej częstości dokumentu (*ang. inverse-document-frequency – IDF*) – polegającą na tym, iż poszczególne wagi w_{ij} wyrażone są za pomocą wyrażenia $\log(N/n_j)$, gdzie: N reprezentuje liczbę wszystkich dokumentów zaś n_j liczbę dokumentów z j-tym wyrażeniem,
d) mieszana TF-IDF – występuje wówczas, gdy pomnożone zostaną przez siebie wagi w_{ij} wyrażone za pomocą ww. schematu TF i IDF, czyli mieszana reprezentacja TF-IDF równa jest $TF \cdot IDF$,
e) logarytmiczna – występuje wówczas, gdy następuje zastąpienie wszystkich niezerowych elementów macierzy A wartościami w_{ij} równymi $1+\log(w_{ij})$,
f) ważona logarytmiczna – występuje wówczas, gdy następuje zastąpienie wszystkich niezerowych elementów macierzy A wartościami w_{ij} obliczonymi za pomocą następującej formuły $(1+\log(w_{ij}))\cdot\log(\frac{N}{n_j})$,
g) okapi BM25 – stosowana jest w przypadkach długich dokumentów tekstowych, gdzie prawdopodobieństwo, że dany wyraz pojawi się wiele razy jest wysokie. Powoduje to wzrost wartości wagi *TF* co w efekcie sprawia, że długie dokumenty są bardziej „faworyzowane". BM25 to rodzina funkcji wykorzystywana do obliczenia wagi w_{ij} z uwzględnieniem długości dokumentów. Mając dokument d (od angielskiego słowa *document*) i wyrażenie t (od angielskiego słowa *term*) można obliczyć wagę korzystając z zależności:

$$bm25(d,t) = idf \cdot \frac{f(t,d)\cdot(k_1+1)}{f(t,d)+k_1\cdot(1-b+b\cdot\frac{|d|}{avg(d)}} \tag{1}$$

Gdzie:
- f(t,d) – liczba wystąpień wyrażenia t w dokumencie d
- |d| – długość dokumentu d
- avg(d) – średnia długość dokumentu w kolekcji
- k_1 i b – wartości stałe (przeważnie przyjmuje się $k_1 = 1.2$ i $b = 0.75$)
- idf – zmodyfikowany schemat IDF wyrażony w postaci formuły $idf(t) = \log\frac{N-n(t)+0.5}{n(t)+0.5}$, N oznacza liczbę wszystkich dokumentów, a $n(t)$ liczbę dokumentów zawierających wyrażenie t.

h) probabilistyczna – występuje wówczas, gdy waga w_{ij} wyrażenia t w dokumencie d zostanie oszacowana na podstawie zdarzenia losowego, polegającego na wystąpieniu danego wyrażenia t w dokumencie d pod warunkiem modelu M. Model M zawiera informacje na temat korpusu A tj. całkowitą ilość wyrażeń oraz częstotliwość występowania poszczególnych wyrażeń w korpusie A. Proste oszacowanie prawdopodobieństwa wystąpienia wyrażenia t można dokonać zgodnie z zasadą estymacji największej wiarygodności (*ang. maximum likelihood estimation – MLE*) [36]:

$$w_{ij} = P_{ML}(Y_i = t \mid d, M) = \frac{TF_{t,d}}{\sum_{t\in d} TF_{t,d}} \tag{2}$$

Wzór na wagę wyrażenia przedstawiony w postaci estymacji największej wiarygodności można interpretować następująco:

$$P_{ML}(Y_i{=}t|d,M){=}\frac{\text{czestotliwosc wystepowania wyrazenia t w dokumencie d}}{\text{suma wszystkich czestotliwosci wyrazen t w dokumencie d}} \tag{3}$$

Na podstawie macierzy A z odpowiednio skonstruowanymi wagami w_{ij} możliwe jest więc wyznaczenie podobieństwa słów oraz dokumentów. Podobieństwo słów wyrażane jest poprzez określenie podobieństwa odpowiadających im kolumn tej macierzy, natomiast o podobieństwie dokumentów wnioskuje się na podstawie analizy podobieństwa wierszy tej

macierzy. Najczęściej wszystkie wagi w_{ij} wektorów macierzy A w zastosowaniach praktycznych są normalizowane do 1.

Wprowadzenie wektorowo-przestrzennego modelu dokumentów umożliwia matematyczną analizę zagadnienia np. wyszukiwania dokumentów tekstowych. Zagadnienie wyszukiwania zostało omówione w podpunkcie 3.2.2.2. Budowanie reprezentacji tekstu na samych wyrażeniach jest jednak często mocno ograniczone. Do zasadniczych wad tego modelu należą:

- utrata wszelkiej informacji na temat struktury dokumentów: tytuł, nagłówki etc.,
- pominięcie informacji na temat kolejności wyrażeń a więc i związków między nimi (występowanie wyrażeń jest niezależne od siebie),
- istnieje konieczność wyboru wyrażeń, dla których zostanie stworzona macierz – liczba wymiarów musi być z góry znana.

Ze względu na ww. ograniczenia, proponowane są takie rozwiązania aby składowymi wektora reprezentującego dokument były automatycznie wydobyte cechy tekstu (jak język, styl, itp.) zamiast wyrażeń kluczowych oraz elementy wydobyte z semantycznego zbioru, a więc wyrażenia i powiązania między nimi skoncentrowane na stronie znaczeniowej tekstu [5]. Ograniczenia związane z reprezentacją przestrzenno-wektorową wymogły stosowanie drugiego sposobu reprezentacji dokumentów tekstowych, a mianowicie ich opis grafowy. Należy zaznaczyć, że reprezentacja przestrzenno-wektorowa mimo ograniczeń posiada też zalety. Sprawiają one, że jest ona dalej powszechnie stosowaną i badaną reprezentacją dokumentów tekstowych. Zaletą wyboru takiej reprezentacji dokumentów jest to, że jest zbieżna z reprezentacją stosowaną typowo w uczeniu maszynowym (obiekty opisane za pomocą atrybutów), dzięki czemu można do niej zastosować istniejące metody. Badania także dowodzą, że niektóre relacje semantyczne mogą zostać wydobyte z tekstu z dużą dokładnością z pominięciem kolejności słów [39], natomiast związek pomiędzy wyrażeniami może zostać ustalony za pomocą analizy współwystępowania wyrażeń [40]. Również wykonywanie operacji, takich jak: liczenie odległości, przeprowadzanych na wektorach, jest aplikacyjnie łatwiejsze w realizacji i bardziej efektywne obliczeniowo od konkurencyjnej reprezentacji, np. opartej na modelu grafowym.

3.2. Metody analizy tekstu

Metody płytkiej analizy tekstu można podzielić ze względu na to czy do ich działania potrzebna jest sformalizowana reprezentacja dokumentu opisana w podpunkcie 3.1 czy też nie. Przykład sformalizowanej reprezentacji tekstu stanowi reprezentacja wektorowa opisana w podpunkcie 3.1.1 oraz grafowa. Niesformalizowana reprezentacja natomiast nie wymaga żadnej z powyższych reprezentacji.

3.2.1. Metody analizy bezpośredniej na tekście

Metodami, które nie wymagają sformalizowanej reprezentacji tekstu, są: wstępne przetwarzanie tekstu, ekstrakcja informacji, automatyczne rozpoznawanie języka, automatyczna translacja tekstów. Metody te zostały omówione w kolejnych sekcjach niniejszego podpunktu.

3.2.1.1. Techniki wstępnego przetwarzania dokumentów tekstowych

W eksploracyjnej analizie tekstu dostępne są dwie metody przetwarzania tekstu: płytkie i głębokie. Pierwsza metoda dotycząca płytkiej analizy tekstu (*ang. shallow text processing – STP*), określa grupę działań polegających 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. text document classification lub text document categorization*) ich grupowania (*ang. text dokument clustering*) i wyszukiwania z nich informacji (*ang. information retrieval – IR*) [5, 41, 42]. Celem tej analizy jest przyporządkowanie nieustrukturyzowanego tekstu wyrażonego za pomocą języka naturalnego do ustalonej reprezentacji (zazwyczaj składającej się ze zbioru obiektów). Przyporządkowanie to odbywa się na drodze procesu wykorzystującego specyficzne dla danej dziedziny algorytmy [42]. 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 [10, 42].

Do technik wstępnego przetwarzania dokumentów tekstowych należą: ekstrakcja rdzeni wyrażeń (*ang. stemming*), tagowanie (*ang. tagging*), lematyzacja, stop lista oraz przycinanie (*ang. pruning*) [10, 41]. Operacje te podejmowane są zanim dokument lub grupa dokumentów tekstowych zostanie przesłana do głównego procesu analizy np. wyszukiwania pełnotekstowego (*ang. full text serach*) [43] czy też innych metod przetwarzania tekstu. Przedstawione terminy, związane ze wstępnym przetwarzaniem tekstu, można zdefiniować w następujący sposób [10, 41]:

a) ekstrakcja rdzeni wyrażeń – określa znajdowanie tematów słów lub tych ich fragmentów, które są niezmienne dla wszystkich form,

b) tagowanie – oznacza wybór opisu morfo-składniowego, który jest właściwy w konkretnym kontekście użycia danej formy,

c) lematyzacja – jest to analiza morfologiczna ograniczana do znalezienia podstawowej formy wyrazu (identyfikacja leksemu),

d) usuwanie słów ze stop listy – na stop liście umieszcza się wyrażenia, które występują zbyt często, by ich użycie jako kluczy wyszukiwania było celowe. Wyrażenia umieszczone na stop liście słów są odrzucane (filtrowane) podczas wczytywania dokumentu,

e) przycinanie – polega na usuwaniu niepotrzebnych słów, operacja ta ma na celu polepszenie skuteczności klasyfikacji. Można usuwać wyrażenia występujące najczęściej (*ang. most frequent*) i najrzadziej (*ang. least frequent*).

Wszystkie wyżej wymienione zabiegi stosuje się w celu ulepszenia przeprowadzanej analizy dokumentów tekstowych oraz ich wydajniejszego indeksowania. Zabiegi te stosowane w kontekście analizy tekstu pozwalają na identyfikację początkowego zestawu cech, który może być później ograniczony (i zoptymalizowany) w procesie wydobywania wyrażeń (podpunkt 3.2.2.1).

3.2.1.2. Ekstrakcja informacji

Ekstrakcja informacji (*ang. information extraction – IE*) jest to identyfikacja, polegająca na odnajdywaniu właściwej informacji w nieustrukturyzowanych danych tekstowych wyrażonych za pomocą języka naturalnego. Proces ten jest zgodny z klasyfikacją polegającą na strukturyzowaniu poprzez nadawanie klas semantycznych dla wybranych elementów tekstu. Proces ten czyni informację zawartą w tekście bardziej właściwą i przydatną w realizowanych zdaniach [26]. Ekstrakcja informacji nazywana jest także ekstrakcją (rozpoznawaniem)

encji i modelowania ich relacji (*ang. concept/entity extraction, named entity recognition*) [44], jednak jest to ograniczenie definicji ekstrakcji informacji tylko do jednego z podstawowych jej zadań. Wymienione zadanie polega na pozyskiwaniu z dokumentów tekstowych nazw obiektów np. osób oraz na wyznaczaniu związków i relacji pomiędzy wydobytymi obiektami. W ogólnym przypadku można pozyskiwać w ten sposób z tekstu nazwy miast, imiona i nazwiska osób, kody pocztowe, numery PESEL itp. W przypadku szczególnym, który stanowią analizy raportów z akcji ratowniczo-gaśniczych, można pozyskać informacje na temat: ilości akcji, w których brała udział dana osoba, ilości ofiar śmiertelnych zarejestrowanych w akcji ratunkowej. Przy pomocy tak wydobytych cech można sprawdzać czy analizowany obiekt np. osoba nie zmieniła rangi (nie awansowała na wyższy stopień), czy nie zaszły jakieś kluczowe zmiany na obiekcie np. niedziałające hydranty, czy też w przestrzeni mediów nie pojawiły się informacje o zdarzeniach określonego typu (katastrofy, wypadki, akty terrorystyczne). Do pozostałych podstawowych zadań z zakresu ekstrakcji informacji należą: rozróżnianie wyrażeń rzeczownikowych z relacją gramatyczną (*ang. noun phrase coreference resolution*), rozpoznawanie ról semantycznych (*ang. semantic role recognition*), rozpoznawanie relacji między encjami (*ang. entity relation recognition*) czy też rozpoznanie czasu oraz określanie linii czasu zachodzenia zdarzeń (*ang. timex and time line recognition*) [26].

Do typowych problemów, które muszą być rozwiązane przez system ekstrakcji informacji należą, następujące zagadnienia [10, 26]:

a) rozpoznanie i utworzenie skryptów (scenariuszy) będących kompleksowym opisem zdarzeń,
b) utworzenie modeli (wzorców) wynikających z tekstu,
c) podział tekstu na ciągi zdań,
d) podział zdań na wyrażenia z przypisanymi wartościami cech gramatycznych,
e) rozpoznawanie skrótów, fraz rzeczownikowych, nazw bez wnikania w ich strukturę wewnętrzną i ich funkcje w zdaniu,
f) budowanie przybliżonej struktury zdania (np. drzewa rozbioru) ze słów i wcześniej rozpoznanych elementów,
g) wypełnienie przygotowanych modeli informacjami z tekstu.

Zadania z punktów c) – f) mają charakter ogólny i ich rozwiązania mogą być stosowane w wielu różnych systemach. Natomiast dwa pierwsze zadania z punktów a) – b) oraz ostatnie zadanie z punktu g) jest ściśle związane z konkretnym zastosowaniem. Wzorce i reguły ich wypełniania zależą od tego, jakich informacji poszukujemy.

Przytoczone wyżej pojęcia ekstrakcji informacji wiążą się najczęściej z normalizacją i identyfikacją w tekście wybranych typów danych oraz ich powiązań. Niemniej w skład tej metody można zaliczyć podejścia i zabiegi stosowane do wydobywania wyrażeń (cech) reprezentatywnych, od jakości których zależą np. wyniki wyszukiwania informacji dla dokumentu czy też ich grupy. W kontekście analizy tekstu i niniejszego opracowania cecha (*ang. feature*) znaczeniowo traktowana jest jako wyrażenie (*ang. term*). W dalszej kolejności, oprócz samego wydobywania wyrażeń, można też ekstrahować semantykę tych wyrażeń za pomocą np. analizy opartej o dane z korpusu lingwistycznego (reprezentacji przestrzenno-wektorowej dokumentów) [32]. Ogólnie do obu tych celów mogą służyć metody klasyfikujące lub grupujące opisane w podpunkcie 3.2.2, jeżeli zadanie klasyfikacji lub grupowania zostanie zdefiniowane na mniejszym poziomie ziarnistości niż dokument, a mianowicie na poziomie wyrażeń.

3.2.1.3. Automatyczne rozpoznawanie języka

Automatyczne rozpoznawanie języka (*ang. automatic language identification – ALI*) polega na identyfikacji wersji językowej dokumentu, w szczególności dokumentu tekstowego, który może zostać napisany w więcej niż jednym języku [45]. Do automatycznej identyfikacji wersji językowej wykorzystywane są głównie dwa rodzaje rozwiązań. Pierwsze rozwiązanie bazuje na statystycznym modelu języka i polega na oszacowaniu prawdopodobieństwa (*ang. estimate the probability*), że dana wejściowa próbka tekstu jest napisana w zadanym języku. Drugie rozwiązanie polega na porównaniu pomiędzy częstotliwością używanych wspólnych słów lub wyrażeń w próbce tekstowej z częstotliwością wydobytą ze statystycznej analizy dużego korpusu służącego jako odniesienie.

Automatyczne rozpoznawanie języka wykorzystywane jest najczęściej w sieci internetowej do analizowania wersji językowych stron internetowych (*ang. World Wide Web – WWW*), czy też korespondencji email. Pewne jego elementy mogą też być wykorzystane we wstępnym procesie tekstowej eksploracji danych w celu polepszenia jakości analizy.

3.2.1.4. Automatyczna translacja tekstów

Automatyczna translacja tekstów nazywana także tłumaczeniem maszynowym TM, polega na dokonywaniu przekładu z jednego języka na drugi. Pierwsze próby TM były podejmowane w latach 50-tych. W latach 70-tych dziedzina ta przeżyła swój rozkwit ze względu na gwałtowny rozwój sprzętu jak i oprogramowania komputerowego. Do automatycznego tłumaczenia tekstu podchodzi się dwojako tj. dokonuje się tłumaczenia „zgrubnego", przeznaczonego do poprawiania przez człowieka (mamy tutaj do czynienia raczej ze wspomaganiem tłumaczenia, a nie z samym tłumaczeniem) oraz tłumaczenia ograniczonego do wąskiego podzbioru języka (np. prognozy pogody, raportów giełdowych) [10].

Największym problemem w tłumaczeniu i kluczem do jego sukcesu jest prawidłowe tłumaczenie słów a raczej ich znaczeń. Mimo pojawiających się problemów związanych z TM, w dalszym ciągu budzi ono wielkie zainteresowanie zarówno w środowisku naukowców jak i biznesowym [46-48].

3.2.2. Metody analizy sformalizowanych reprezentacji tekstu

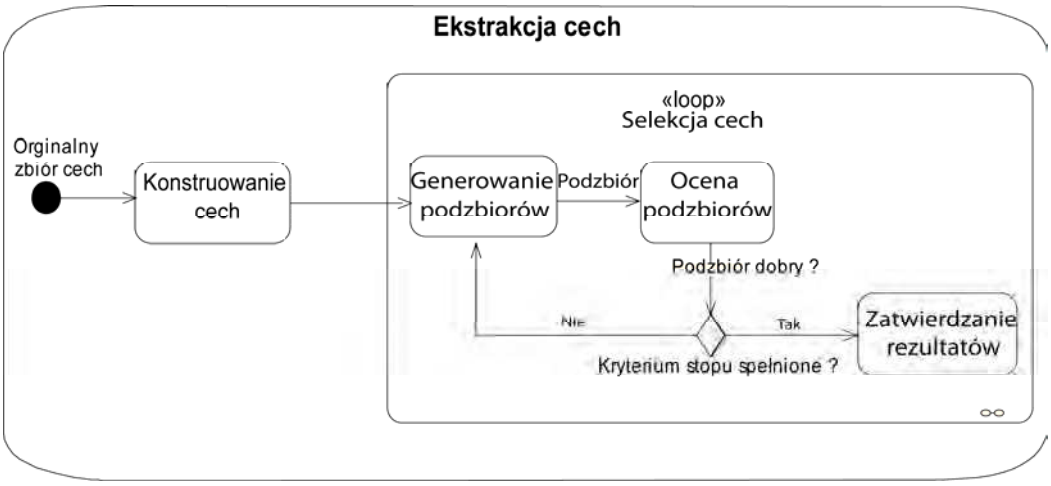
W przypadku sformalizowanych reprezentacji tekstu do metod jego analizy zaliczane są: wydobywanie wyrażeń z tekstów, wyszukiwanie informacji w szczególności wyszukiwanie informacji w reprezentacji przestrzenno wektorowej oraz grafowej, klasyfikacja oraz grupowanie. Metody te zostały omówione w kolejnych sekcjach niniejszego podpunktu.

3.2.2.1. Wydobywanie wyrażeń

Wydobywanie wyrażeń może następować poprzez ich ekstrakcje (*ang. feature extraction*). Ekstrakcja cech w literaturze określana jest także jako transformacja cech (*ang. feature transform*) czy też generowanie, uogólnianie cech (*ang. feature generation*). Proces ekstrakcji cech podzielony jest na dwa etapy: konstruowania cech (*ang. feature construction*) a następnie ich selekcji (*ang. feature selection*) [49, 50]. Selekcja cech w literaturze określana jest także jako: selekcja zmiennych (*ang. variable selection*), redukcja cech (*ang. feature reduction*), selekcja atrybutów (*ang. attribute selection*), lub selekcja podzbioru zmiennych (*ang. variable subset selection*). Metody selekcji cech można rozpatrywać w kontekście dziedziny nauki związanej z uczeniem maszynowym (*ang. machine learning*), wówczas otrzymany zostanie dodatkowy podział (ze względu na zastosowane kryterium oceny podzbioru cech), na który składają się podkategorie: filtry (*ang. filters*), opakowywacze (*ang. wrapper*) i metody wbudowane (*ang. embedded methods*).

Metody ekstrakcji, nie tylko samych wyrażeń lecz i ich semantyki, są oparte na *hipotezie dystrybucyjnej* [51] i stanowią specyficzną odmianę metod ekstrakcji specjalnie stworzonych na potrzeby analizy tekstów. Metody wydobywania podobieństwa semantycznego wyrażeń z tekstów opierają się na uzyskaniu funkcji podobieństwa semantycznego. Przegląd takich metod, odwołania do nich i opisy można znaleźć w pracy [32].

Powyżej zostały opisane klasyfikacyjne „statyczne" aspekty ekstrakcji cech. Na proces ekstrakcji cech można spojrzeć w sposób dynamiczny, wyrażony w postaci algorytmu i automatu z określoną ilością sekwencji (stanów, etapów), którego działanie ma przynieść wydobycie interesujących składowych. Kluczowe etapy tego procesu przedstawia rysunek 8.



Rysunek 8. Kluczowe etapy ekstrakcji cech. Źródło: [rozszerzone opracowanie na podstawie [52]]

Rysunek 8 prezentuje kompleksowy proces ekstrakcji cech obiektów, który w przypadku analizy tekstu obejmuje: konstruowanie cech, generowanie ich podzbioru, ocenianie otrzymanych podzbiorów oraz zatwierdzanie rezultatów jeśli uprzednio zostało spełnione kryterium stopu. W niektórych zastosowaniach pierwszy etap – konstruowania cech – nazywany jest etapem wstępnego przetwarzania (*ang. preprocessing*). Konstruowanie cech w analizie tekstu zawiera takie działania, jak standaryzacja (*ang. standadization*), normalizacja (*ang. normalization*), wydobycie lokalnych cech (*ang. extraction of local features*) [53]. Dodatkowo, do działań tych można zaliczyć techniki wstępnego przetwarzania dokumentów tekstowych wymienione w podpunkcie 3.2.1.1. Konstruowanie cech polega więc na wykorzystaniu całej dostępnej informacji w celu przejścia do nowej przestrzeni. Nowo uzyskana przestrzeń może być, w zależności od wykorzystanych metod, zredukowana, rozszerzona, pozostawiona bez zmian lub wewnętrznie zmieniana w różnych kierunkach. Redukcja wymiaru dotyczy zastosowania metod wbudowanych, które również powodują skonstruowanie nowych cech (pseudo wyrażeń) z cech wyjściowych (podstawowych, bazowych) [54-56]. Transformacja redukująca odbywa się na drodze przekształcenia liniowego bądź nieliniowego. Do liniowych przekształceń należą: analiza składowych głównych (*ang. principal components analysis – PCA)* lub rozkład na wartości osobliwe (*ang. singular value decomposition – SVD*) wykorzystywane w ukrytym indeksowaniu semantycznym (*ang. latent semantic indexing – LSI*) [33, 57]. Natomiast do nieliniowych przekształceń można zaliczyć odwzorowanie Sammona oraz skalowanie wielowymiarowe (*ang. multi dimensional scaling – MDS*) [55]. Rozszerzanie przestrzeni w przypadku analizy tekstu (ekstrakcji wyrażeń) nie znajduje zastosowania. Metodą, która działa i modyfikuje w różnych kierunkach zbiór cech, jest metoda wydobywania cech lokalnych. Przypadek, gdy przestrzeń cech (jej wymiaro-

wość) pozostaje bez zmian, świadczy o zastosowaniu metod z zakresu standaryzacji, normalizacji i zabiegów semantycznych omówionych w podpunkcie 3.2.1.

Etapem, który następuje po konstruowaniu cech obiektów, jest etap ich selekcji. Selekcja polega na wyborze możliwie małego podzbioru cech, który da jak największą możliwość rozróżnienia obiektów (dokumentów lub wyrażeń w korpusie lingwistycznym). Należy przy tym zaznaczyć, że może być wiele różnych kryteriów oceny, zależnych od specyficznego zastosowania (zwłaszcza w przypadku podejścia typu wrapper). Wybór cech polega więc na zachowaniu jedynie tych użytecznych, które niosą największą ilość informacji i wyeliminowaniu pozostałych [54]. Proces selekcji z oryginalnego zbioru cech dąży do otrzymania optymalnego ich podzbioru, który zazwyczaj jest niemożliwy do osiągnięcia. Podzbiór ten otrzymywany jest w wyniku procesu (rysunek 8) składającego się z kliku podetapów generowania podzbioru cech na drodze pomiaru i związanego z nim przyjętego kryterium oceny, oraz decyzji czy wygenerowany podzbiór cech jest odpowiedni po spełnieniu zadanego kryterium stopu [52, 53].

Po sparametryzowaniu i wykonaniu etapu generującego podzbiory dochodzi się do ich oceny. Posługując się kryterium oceny podzbiorów, można podzielić algorytmy selekcji cech na cztery kategorie: *filtry*, *wrapper*, *metody wbudowane* (*ang. embedded methods*) oraz *hybrydy* [53, 56, 58].

Przy użyciu wrappera oraz metod wbudowanych można otrzymać różne podzbiory cech z małymi peturbacjami w zbiorze danych. W celu zminimalizowania tego efektu wykorzystuje się zbiór różnych metod (*ang. ensemble learning*) [59]. Dodatkowo, oprócz ww. podziału na filtry, wrappery, metody wbudowane i hybrydy, wprowadzane są kryteria niezależne (*ang. independent criteria*) oraz zależne (*ang. dependent criteria*) [52]. Kryteria niezależne zazwyczaj związane są z modelem filtrów i do oceny podzbioru cech nie wykorzystują żadnego algorytmu eksploracji danych. Kryteria te posługują się pomiarem odległości (*ang. distance measures*), zawartości informacji (*ang. information measures*), zależności (*ang. dependency measures*) i spójności zmiennych (*ang. consistency measures*). Drugie kryterium – zależne, odnosi się do modelu wrappera i wykorzystuje predefiniowane i wydajne algorytmy eksploracji danych w selekcji cech. Niezależnie od podziału, w przypadku wrapperów oraz rozwiązań hybrydowych, wybrane cechy są dobierane w taki sposób, aby zapewnić możliwe najlepsze wyniki działania docelowej metody (np. grupowanie, klasyfikacja), podczas gdy filtr jest niezależny od stosowanej później metody przetwarzania dokumentów.

Wybór podzbiorów odpowiednich cech i ich ocenianie trwa dopóki nie zostanie spełniony warunek stopu. Rysunek 8 prezentuje ten warunek jako akcję decyzyjną pt. „Kryterium stopu spełnione ?”. Warunek stopu jest spełniony gdy spełnione są następujące warunki:

a) przeszukiwanie jest kompletne tj. zbadano całą przestrzeń za pomocą algorytmu przeszukiwania,
b) osiągnięta została specyficzna granica np. ilości iteracji czy też ilości cech,
c) dodawanie lub usuwanie cech nie polepsza i nie generuje ich podzbiorów o lepszych parametrach,
d) określony błąd pomiaru spadł poniżej wyznaczonej granicy.

Ostatnim etapem selekcji cech, choć nie koniecznie kończącym ten proces, jest faza zatwierdzania rezultatów. Bezpośrednio jakość wybranego podzbioru cech można ocenić *a priori* na podstawie jego porównania z cechami jakie się oczekuje. Zazwyczaj taka wiedza *a priori* nie jest dana, wówczas wykorzystywane są metody pośrednie polegające na badaniu jakości osiągów (zwiększanie, bądź zmniejszanie np. celności klasyfikacji) algorytmów eksploracyjnych do wyznaczonego zadania np. klasyfikacji.

W ogólnym przypadku zastosowanie selekcji cech, czy też ogólniej – ekstrakcji cech, ma dodatkowo za zadanie [53]: zredukować dane, zmniejszyć ilość potrzebnej pamięci i tym samym przyczynić się do przyśpieszenia algorytmów operujących na tych danych, zredukować zbiór cech, ulepszyć przetwarzanie (osiągi) związane z dokładnością przewidywania oraz doprowadzić do zrozumienia danych poprzez pozyskanie wiedzy o procesie, który generuje dane i dostarczyć możliwość ich wizualizowania.

Koncepcję podziału selekcji cech wyrażoną w postaci trójwymiarowego szkieletu (*ang. three-dimensional framework*), oraz uogólnione, algorytmiczne modele filtrów zostały przedstawione w pracy [52].

3.2.2.2. Wyszukiwanie informacji

Termin wyszukiwanie informacji określa i odnosi się do procesów oraz metod i technik wykorzystywanych do wyszukiwania żądanej informacji w zbiorze dokumentów tekstowych) [10, 33, 60, 61]. Wyszukiwanie to odbywa się na podstawie zadanych zapytań składających się z wyrażeń *t* (*ang. terms*). Z dziedziny wyszukiwania informacji wywodzą się też koncepcje dotyczące m.in. budowy i reprezentacji dokumentów tekstowych, ich indeksowania oraz oceny zastosowanego rozwiązania. Koncepcje te stosowane są przy analizach dokumentów tekstowych opisanych w niniejszym opracowaniu.

3.2.2.3. Wyszukiwanie informacji – reprezentacja przestrzenno-wektorowa

Na podstawie macierzy A z odpowiednio skonstruowanymi wagami w_{ij} możliwe jest wyznaczenie podobieństwa słów oraz dokumentów. Podobieństwo słów wyrażane jest poprzez określenie podobieństwa odpowiadających im kolumn tej macierzy, natomiast o podobieństwie dokumentów wnioskuje się na podstawie analizy podobieństwa wierszy tej macierzy. Najczęściej wszystkie wagi w_{ij} wektorów macierzy A w zastosowaniach praktycznych są normalizowane do 1.

W celu określenia miary podobieństwa (dokumentów jak i wyrażeń) stosuje się metryki jak np.: euklidesową, blokową (Manhattanowi), L∞, uogólnioną Minkowskiego L_λ, cosinusową, Jaccarda czy też Dicea [10, 33, 36]. Podobieństwo dokumentów ustala się na podstawie pomiaru odległości. W wyszukiwaniu należy minimalizować odległość maksymalizując w ten sposób podobieństwo. Najpopularniejsze w zastosowaniach metryki, określające podobieństwo dokumentów wrażane są w następujący sposób:

a) miara Euklidesowa, wyrażana jest w postaci wzoru:

$$d_E(i,j) = \left[\sum_{k=1}^{n} (w_k(i) - w_k(j))^2 \right]^{\frac{1}{2}} \tag{4}$$

Gdzie:
- i oraz j – oznaczają *i*-ty i *j*-ty dokument między którymi wyznaczana jest odległość (odpowiednie wiersze macierzy A z reprezentacji, którą przedstawia rysunek 7)
- n – ilość składowych (wyrażeń) występujących w macierzy A
- $w_k(i)$ i $w_k(j)$ – kolejne *k*-te wagi (wartości obserwacji) dla *i*-tego oraz *j*-tego dokumentu

b) miara Manhattanu (L_1) nazywana także *miarą miejską*, wyrażana jest w postaci wzoru:

$$d_M(i,j) = \sum_{k=1}^{n} |(w_k(i) - w_k(j)) \tag{5}$$

c) miara L∞, wyrażana w jest postaci wzoru:

$$d_\infty(i,j) = \max_k |(w_k(i) - w_k(j)| \tag{6}$$

d) miara uogólniona Minkowskiego L_λ, wyrażana jest w postaci wzoru:

$$d_\lambda(i,j) = (\sum_{k=1}^{n}(w_k(i) - w_k(j))^\lambda)^{\frac{1}{\lambda}} \tag{7}$$

Gdzie:
- $\lambda \geq 1$ – jeśli za λ przyjęte zostanie: $\lambda = 2$ uzyskana zostanie metryka Euklidesowa, $\lambda = 1$ to uzyskana zostanie metryka Manhattanu i $\lambda \rightarrow \infty$ to uzyskana zostanie metryka $L\infty$

e) miara odległości kosinusowa, wyrażana jest w postaci wzoru:

$$d_C(i,j) = \frac{\sum_{k=1}^{n} w_k(i) \cdot w_k(j)}{\sqrt{\sum_{k=1}^{n} w_k(i)^2 \sum_{k=1}^{n} w_k(j)^2}} \tag{8}$$

f) miara Jaccarda, wyrażana jest w postaci wzoru:

$$d_J(i,j) = \frac{2\sum_{k=1}^{n} w_k(i) \cdot w_k(j)}{\sum_{k=1}^{n} w_k(i)^2 + \sum_{k=1}^{n} w_k(j)^2} \tag{9}$$

g) miara współczynnika Dicea, wyrażany jest w postaci wzoru:

$$d_D(i,j) = \frac{2 \cdot |d_i \cap d_j|}{(|d_i| + |d_j|)} \tag{10}$$

Wzór (współczynnik) Dicea można interpretować następująco:

$$d_D(i,j) = \frac{2 \cdot \text{liczba wspolnych wyrazen w dokumencie } d_i \text{ i } d_j}{\text{liczba wyrazen w dokumencie } d_i + \text{liczba wyrazen w dokumencie } d_j} \tag{11}$$

h) miara oszacowania pokrycia (*ang. expected overlap measure*) [36], wykorzystywana gdy wagi wyrażeń w_{ij} zostały wyrażone probabilistycznie (równanie 2). Miara ta wyrażana jest w postaci wzoru:

$$d_{EO}(d_i, d_j, A) = \sum_{t \in D_i \cap D_j} \left[P(Y_i = t \mid d_i, M) \cdot P(Y_j = t \mid d_j, M) \right] \tag{12}$$

W przestrzeni wektorowej wykorzystując ww. miary podobieństwa istnieje możliwość wyszukiwania dokumentów na podstawie zapytania *Q*. Wyszukiwanie to polega na wnioskowaniu opierającym się na zapytaniu *Q*, prowadzącym do odnalezienia najbardziej podobnych do niego obiektów. Obiekty te w opisywanym przypadku stanowią zbiór dokumentów tekstowych. Zapytanie *Q* może zostać wyrażone w postaci:

a) Boolowskiej funkcji logicznej na zbiorze dostępnych wyrażeń np. *pożar AND mocne zadymienie AND prąd gaśniczy AND NOT (prąd elektryczny),*

b) wektora wag $Q = (q_1,\ldots,q_j)$, gdzie q_j stanowi wagę wyrażenia w zapytaniu i $q_j \epsilon \langle 0,1 \rangle$.

Jeżeli zapytanie *Q* będzie składało się z poszukiwanych wyrażeń i w przypadku zastosowania innej reprezentacji ich wag niż Boolowska, to otrzymany zostanie ranking poszukiwanych dokumentów. Zastosowanie zapytania *Q* w wektorowej wagowej postaci wyrażeń i zastosowanie jednolitego zapisu, tj. takiej samej wektorowej reprezentacji dla zbioru dokumentów i wyrażeń w postaci macierzy *A* oraz wektora *Q,* umożliwia stworzenie rankingu poszukiwanych dokumentów. Wyszukiwanie w tym przypadku opiera się na badaniu odległości, która jest określona za pomocą opisanych powyżej miar między wektorem zapytań *Q* składającym się z wybranych wyrażeń i ich wag a macierzą *A* (wierszami w przyjętej w opracowaniu reprezentacji).

W przypadku zastosowania reprezentacji Boolowskiej zarówno dla *A* jak i *Q* przy wyszukiwaniu nie opartym na mierze lecz na dopasowaniu, istnieje szereg problemów m.in.:

a) brak jest naturalnego znaczenia pojęcia odległości między zapytaniem a dokumentem. W wyniku wyszukiwania uzyskiwany jest nieuporządkowany zbiór (względem miary) dokumentów, pasujących dokładnie do zapytania *Q*,
b) brak jest możliwości wprowadzenia rankingu dokumentów,
c) powstaje problem z konstruowaniem wyrażeń boolowskich, stąd pojawia się problem użyteczności (*ang. usability*) polegający na zrozumieniu przez użytkownika sposobu formułowania tych wyrażeń i ich stosowaniu.

Mimo tych wad rozwiązanie oparte o reprezentacje Boolowską jest dalej popularne i szeroko stosowane ze względu na implementacyjną prostotę i efektywność. W celu przezwyciężenia ww. problemów stosuje się rozszerzone podejścia boolowskie do reprezentacji i wyszukiwania dokumentów, które pozwalają na uzyskanie rankingu (zakładają one częściowe dopasowanie dokumentów do zapytania). Wykorzystuje się również pozostałe wyżej wymienione odmiany reprezentacji dokumentów tekstowych tj.: częstotliwościową występowania wyrażeń, odwrotną częstość etc. Ich głównym atutem jest to iż umożliwiają tworzenie rankingu istotności zwracanych dokumentów na podstawie zadanego wzorca *Q*.

3.2.2.4. Klasyfikacja dokumentów tekstowych

Klasyfikacja, nazywana także kategoryzacją, dokumentów tekstowych polega na określeniu do jakiej klasy dokumentów można zaliczyć wybrany tekst [41, 62-65] lub jego fragment [66, 67]. Klasyfikacja odbywa się za pomocą wyznaczonego w procesie uczenia klasyfikatora, który będzie dokonywał przyporządkowania dokumentów do jednej lub kilku uprzednio zdefiniowanych klas. Klasy te nie są definiowane wprost, lecz poprzez zbiór trenujący, który stanowi grupa dokumentów już odpowiednio zaklasyfikowana ręcznie np. przez ekspertów. W większości przypadków klasy nie są zagnieżdżane, natomiast przyjmuje się, iż jeden dokument może należeć do więcej niż jednej klasy. Do kategoryzacji dokumentów tekstowych używane są takie techniki, jak [68]: drzewa decyzyjne (*ang. decission tree*), reguły decyzyjne, algorytmy najbliższych sąsiadów i związane z nimi różnego rodzaju metryki (m.in. przedstawione w podpunkcie 3.2.2.3), klasyfikator bayesowski, sieci neuronowe, metody regresyjne czy też techniki z zakresu maszyn wektorów wspierających (*ang. suport vector machines – SVM*), oraz metody odnajdywania wspólnych podgrafów opartej na metodzie najbliższych sąsiadów ze specjalizowaną miarą odległości, w przypadku zastosowania modelu grafowego dokumentów [31].

3.2.2.5. Grupowanie dokumentów tekstowych

Grupowanie dokumentów tekstowych polega na wyznaczeniu grup podobnych dokumentów np. ze względu na ich tematykę, m.in. za pomocą analizy statystycznej słów występujących w tekście [5, 32, 41, 69-71]. Grupowanie dokumentów tekstowych jest zadaniem pokrewnym do klasyfikacji. W tym przypadku jednak system nie posiada wejściowej wiedzy w postaci już zakwalifikowanych dokumentów, czy też klas wyznaczonych przez ekspertów. Zadaniem tej metody jest takie pogrupowanie dokumentów, by dokumenty należące do jednej klasy były do siebie jak najbardziej podobne i jednocześnie różniły się znacząco od tych należących do innych klas. Do grupowania dokumentów tekstowych używane są takie techniki, jak: analiza skupień, klastrowanie (*ang. clustering*) [72], samoorganizujące się mapy (*ang. self-organization map*) [73], algorytmy aproksymacji wartości oczekiwanej (*ang. expectation-maximization*) [74] czy też zbiory przybliżone [75].

3.3. Wizualizacja

Wizualizacja to metoda związana z końcową realizacją analizy tekstu i wykonywana jest w celu wizualizowania i lepszego zrozumienia otrzymanych wyników. Głównym celem wizualizacji jest zapewnienie inżynierowi wiedzy lub oprogramowania prostej metody interpretacji uzyskanych wyników. Najczęściej wizualizacji poddawane są związki zachodzące pomiędzy wyodrębnionymi faktami lub zależności zachodzące w strukturze rozpatrywanego zbioru dokumentów tekstowych [76]. Oprócz zaproponowanych metod wizualizacji danych tekstowych stosowanych podczas omawianego procesu przetwarzania raportów, istnieje szereg innych metod wizualizacyjnych. Związane są one zarówno z inżynierią wiedzy jak i eksploracyjną analizą tekstu. Do najbardziej znanych metod reprezentacji (wizualizacji) wyników (danych), należą: sieci semantyczne związane z ontologiami, kraty pojęć wykorzystywane w formalnej analizie pojęć, histogramy, wykresy słupkowe, kolumnowe, mapy znaczeń, wykresy gwiazdowe oraz macierze korelacji narysowane jako obrazy pikselowe wykorzystywane w wyszukiwaniu, klasyfikowaniu oraz grupowaniu dokumentów tekstowych [17, 33, 71, 77, 78].

4. Podsumowanie

Przedstawione metody analizy dokumentów tekstowych mogą służyć do przetwarzania raportów składowanych w systemie EWID PSP, sporządzanych z akacji ratowniczo-gaśniczych. Dalsze badania autorów będą dotyczyć procesu ekstrakcji wyrażeń, klasyfikacji oraz ekstrakcji informacji z dostępnych raportów. Aktualnie autorzy utworzyli korpus raportów składający się z oznakowanych zdań przydzielonych manualnie do wybranych klas, który jest gotowy do dalszego przetwarzania. Przewidywana jest ich reprezentacja w przestrzeni wektorowej, której wagi kolejnych wyrażeń będą zakodowane za pomocą jednego z opisywanych w artykule schematów wag wyrażeń. Aktualnie przewiduje się wykorzystanie wszystkich wymienionych w artykule reprezentacji, od boolowskiej do probabilistycznej, w celu znalezienia odpowiedniego kodowania do wyszukiwania rozwiązań w bazie *półustrukturyzowanych użytecznych przypadków zdarzeń* i klasyfikowania części raportów. Ekstrakcja wyrażeń będzie miała na celu wydobycie ze zdań raportów kluczowych wyrażeń charakterystycznych dla wyznaczonych klas. W następnym kroku zbudowany zostanie klasyfikator lub ich grupa w celu wyboru najlepszego rozwiązania oraz przeprowadzony zostanie proces klasyfikacji pozostałej, nieoznakowanej części raportów do utworzonych podpunktów (klas semantycznych). Za pomocą wybranych metod wizualizacyjnych oraz grupujących zostaną przeprowadzone numeryczne analizy zdań z wybranych klas. Działania te mają na celu utworzenie formalnego słownika pojęć (ontologii), w oparciu o który inżynierowie oprogramowania będą mogli przygotowywać częściową reprezentację danych wyrażoną np. w notacji obiektowej. Ontologie budowane na podstawie tych analiz mają stanowić „język", który w przyszłości mógłby być zastosowany do przygotowywania meldunków. Na końcu przeprowadzony zostanie proces ekstrakcji informacji do utworzonych modeli.

Autorzy dopuszczają możliwość uczenia przyrostowego i modyfikację utworzonych ontologii i modeli po zebraniu nowych danych w postaci nowo tworzonych raportów po interwencji sił i środków PSP. Ograniczenia i problemy w proponowanym procesie mogą wiązać się z elementem wykorzystywanym do translacji, ze względu na dwukierunkowe tłumaczenie. Niemniej rozwiązanie takie można zaproponować gdyż techniki te są cały czas rozwijane i doskonalone. Natomiast charakter zagrożeń jest międzynarodowy a ewentualne działania i kooperacje transgraniczne jednostek ratowniczych powoduje, że takie rozwiąza-

nia powinny być brane pod uwagę. Problem ustandaryzowanego słownictwa opisującego zdarzenia w tworzonych meldunkach został rozwiązany za pomocą manualnie lub półautomatycznie tworzonych ontologii wraz z ich obiektową realizacją. Zastosowanie praktyczne w dziedzinie ratownictwa rezultatów badań, proponowanych rozwiązań, jest aktualnie niezależne od badaczy i w większym stopniu zależy od szczebli decyzyjnych w komendzie głównej PSP.

Bibliografia

[1] 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.

[2] *Abakus: System EWID99*. [on-line] [dostęp: 1 maja 2009] Dostępny w Internecie: *http://www.ewid.pl/?set=rozw_ewid&gr=roz*.

[3] *Abakus: System EWIDSTAT*. [on-line] [dostęp: 1 maja 2009] Dostępny w Internecie: *http://www.ewid.pl/?set=ewidstat&gr=prod*.

[4] *Strona firmy abakus*. [on-line] [dostęp: 1 marca 2009] Dostępny w Internecie: *http://www.ewid.pl/?set=main&gr=aba*.

[5] Kozłowski J., 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*.

[6] Krasuski A., Maciak T. *Wykorzystanie rozproszonej bazy danych oraz wnioskowania na podstawie przypadków w procesach decyzyjnych Państwowej Straży Pożarnej*. Zeszyty Naukowe SGSP, No 36, 2008, s. 17-35.

[7] Mirończuk M., Maciak T. *Problematyka projektowania modelu hybrydowego systemu wspomagania decyzji dla Państwowej Straży Pożarnej*. Zeszyty Naukowe SGSP, No 39, 2009.

[8] Kempa A. *Zastosowanie rozszerzonej metodologii wnioskowania na podstawie przypadków - textual cbr w pracy z dokumentami tekstowymi*. Katowice: Systemy Wspomagania Organizacji, 2005. [dostęp: 1 stycznia 2008] Dostępny w Internecie: *http://www.swo.ae.katowice.pl/content/view/221/32/*.

[9] Zhou N., Cheng H., Chen H., Xiao S. *The Framework of Text-Driven Business Intelligence*. Wireless Communications, Networking and Mobile Computing, WiCom 2007 International Conference, 2007.

[10] Mykowiecka A. *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*. Warszawa: PJWSTK, 2007.

[11] Mirończuk M. *Zmodyfikowana analiza FMEA z elementami SFTA w projektowaniu systemu wyszukiwania informacji na temat obiektów hydrotechnicznych w nierelacyjnym katalogowym rejestrze*. Studia Informatica, No 2, 2011.

[12] Paul F., Fearzana H., Enid H. *The efficacy of the 'mind map' study technique*. Medical Education, 2002. s. 426-431.

[13] *FreeMind - free mind mapping software*. [dostęp: 12 lipca 2011] Dostępny w Internecie: *http://freemind.sourceforge.net/wiki/index.php/Main_Page*.

[14] Mirończuk M. *Systemy zarządzania bazą danych i architektura agentowa w służbach ratowniczych Państwowej Straży Pożarnej*. CNBOP „Bezpieczeństwo i Technika Pożarnicza", No 1, 2011.

[15] Doctrine MongoDB Object Document Mapper. [dostęp: 1 stycznia 2010] Dostępny w Internecie: *http://www.doctrine-project.org/blog/doctrine-mongodb-object-document-mapper*.
[16] Hibernate. [dostęp: 10 marca 2011] Dostępny w Internecie: *http://www.hibernate.org*.
[17] Wolff K. E. *A first course in formal concept analysis*. [dostęp: 22 grudnia 2009] Dostępny w Internecie: *http://www.fbmn.fh-darmstadt.de/ home/wolff/Publikationen /A_First_Course_in_Formal_Concept_Analysis.pdf*.
[18] Patil P. *Applying Formal Concept Analysis to Object Oriented Design and Refactoring.* Bombay: Department Of Computer Science and Engineering Indian Institute Of Technology, 2009.
[19] Priss U. *Formal concept analysis in information science*. Annual review of information science and technology, No 40, 2006, s. 521-543.
[20] Hwang S. H., Kim H. G., Yang H. S. *A FCA-Based Ontology Construction for the Design of Class Hierarchy*. In: Gervasi O., Gavrilova M., Kumar V., Laganà A., Lee H., Mun Y., et al., editors. *Computational Science and Its Applications – ICCSA 2005*: Springer Berlin/Heidelberg, 2005. s. 307-320.
[21] Carpineto C., Romano G. *Using Concept Lattices for Text Retrieval and Mining*. In: Ganter B., Stumme G., Wille R., editors. *Formal Concept Analysis*: Springer Berlin / Heidelberg, 2005. s. 3-45.
[22] Maedche A., Staab S. *Mining Ontologies from Text*. Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management, 2000.
[23] Brewster C., Jupp S., Luciano J., Shotton D., Stevens R., Zhang Z. *Issues in learning an ontology from text*. BMC Bioinformatics, No 10, 2009, s. 25-29.
[24] Jiang G., Ogasawara K., Endoh A., Sakurai T. *Context-based ontology building support in clinical domains using formal concept analysis*. International Journal of Medical Informatics, No 71, 2003, s. 71-81.
[25] Radvansky M. *Formal concept analyse*. [dostęp: 1 maja 2011] Dostępny w Internecie: *http://www.fca.radvansky.net/news.php*.
[26] Moens M. F. *Information Extraction: Algorithms and Prospects in a Retrieval Context* (The Information Retrieval Series). Springer, 2006.
[27] Feldman R., Dagan I., Hirsh H. *Mining Text Using Keyword Distributions*. Journal of Intelligent Information Systems, No 10, 1998.
[28] Witten I. H., Don K. J., Dewsnip M., Tablan V. *Text mining in a digital library*. International Journal on Digital Libraries, No 4, 2004, s. 56-59.
[29] Maciołek P., Dobrowolski G. *Propozycja metody klasyfikacji dokumentów w języku polskim*. In: Grzech A., Juszczyszyn K., Kwaśnicka H., Nguyes N. T., editors. *Inżynieria wiedzy i systemy ekspertowe*. Warszawa: Akademicka oficyna wydawnicza EXIT, 2009.
[30] Chow T. W. S., Haijun Z., Rahman M. K. M. *A new document representation using term frequency and vectorized graph connectionists with application to document retrieval*. Expert Systems with Applications, No 36, 2009, s. 12023-12035.
[31] Schenker A., Kandel A., Bunke H., Last M. *Graph-Theoretic Techniques for Web Content Mining*. World Scientific Publishing Co, 2005.
[32] Broda B. *Mechanizmy grupowania dokumentów w automatycznej ekstrakcji sieci semantycznych dla języka polskiego*. Wydział Informatyki i Zarządzania. Wrocław: Politechnika Wrocławska, 2007.
[33] Hand D., Mannila H., Smith P. *Eksploracja danych. Wydanie 1*. Warszawa: Wydawnictwo Naukowo-Techniczne, 2005.
[34] Morzy M., Królikowski Z. *Metody indeksowania atrybutów zawierających zbiory*. Pro Dialog, No 15, 2003, s. 87-106.

[35] Dudczak A. *Zastosowanie wybranych metod eksploracji danych do tworzenia streszczeń tekstów prasowych dla języka polskiego*. Wydział Informatyki i Zarządzania Instytut Informatyki. Poznań: Politechnika Poznańska 2007.
[36] Goldszmidt M., Sahami M. *A Probabilistic Approach to Full-Text Document Clustering*. 1998.
[37] Singhal A., Buckley C., Mitra M., Mitra A. *Pivoted Document Length Normalization*. ACM Press, 1996, s. 21-29.
[38] Robertson S. E., Walker S., Jones S., Hancock-Beaulieu M. M., Gatford M. Okapi at TREC-3. 1996, s. 109-126.
[39] Lin D. *Using syntactic dependency as local context to resolve word sense ambiguity*. [Madrid, Spain]: Annual Meeting of the ACL Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 1997.
[40] Matsuo Y., Ishizuka M. *Keyword Extraction From A Single Document Using Word Co-Occurrence Statistical Information*. International Journal on Artificial Intelligence Tools, No 13, 2004, s. 157-169.
[41] Borycki Ł., 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*.
[42] Neumann G., Piskorski J. *A Shallow Text Processing Core Engine. Computational Intelligence*, No 18, 2002, s. 451-476.
[43] Praca zbiorowa Wikipedia. *Full text search*. [dostęp: 10 czerwca 2009] Dostępny w Internecie: *http://en.wikipedia.org/wiki/Full_text_search*.
[44] Bikel D. M., Schwartz R., Weischedel R. M. *An Algorithm that Learns What's in a Name*. Machne Learning, 1999, s. 211-231.
[45] McNamee P. *Language identification: a solved problem suitable for undergraduate instruction*. Journal of Computing Sciences in Colleges, No 20, 2005, s. 94-101
[46] He X., Yang M., Gao J., Nguyen P., Moore R. *Improved Monolingual Hypothesis Alignment for Machine Translation System Combination*. No 8, 2009, s. 1-19.
[47] Feng Y., Liu Y., Mi H., Liu Q., Lü Y. *Lattice-based system combination for statistical machine translation*. [Singapore]: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Volume 3, 2009.
[48] He X., Toutanova K. *Joint optimization for machine translation system combination*. [Singapore]: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Volume 3, 2009.
[49] Dasgupta A., Drineas P., Harb B., Josifovski V., Mahoney M. W. *Feature selection methods for text classification*. [San Jose, California, USA]: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007.
[50] Li S., Xia R., Zong C., Huang C. R. *A framework of feature selection methods for text categorization*. [Suntec, Singapore]: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Volume 2, 2009.
[51] Karlgren J., Sahlgren M. *From Words to Understanding. 2001*. [dostęp: 10 stycznia 2010] Dostępny w Internecie: *http://www.sics.se/~mange/papers/KarlgrenSahlgren2001.pdf*.
[52] Liu H., Yu L. *Toward integrating feature selection algorithms for classification and clustering*. Knowledge and Data Engineering, IEEE Transactions, No 17, 2005, s. 491-502.

[53] Guyon I., Elisseeff A. *Introduction to Feature Extraction. Studies in Fuzziness and Soft Computing*. Berlin/Heidelberg: Springer 2006.
[54] Torkkola K. *Feature extraction by non parametric mutual information maximization*. The Journal of Machine Learning Research, No 3, 2003, s. 1415-1438
[55] Pal S. K., Mitra P. *Pattern Recognition Algorithms for Data Mining Scalability, Knowledge Discovery and Soft Granular Computing*. London New York Washington, D.C.: Chapman & Hall , 2004.
[56] Praca zbiorowa *JMLR Special Issue on Variable and Feature Selection*. [dostęp: 5 stycznia 2010] Dostępny w Internecie: *http://jmlr.csail.mit.edu/papers/special/feature03.html*.
[57] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. *Indexing by latent semantic analysis*. Journal of the American Society for Information Science, No 41, 1990, s. 391-407.
[58] 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*.
[59] Tuv E. *Ensemble Learning*. In: Guyon I., Gunn S., Nikravesh M., Zadeh L. A., editors. *Feature Extraction: Foundations and Applications* (Studies in Fuzziness and Soft Computing) (Hardcover): Springer, 2006.
[60] Baeza-Yates R., Ribeiro-Neto B. *Modern Information Retrieval*. Boston: Addison-Wesley Longman Publishing, 1999.
[61] Manning C. D., Raghavan P., Schtze H. *Introduction to Information Retrieval*. Cambridge University Press India, 2008.
[62] Song F., Liu S., Yang J. *A comparative study on text representation schemes in text categorization*. Pattern Analysis & Applications, No 8, 2005, s. 199-209
[63] Weigend A. S., Wiener E. D., Pedersen J. O. *Exploiting Hierarchy in Text Categorization*. Information Retrieval, No 1, 1999.
[64] Yang Y., Liu X. *A re-examination of text categorization methods*. [New York]: ACM SIGIR Conference of Research and Development in Information Retrieval, 1998.
[65] Łażewski Ł., Pikuła M., Siemion A., Szklarzewski M. *Klasyfikacja dokumentów tekstowych*. Warszawa: PJWSTK 2005. Dostępny w Internecie: *http://www.scribd.com/doc/2242106/Klasyfikacja-dokumentow-tekstowych*.
[66] Agarwal S., Yu H. *Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. Bioinformatics*, No 25, 2009, s. 3174-3180.
[67] Sebastiani F. *Machine learning in automated text categorization*. ACM Comput Surv, No 34, 2002, s. 1-47.
[68] Aas K., Eikvil L. *Text Categorisation: A Survey*. Technical Report, Norwegian Computing Center, 1999.
[69] Weiss S., White B., Apte C., Weiss S. M., White B. F., Apte V. *Lightweight Document Clustering*. 2000.
[70] Domeniconi C., Gunopulos D., Ma S., Papadopoulos D., Yan B. *Locally adaptive metrics for clustering high dimensional data*. Data Mining and Knowledge Discovery, No 1, 2006, s. 63-97.
[71] Solka J. L. *Text Data Mining: Theory and Methods. Statistic Survey*.
[72] Everitt B. S., Landau S., Leese M. *Cluster Analysis*. 2001.
[73] Kohonen T. *Self-Organizing Maps*. In: Sciences S.S.i.I., editor. Wydanie 3. Berlin: Springer, 2001.

[74] Dempster A. P., Laird N. M., Rdin D. B. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, No 39, 1977, s. 1-38.
[75] Rutkowski L. *Metody i techniki sztucznej inteligencji*. Wydawnictwo Naukowe PWN, 2005.
[76] Lula P. *Text mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*. StatSoft, 2005.
[77] Friedman V. *Data Visualization: Modern Approaches*. [dostęp: 29 grudnia 2009] Dostępny w Internecie: *http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches/*.
[78] Piwowar K. *Wizualizacja danych a ich używalność – czyli pokazać to tak, aby inni to zrozumieli*. [dostęp: 29 grudnia 2009] Dostępny w Internecie: *http://interaktywnie.com/biznes/blog-ekspercki/blogi/wizualizacja-danych-a-ich-uzywalnosc-8211-czyli-pokazac-to-tak-aby-inni-to-zrozumieli-384*.

Projekt współfinansowany ze środków Europejskiego Funduszu Społecznego w ramach Programu Operacyjnego Kapitał Ludzki Działanie 8.2 Transfer wiedzy, Poddziałanie 8.2.2 Regionalne strategie innowacji, budżetu państwa oraz środków Samorządu Województwa Podlaskiego.

Artykuł finansowany w ramach pracy statutowej S/WI/4/08