ZACHODNIOPOMORSKI UNIWERSYTET TECHNOLOGICZNY W SZCZECINIE

WYDZIAŁ INFORMATYKI

mgr inż. Marcin Gibert

Autoreferat rozprawy doktorskiej

Integracja metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji

Promotor rozprawy:
dr hab. Bożena Śmiałkowska, prof. ZUT

Promotor pomocniczy rozprawy:
dr inż. Jarosław Jankowski

Recenzenci rozprawy:

dr hab. inż. Arkadiusz Orłowski, prof. SGGW
Wydział Zastosowań Informatyki i Matematyki
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

dr hab. inż. Robert Burduk
Wydział Elektroniki
Politechnika Wrocławska

Szczecin 2016

Spis treści

1. Aktualność problemu

Proces podejmowania decyzji *PD* (inaczej proces decyzyjny) wspierany jest różnymi metodami i technikami często bazującymi na komputerowym wspomaganiu. Szerokie zastosowanie mają tu komputerowe systemy wspomagania decyzji DSS (ang. Decision Suport Systems) [10, s. 24], często uzupełniane przez systemy odkrywania wiedzy z danych KDD (ang. Knowledge Discovery in Databases) [1, s. 1], w których wykorzystywane są metody eksploracji danych (ang. Data mining) [21, ss. 153–154].

Problem decyzyjny zdefiniowany w procesie *PD* rozważany jest w oparciu o dane (cechy i charakterystyki) opisujące obiekty analizowane w tym procesie. Mogą one być wyrażone za pomocą danych numerycznych (za pomocą liczb), mogą wynikać z opisu sformułowanego za pomocą języka naturalnego, a także mogą być wyrażone innymi typami danych np. multimedialnymi, które w oryginalnej formie nie są ani danymi numerycznymi ani tekstowymi.

Metody eksploracji danych wspomagające proces decyzyjny operują na zbiorze wszystkich danych zgromadzonych w tym procesie, jednak zgodnie z literaturą [11] [18] [14] metody eksploracji danych koncentrują się przede wszystkim na danych numerycznych oraz danych tekstowych. Z kolei eksploracja innych danych może być realizowana przez transformację tych danych do reprezentacji wyrażonych przez zbiór ustrukturyzowanych danych numerycznych [31, s. 20] lub opisanych za pomocą zbioru danych tekstowych [25, s. 108]. Zatem zbiór Z_E, będący przedmiotem eksploracji jest sumą mnogościową dwóch typów danych zgodnie ze wzorem *(1)*.

$$Z_E = Z_N \cup Z_T \qquad (1)$$

gdzie:

Z_E – zbiór dostępnych danych (zbiór danych w procesie *PD* poddanych eksploracji, wyrażonych za pomocą danych tekstowych lub numerycznych),

Z_N – zbiór danych numerycznych,

Z_T – zbiór danych tekstowych.

Jeśli proces *PD* oparty jest wyłącznie na jednym z tych zbiorów (danych numerycznych lub tekstowych) to w literaturze przedmiotu [31] [2, ss. 163–213] dostępnych jest wiele metod eksploracji tych danych.

Głównym problemem jest taka sytuacja, w której eksploracja realizowana jest jednocześnie w oparciu o zbiór danych numerycznych oraz tekstowych. W literaturze [7, ss.99-100] zauważa się, że integracja metod eksploracji danych może wpłynąć na osiąganie korzystniejszego wyniku w sensie kryterium nośności informacyjnej danych, wskaźników jakości eksploracji, a co za tym idzie jakości decyzji w procesie wspomagania decyzji w stosunku do wyników osiąganych przez metody indywidualnie, dedykowane odrębnie danym numerycznym lub tekstowym [12, s. 2]. Dodatkowo w literaturze podkreślono istotne znaczenie problemu integracji metod eksploracji danych tekstowych i numerycznych w wielu dziedzinach, takich jak finanse, medycyna czy web mining [6, s. 310], [29, s. 151], [23, s. 18] oraz [16]. Ponadto w opracowaniach badawczych wskazano na możliwość pozyskania bardziej wartościowej wiedzy, gdy uwzględnia się jednocześnie w procesie *PD* eksplorację danych numerycznych i tekstowych [6, s. 314] [29, s. 368] [26, s. 4] oraz, że brakuje tu metody, która jednocześnie w sposób wieloaspektowy i systemowy umożliwiałaby eksplorację obu wyróżnionych typów danych w sposób adekwatny do tych typów.

2. Przedmiot badań

Przedmiotem badań jest opracowanie procedury integracji metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji, która zwiększy nośność informacyjną danych mierzoną w oparciu o jakość decyzji w procesie *PD*. W pracy skoncentrowano się na szczególnie przydatnej i popularnej z praktycznego punktu widzenia procesów *PD* metodzie wykorzystywanej w eksploracji danych, jaką jest klasyfikacja [9, ss. 119–121] [31, s. 159]. W celu uwzględnienia specyfiki języka polskiego (fleksyjnego), w eksploracji danych tekstowych w modelu przestrzeni wektorowej VSM [15, ss. 45–54] wykorzystano reprezentację danych uwzględniająca tę specyfikę. W przypadku eksploracji danych numerycznych użyto metody Teorii Zbiorów Przybliżonych [20].

3. Metody naukowe stosowane podczas wykonywania badań

W eksploracji danych tekstowych wykorzystano analizę fleksyjną języka polskiego po to by przy opracowaniu elementów reprezentacji tych danych zwiększyć ich możliwości interpretacyjne w kontekście postawionego problemu decyzyjnego oraz zweryfikować ich poprawność. Zaproponowano również metodę opracowywania γ-gramowej [7, s. 45] reprezentacji tekstu, której elementy tzw. rzeczowe informacje, są ekstrahowane na podstawie wzorców informacyjnych. W pracy wykorzystano również takie metody jak:

- Analizę systemową,
- Analizę SWOT (mocne strony-szanse i zagrożenia),
- Analizę danych źródłowych ukierunkowaną na badanie istotności i wiarygodności tych danych (metoda Teoria Zbiorów Przybliżonych),
- Analizę przypadków użycia,
- Metodę statystyczną (test zgodności pomiędzy wynikami pomiarów McNemara).

4. Główny cel rozprawy

W związku ze zidentyfikowanymi brakami metod eksploracji danych w procesie podejmowania decyzji *PD*, za główny cel pracy przyjęto *opracowanie procedury integracji metod analizy fleksyjnej tekstu oraz metod eksploracji danych numerycznych*.

W pracy sformułowano następującą hipotezę: *integracja metod analizy fleksyjnej tekstu oraz eksploracji danych numerycznych zwiększy nośność informacyjną danych w wielokryterialnym procesie wspomagania decyzji.*

Aby osiągnąć cel w pracy przyjęto następujące założenia:

- Proces decyzyjny *PD* jest oparty na eksploracji danych tekstowych i numerycznych,
- Dane poddawane eksploracji stanowią zbiór (zgodny ze wzorem *(1)*) wszystkich dostępnych danych w procesie *PD*,

- Podejmowana decyzja w procesie *PD* jest decyzją wielokryterialną (w szczególności jednokryterialną), a kryteria jej wyboru wynikają z dostępnych dla procesu eksploracji danych,
- Dane tekstowe w procesie *PD* są danymi opisanymi w języku fleksyjnym polskim, w którym ze względu na jego specyfikę możliwe jest występowania przestawnego szyku wyrazów w zdaniu zawartym w danych (dokumentach) tekstowych,
- Reprezentacja danych numerycznych wykorzystywana w eksploracji (wartości dyskretne oraz nominalne atrybutów) jest definiowana z uwzględnieniem struktury dziedziny wartości atrybutów, która odpowiada specyfice rozważanego problemu decyzyjnego,
- Elementami reprezentacji danych (dokumentów) tekstowych, która charakteryzuje dokument tekstowy, są rzeczowe informacje (ang. factual information), które mają bezpośredni wpływ na podejmowaną w procesie *PD* decyzję, przy czym rzeczowe informacje są tu rozumiane jako sekwencje wyrazów o zmiennej długości, które są ekstrahowane z dokumentów tekstowych na podstawie zdefiniowanych przez eksperta dziedzinowego wzorców informacyjnych,
- W pracy skoncentrowano się na metodzie eksploracji danych zwanej klasyfikacją, a to ze względu na mnogość oraz szeroki wachlarz zastosowań tych metod w procesach decyzyjnych *PD,*
- Ze względu na specyfikę wyniku klasyfikacji tj. występowanie zmiennych posiadających wyłącznie dwie kategorie (zmienne dychotomiczne), badanie prób zależnych danych oraz występowanie skali nominalnej zmiennych, do badań istotności i zgodności wyników klasyfikacji (weryfikacja statystyczna) zastosowano test McNemara.
- Z powodu istnienia w procesach decyzyjnych *PD* luki informacyjnej, brak jest możliwości dokładnego oszacowania nośności informacyjnej danych. Ponieważ nośność informacyjna nie tylko zależy od samych danych, ale również od wiedzy decydenta (możliwości interpretacyjnych tych danych) to wydaje się iż istnieje ścisła zależność iż im lepsza jest jakość eksploracji danych (np. klasyfikacji) tym nośność informacyjna danych na bazie których przeprowadzono tę eksplorację (klasyfikację) będzie wyższa. Dlatego w do weryfikacji hipotezy przyjęto założenie, że nośność informacyjna danych w procesie *PD* może być szacowana za pomocą wybranych miar jakości klasyfikacji.

5. Zadania do rozwiązania

W celu opracowania procedury integracji metod eksploracji danych tekstowych i numerycznych w procesie *PD* należało zrealizować następujące zadania cząstkowe:

1. przeprowadzić analizę i klasyfikację metod wykorzystywanych w eksploracji danych tekstowych oraz eksploracji danych numerycznych pod kątem możliwości ich integracji,

2. dokonać wyboru integrowanych metod eksploracji danych tekstowych i numerycznych,
3. opracować metodę budowania reprezentacji danych tekstowych wykorzystującą specyfikę języka polskiego (fleksyjnego),
4. opracować procedurę integracji wybranych metod eksploracji danych tekstowych i numerycznych oraz odpowiednią metodykę jej stosowania,
5. zweryfikować hipotezę postawioną w pracy empirycznie (studium przypadków dotyczących różnych dziedzin zastosowań opracowanej w pracy metody do różnych procesów PD).

6. Akceptacja wyników przez społeczność naukową

Czasopisma i materiały konferencyjne, w których recenzenci oceniali wynik cząstkowych badań:

- *Metody Informatyki Stosowanej*, Polska Akademia Nauk, 2010,
- *Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedzą*, 2010 oraz 2011,
- *Monografia oraz materiały konferencyjne Internet in the Information Society*, Wydawnictwo Wyższej Szkoły Biznesu w Dąbrowie Górniczej w Gliwicach, 2013 oraz 2015,
- *Theoretical and Applied Informatics*, Instytut Informatyki Teoretycznej i Stosowanej Polskiej Akademii Nauk w Gliwicach, 2013,
- *Ekonomiczne Problemy Usług*, Zeszytach Naukowych Uniwersytetu Szczecińskiego, 2013,
- *Monografia anglojęzyczna Data analysis. Selected Problems*, Sejmik Młodych Informatyków 2013,
- *Materiały konferencyjne Advances in Data Mining*, Ibai-publishing, 2015,
- *Lecture Notes in Artificial Intelligence*, Springer, 2015.

Konferencje na których wyniki badań zostały przedyskutowane:

- *VIII edycja konferencji Sejmik Młodych Informatyków*, Szczecin, 2013,
- *Internet w społeczeństwie informacyjnym*, Dąbrowa Górnicza, 2013,
- *15th Industrial Conference on Data Mining*, Hamburg, 2015,
- *10th International Conference Internet in the Information Society 2015*, Dąbrowa Górnicza, 2015,
- *7th International Conference Computational Collective Intelligence 2015*, Madryt, 2015.

7. Główne osiągnięcia pracy

W ramach pracy wniesiony został następujący wkład autorski:

- opracowano wieloaspektową i systemową procedurę integracji uwzględniającą metodę eksploracji (klasyfikacji) danych tekstowych i numerycznych w kontekście procesu *PD*, w którym dostępne są dwa typy danych,
- opracowano metodę budowy reprezentacji danych tekstowych w modelu przestrzeni wektorowej VSM z wykorzystaniem zintegrowanych metod bazujących na wiedzy eksperta (definiowanie wzorców informacyjnych) oraz uczeniu maszynowym (ekstrakcja i weryfikacja rzeczowych informacji na podstawie wzorców),
- opracowano metodę analizy fleksyjnej danych tekstowych wyrażonych w języku fleksyjnym polskim wykorzystywaną do weryfikacji poprawności wyekstrahowanych za pomocą wzorców rzeczowych informacji (elementów reprezentacji γ-gramowej).

8. Wartość teoretyczna

Będąca przedmiotem pracy doktorskiej procedura integracji metod eksploracji danych tekstowych i numerycznych umożliwia eksplorację obu wymienionych typów danych (wzór 1) w procesie podejmowania decyzji w sposób wieloaspektowy i systemowy, który jest jednocześnie adekwatny do tych typów. Opracowaną procedurę integracji można scharakteryzować następującymi własnościami:

- uwzględnia w eksploracji danych tekstowych specyfikę języka polskiego (fleksyjnego),
- dostosowuje budowę reprezentacji danych tekstowych do rozważanego problemu decyzyjnego,
- wykorzystuje zarówno podejście oparte na metodach uczenia maszynowego jak również metodach bazujących na wiedzy eksperta,
- wykorzystuje metody eliminujące szum informacyjny przy eksploracji danych tekstowych oraz numerycznych,
- wpływa na zwiększenie nośności informacyjnej eksplorowanych danych oraz jakość podejmowanych w procesie *PD* decyzji.

Opracowana w ramach rozprawy procedura integracji może być wykorzystywana jako standardowe podejście w różnego rodzaju procesach decyzyjnych, w których analizowane są różne typy danych tj. dane tekstowe, numeryczne oraz po transformacji na powyżej wymienione typy, również inne dane np. multimedialne.

W ramach pracy dokonano również analizy stanu wiedzy, w tym dostępnych metod wykorzystywanych do:

- tzw. wstępnej eksploracji danych, za pomocą której opracowywana jest reprezentacja danych tekstowych,
- właściwej eksploracji danych tekstowych z użyciem technik uczenia maszynowego,
- właściwej eksploracji danych tekstowych, bazujących na wiedzy eksperta,
- eksploracji danych numerycznych przy użyciu Teorii Zbiorów Przybliżonych,
- eliminacji tzw. szumu informacyjnego zarówno w eksploracji danych tekstowych jak i numerycznych,
- oceny nośności informacyjnej danych.

9. Wartość praktyczna

Opracowana w ramach rozprawy doktorskiej procedura integracji metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji może być z powodzeniem wykorzystywana w systemach komputerowego wspomagania decyzji DSS do rozwiązywania różnorodnych, rzeczywistych problemów decyzyjnych, co udowodniono na przykładzie trzech przypadków użycia.

W ramach badań testowych zweryfikowano również wpływ opracowanej procedury integracji na poprawę nośności informacyjnej eksplorowanych danych oraz jakości podejmowanych decyzji na podstawie przykładowych procesów *PD*.

10. Struktura pracy

Rozprawa doktorska składa się z siedmiu rozdziałów oraz spisu referencji, rysunków, tabel i symboli. Praca liczy 164 strony.

W rozprawie zamieszczono wykaz użytych referencji, który obejmuje 110 pozycji. Należą do nich artykuły konferencyjne, strony internetowe oraz pozycje książkowe (głównie anglojęzyczne). Praca zawiera 149 równań oraz 42 rysunki, 74 tabele i 124 symboli.

11. Ogólna charakterystyka zawartości pracy doktorskiej

W rozdziale pierwszym pracy przedstawiono rolę eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji. Podkreślono znaczenie problemu integracji metod eksploracji danych tekstowych i numerycznych oraz przedstawiono cel pracy i hipotezę badawczą.

Rozdział drugi zawiera opis metod eksploracji danych tekstowych bazujących na klasyfikacji przeprowadzonej z wykorzystaniem modelu przestrzeni wektorowej VSM. Rozdział rozpoczyna się wprowadzeniem do analizy tekstu i omówieniem podstawowych zagadnień związanych eksploracją danych tekstowych. W szczególności scharakteryzowano tu różne podejścia do klasyfikacji dokumentów tekstowych, zarówno metody bazujące na uczeniu maszynowym jak i metody wykorzystujące wiedzę eksperta.

W rozdziale trzecim opisano metody klasyfikacji danych numerycznych. W rozdziale tym skoncentrowano się na Teorii Zbiorów Przybliżonych, która umożliwia budowanie wiedzy wykorzystywanej do podejmowania decyzji w procesie *PD* na bazie reguł decyzyjnych. W szczególności opisano metody opracowania reprezentacji danych numerycznych (między innymi dyskretyzację danych numerycznych) oraz szczegółowo opisano metody wykorzystywane przy eliminacji szumu informacyjnego.

Przedmiotem rozdziału czwartego jest autorska procedura integracji metod klasyfikacji danych tekstowych i numerycznych w procesie podejmowania decyzji. W pierwszej kolejności został przedstawiony ogólny schemat procedury po czym szczegółowo opisano poszczególne etapy tej procedury. Zaprezentowano tu metody budowy reprezentacji danych

tekstowych oraz numerycznych. W szczególności opisano metodę opracowywania γ-gramowej reprezentacji danych tekstowych bazującą na wzorcach informacyjnych definiowanych przez eksperta dziedzinowego oraz analizie fleksyjnej rzeczowych informacji wyekstrahowanych z tekstu za pomocą wzorców. Kolejno opisano etap budowania systemu informacyjnego *SI*, na podstawie, którego generowana jest wiedza w procesie decyzyjnym *PD*.

W rozdziale piątym dokonano oceny różnych wariantów eksploracji (wariant A – z wykorzystaniem zintegrowanych metod eksploracji danych tekstowych i numerycznych, wariant B – z wykorzystaniem wyłącznie metody eksploracji danych tekstowych, wariant C – z wykorzystaniem wyłącznie metody eksploracji danych numerycznych, wariant D – z wykorzystaniem zintegrowanych wyników eksploracji z wariantów B i C), w oparciu wyniki badań testowych dotyczących przykładowych procesów podejmowania decyzji *PD*. W pierwszej części analizę poddano przykład, którego celem było wyszukiwanie rentownych zamówień publicznych w Biuletynie Zamówień Publicznych, kolejno przykład dotyczący inwestowania na Giełdzie Papierów Wartościowych oraz przykład związany z wyszukiwaniem atrakcyjnych ofert pracy.

W rozdziale szóstym przeprowadzono dyskusję wyników uzyskanych w badaniach przypadków użycia.

Zakończenie pracy stanowi rozdział siódmy zawierający podsumowanie badań nad opracowaną procedurą integracyjną, sformułowano wnioski z realizacji celu pracy i weryfikacji postawionej hipotezy**.**

12. Procedura integracji metod klasyfikacji danych tekstowych i numerycznych w procesie podejmowania decyzji

12.1. Ogólny schemat procedury

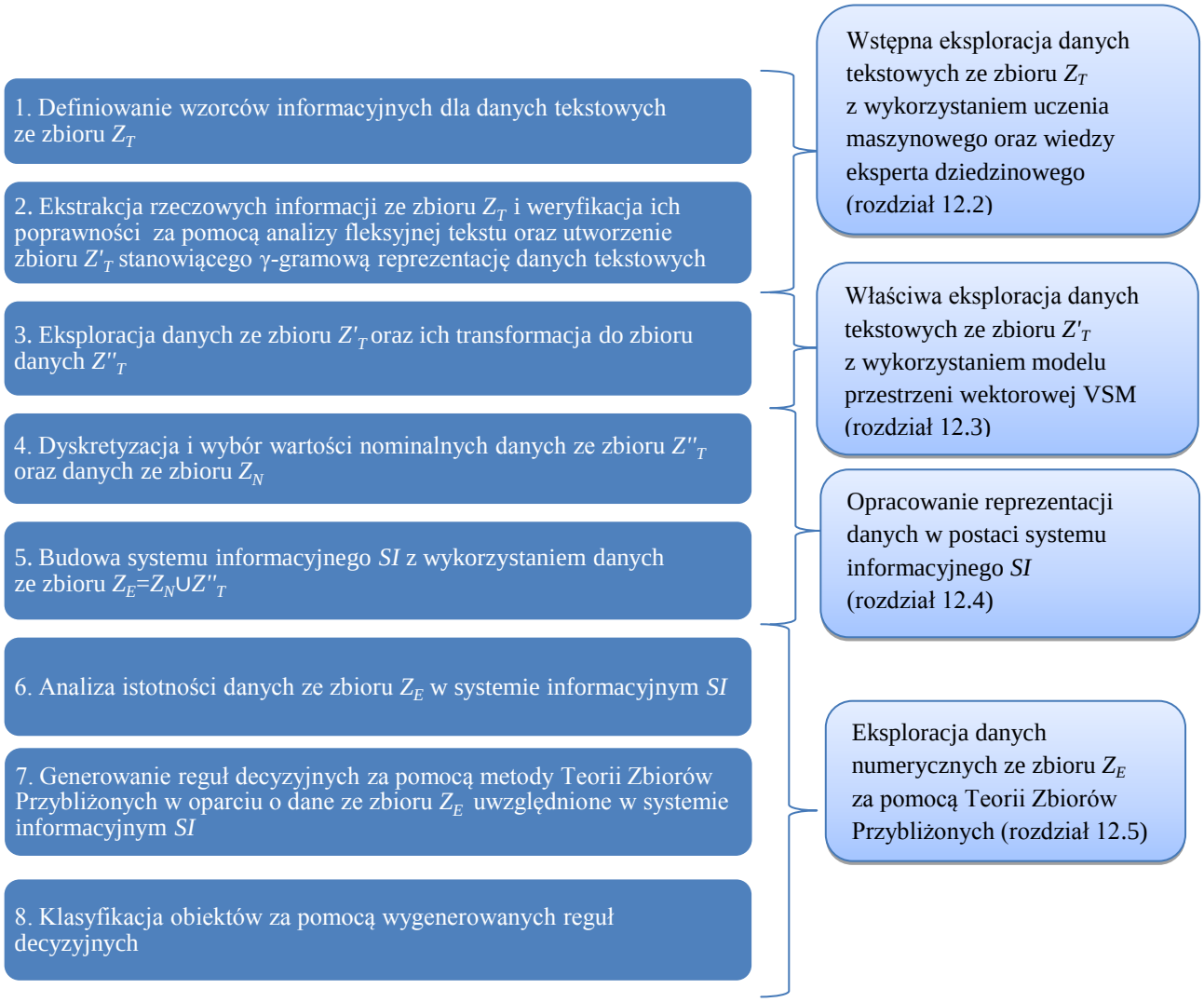
Procedura integracji metod klasyfikacji danych tekstowych i numerycznych w procesie podejmowania decyzji składa się z 8 podstawowych etapów, które zaprezentowano na rysunku 1.

W dalszej części autoreferatu przedstawiono przykładową implementację procedury z rysunku 1 z uwzględnieniem opisu poszczególnych etapów tej procedury.

12.2. Wstępna eksploracja danych tekstowych

Na wynik eksploracji duży wpływ ma zastosowanie odpowiedniej reprezentacji danych. W przypadku eksploracji danych tekstowych bazującej na modelu przestrzeni wektorowej VSM kluczowe jest ustrukturyzowanie danych polegające na wyborze odpowiedniej ich reprezentacji. Z tego względu etapy 1 i 2 z rysunku 1, realizowane w ramach tzw. wstępnej eksploracji danych tekstowych ze zbioru Z_T, są wykorzystywane do opracowania właściwej względem rozpatrywanego problemu decyzyjnego reprezentacji

danych tekstowych. W przygotowaniu ustrukturyzowanej reprezentacji danych tekstowych w modelu przestrzeni wektorowej VSM przeważnie wykorzystywane są automatyczne metody uczenia maszynowego. Ze względu na to, że w literaturze podkreśla się korzyści wynikające również z zastosowania przy eksploracji danych tekstowych metod bazujących na wiedzy eksperta np. wzorców informacyjnych [9, s. 127], co potwierdzają również wyniki analizy SWOT z tabeli 1, w procedurze zgodnej z rysunkiem 1 wykorzystano reprezentację γ-gramową, która jest opracowywana z wykorzystaniem zarówno metod bazujących na wiedzy eksperta jak i metod uczenia maszynowego.



Rysunek 1. Etapy procedury integracji metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji.
Źródło: opracowanie własne

W pierwszej kolejności dla danych ze zbioru Z_T, uwzględniając kontekst decyzyjny, definiowane są przez eksperta dziedzinowego wzorce informacyjne. Wzorce są ogólnym modelem najbardziej istotnej informacji semantycznej (rzeczowych informacji) przenoszonej przez tekst względem rozpatrywanego problemu decyzyjnego, która traktowana jest jako układ wybranych wyrazów [14, s. 137]. Formalny zapis wzorców stanowi mechanizm, który automatyzuje i ułatwia definiowanie różnych sekwencji wyrazów, na podstawie których

budowana jest reprezentacja tekstu. W literaturze przedstawiono kilka propozycji formalizacji zapisu wzorców informacyjnych [25, ss. 345–358] [14, ss. 155–164]. W pracy wykorzystano do tego celu języka OWL (ang. Web Ontology Language).

Standardowo elementy reprezentacji γ-gramowej budowane są w oparciu o funkcję oceniającą $\gamma(w_1,\ldots,w_{lw})$, której wartości odpowiadają przydatności danej sekwencji wyrazów $w_1,\ldots,w_{lw}$ w analizie dokumentu tekstowego, natomiast w pracy elementy reprezentacji budowane są z rzeczowych informacji (sekwencji wyrazów o zmiennej długości) wyekstrahowanych z tekstów za pomocą wzorców informacyjnych. W opracowanej procedurze z rysunku 1 wykorzystano zatem metody dotyczące budowy wzorców informacyjnych oraz ekstrakcji na ich podstawie tzw. rzeczowych informacji, które stanowią elementy nowej reprezentacji danych tekstowych w modelu przestrzeni wektorowej VSM [5, s. 1].

Po zdefiniowaniu wzorców informacyjnych przez eksperta dziedzinowego realizowany jest etap 2 procedury z rysunku 1, który schematycznie został przedstawiony na rysunku 2.

W pracy segmentacja polegająca na przekształceniu tekstów z formy ciągłej na zdania oraz pojedyncze wyrazy wykonywana jest na podstawie zdefiniowanych wyrażeń regularnych.

Lematyzacja bazuje na słowniku fleksyjnym i polega na wyszukiwaniu kolejnych wyrazów z tekstu w bazie wyrazów Słownika Języka Polskiego - SJP.PL [34], a następnie pobraniu jego formy podstawowej.

Po przeprowadzeniu lematyzacji budowany jest zbiór wszystkich wyrazów występujących w dokumentach tekstowych przy jednoczesnym zachowaniu ich podstawowych (po lematyzacji) i oryginalnych form fleksyjnych oraz informacji o granicach zdań. Następnie, w etapie 2.3 z rysunku 2, utworzony zbiór wyrazów redukowany jest wyłącznie do wyrazów występujących we wzorcach. Pozostałe wyrazy zostają uznane za szum informacyjny i zostają pominięte.

W etapie 2.4 z rysunku 2 w utworzonym zbiorze wyszukiwane są wyrazy odpowiadające nazwom klas określonym we wzorcach zdefiniowanych przy użyciu języka OWL. Odbywa się to na podstawie listy wszystkich unikalnych nazw klas występujących we wzorcach. Po zidentyfikowaniu danej nazwy w zbiorze wyrazów reprezentujących dokument tekstowy wyszukiwane są zawierające ją wzorce.

W etapie 2.5 z rysunku 2 dla wzorców generowane są oczekiwania, czyli kolejne elementy, które znajdują się w ich pełnej definicji. Z listy wyszukanych wzorców usuwane są te, dla których nie odnaleziono pełnej listy oczekiwań zgodnej z ich całkowitą definicją.

W kolejnym etapie 2.6 z rysunku 2, na podstawie zweryfikowanej listy wzorców, zostają wyekstrahowane z poszczególnych zdań wszystkie możliwe rzeczowe informacje (sekwencje wyrazów zdefiniowane za pomocą wzorców) z uwzględnieniem podstawowych oraz oryginalnych form fleksyjnych wyrazów.

W etapie 2.7 z rysunku 2, przy użyciu technik uczenia maszynowego dla wydobytych z tekstu rzeczowych informacji określona zostaje poprawności dopasowania do siebie form fleksyjnych poszczególnych wyrazów wchodzących w ich skład. Jest to analiza fleksyjna wyekstrahowanych rzeczowych informacji, która polega na automatycznym generowaniu list skojarzeniowych dla wyekstrahowanych z tekstu oryginalnych form fleksyjnych wyrazów.

Tabela 1. Analiza SWOT metod klasyfikacji danych tekstowych.

Mocne strony (zalety)	**Słabe strony (wady)**
Metody automatycznej eksploracji tekstów (VSM): ▪ krótki czas opracowania modelu, ▪ automatyczna ocena podobieństwa, ▪ możliwość zastosowania funkcji istotności w stosunku do elementów reprezentacji tekstu, co wpływa na poprawę jakości wyniku eksploracji. Metody automatycznej eksploracji tekstów (bez VSM): ▪ mała złożoność obliczeniowa, ▪ prosta implementacja. Metoda definiowania przez eksperta wzorców i reguł dopasowania (metody bazujące na wiedzy eksperta): ▪ dokładna analiza znaczeniowa tekstu, ▪ precyzyjne porównanie na poziomie rzeczowych informacji wyekstrahowany za pomocą wzorców informacyjnych, ▪ relatywnie mała ilość rzeczowych informacji i reguł dopasowania w stosunku do reprezentacji tekstu w metodach automatycznych.	Metody automatycznej eksploracji tekstów (VSM): ▪ niedokładna analiza znaczeniowa tekstu, ▪ nieprecyzyjne porównanie dokumentów tekstowych na poziomie pojedynczych wyrazów lub automatycznie wykrytych struktur semantycznych, ▪ duża wielkość reprezentacji i występowanie związanego z nią szumu informacyjnego wpływającego na obniżenie jakości wyniku eksploracji. Metody automatycznej eksploracji tekstów (bez VSM): ▪ ostre wyszukiwanie - nieodnajdywanie tekstów, które prawie pasują do zapytania, przypadkowa kolejność wyszukanych dokumentów tekstowych, trudność ograniczenia wielkości odpowiedzi, ▪ niska skuteczność przy długich tekstach Metoda definiowania przez eksperta wzorców i reguł dopasowania (metody bazujące na wiedzy eksperta): ▪ długi czas opracowania modelu wynikający z ręcznego definiowania reguł dopasowania, ▪ intuicyjne (nieprecyzyjne przy dużej ilości reguł) definiowanie reguł dopasowania, ▪ ze względu na ręczną definicję brak możliwości zastosowania funkcji istotności w stosunku do elementów wzorców.
Szanse	**Zagrożenia**
▪ możliwość integracji różnych rodzajów metod w celu zwiększenia ich użyteczności oraz jakości wyniku eksploracji	▪ uzyskanie wyniku eksploracji o niskiej jakości.

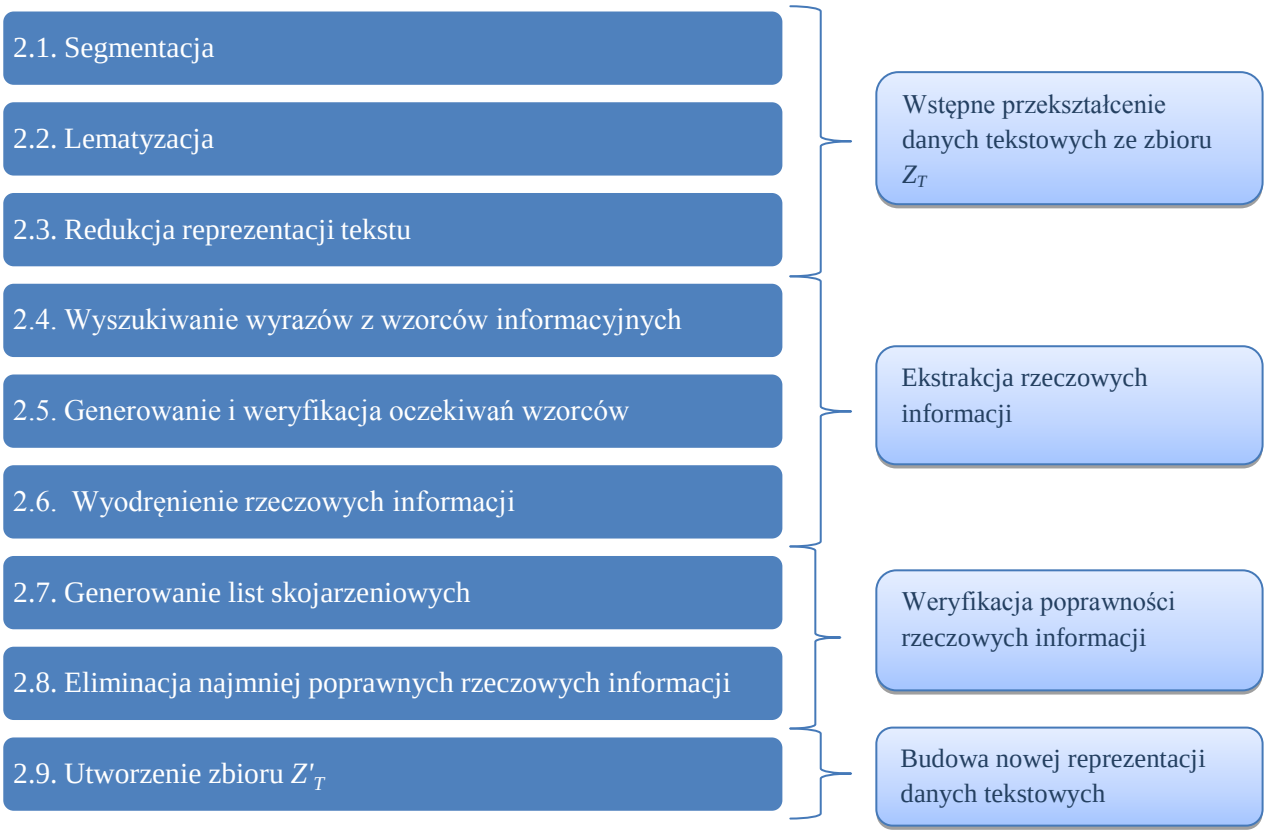
Źródło: opracowanie własne

Podstawą metody generowania list skojarzeniowych jest statystyczno-matematyczne wyliczenie miary skojarzeniowej dla form fleksyjnych elementów z kolejnych trójek (podmiot, orzeczenie/właściwość, obiekt) zawierających się w poszczególnych wzorcach, zgodnie ze wzorem (2).

$$sk = \frac{cw}{lw} \qquad (2)$$

gdzie:

sk – wyrażona w procentach miara skojarzenia,
cw – częstość względna określonej formy fleksyjnej nazwy podmiotu i obiektu występujących razem w zdaniach,
lw – częstość bezwzględna uwzględniająca wszystkie formy fleksyjne nazwy obiektu, które występują w zadaniach z określoną formą fleksyjną nazwy podmiotu.



Rysunek 1. Części składowe etapu 2 procedury z rysunku 1
Źródło: opracowanie własne

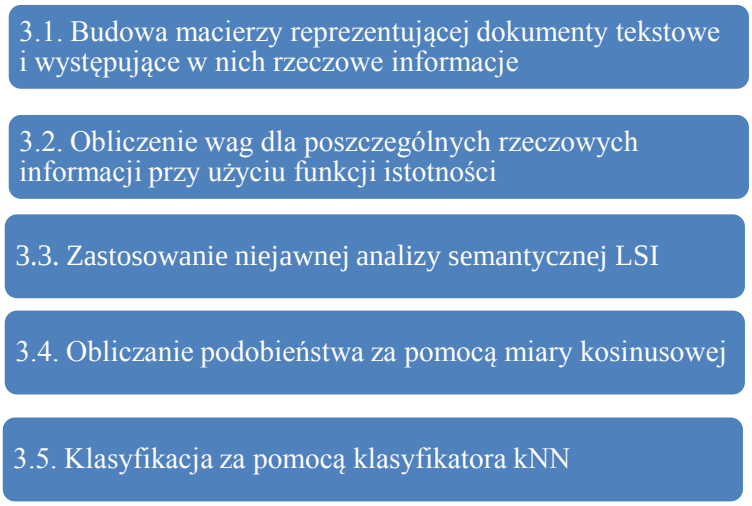
Innymi słowy miara skojarzeniowa uwzględnia liczbę par wyrazów odpowiadającym nazwom podmiotów i obiektów w konkretnych formach fleksyjnych występujących w zdaniach. Obliczona miara skojarzeniowa jest wskazaniem na najbardziej poprawne formy fleksyjne wyrazów dla danej rzeczowej informacji.

W etapie 2.8 z rysunku 2 w celu wyeliminowania najmniej poprawnych form fleksyjnych wyrazów wyekstrahowanych według poszczególnych wzorców informacyjnych eksperymentalnie dopierany jest próg (wartość graniczna miary skojarzeniowej), który decyduje o uwzględnieniu lub odrzuceniu rzeczowej informacji w dalszej części procedury eksploracji.

Rezultatem działania etapu 9 z rysunku 2 jest γ-gramowa reprezentacja danych tekstowych tworząca zbiór Z'_T.

12.3. Właściwa eksploracja danych tekstowych

Schematycznie kolejne podetapy etapu 3 procedury z rysunku 1 zaprezentowano na rysunku 3.



Rysunek 2. Części składowe etapu 3 procedury z rysunku 1
Źródło: opracowanie własne

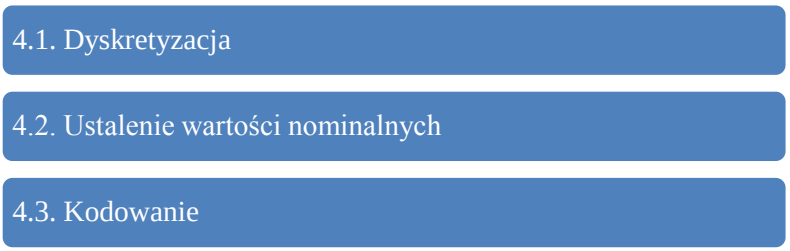
W celu przeprowadzenia właściwej eksploracji danych tekstowych w pierwszej kolejności (etap 3.1 z rysunku 3) opracowywana jest macierz reprezentująca rzeczowe informacyjne oraz dokumenty tekstowe. Wielkość macierzy określa liczba dokumentów tekstowych oraz liczba wyekstrahowanych z dokumentów tekstowych rzeczowych informacji, które uwzględniają w swojej budowie różne formy fleksyjnej wyrazów.

Kolejno dobierane są odpowiednie metody i techniki eksploracji, które umożliwią uzyskanie wyniku o jak najwyższej jakości. Na wynik uzyskany za pomocą określonych metody eksploracji danych tekstowych mają wpływ poszczególne techniki wykorzystywane w ramach tych metod. Jest to szczególnie istotne w przypadku eksploracji danych tekstowych z wykorzystaniem modelu przestrzeni wektorowej VSM, w której istnieje możliwość wykorzystania wielu różnych technik wspomagających eksplorację. Wybór poszczególnych technik może być zrealizowany za pomocą metody eksperymentalnej (na podstawie wyniku eksploracji) lub poprzez wskazanie eksperta wynikające z określonych zależności, wymagań lub ograniczeń np. ograniczeń wydajnościowych jednostki obliczeniowej. Przykładowe techniki wykorzystywane w eksploracji danych tekstowych to: funkcja istotności nadająca odpowiednie wagi elementom reprezentacji danych tekstowych, niejawna indeksacja semantyczna, obliczanie podobieństwa dokumentów tekstowych za pomocą funkcji kosinusowej czy klasyfikacja dokumentów tekstowych za pomocą klasyfikatora kNN.

Ostatecznie, po wybraniu odpowiednich technik eksploracji danych tekstowych przeprowadzana jest klasyfikacja danych tekstowych, w ramach której otrzymywany jest zbiór danych Z''_T stanowiący wynik eksploracji danych tekstowych ze zbioru Z'_T.

12.4. Opracowanie reprezentacji danych numerycznych

Opracowanie reprezentacji przetransformowanych danych tekstowych (dane ze zbioru Z''_T) oraz danych numerycznych ze zbioru Z_N w etapie 4 procedury z rysunku 1 przebiega zgodnie ze schematem z rysunku 4.



Rysunek 3. Części składowe etapu 4 procedury z rysunku 1
Źródło: opracowanie własne

W pierwszej kolejności, w podetapie 4.1 z rysunku 4, dane numeryczne w formie ciągłej zostają poddane dyskretyzacji. W opracowaniu przyjęto metodę dyskretyzacji opartą o podział o równej częstości, czyli podział dziedziny atrybutów (danych numerycznych) na przedziały zawierające równą liczbę przypadków (obiektów), która w niektórych sytuacjach została doprecyzowana przez eksperta.

W podetapie 4.2 z rysunku 4, w przypadku atrybutów, które nie podlegają dyskretyzacji przypisywane są wartości nominalne.

Na podstawie nowej reprezentacji danych w postaci zakodowanych wartości nominalnych i dyskretnych (dane ze zbiorów Z''_T oraz Z_N) wszystkich atrybutów poddawanych eksploracji zostaje utworzony zbiór Z_E.

W etapie 5 procedury z rysunku 1, dane ze zbioru Z_E zostają uwzględnione w systemie informacyjnym *SI* w postaci tablicy decyzyjnej *TD*. Tablica decyzyjna *TD* jest zdefiniowana za pomocą wzoru [19] *(3)*.

$$TD = (U, Aw, ad, V, f), gdzie\ Aw, ad \in A; Aw \neq \emptyset; Aw \cup ad = A;\ Aw \cap ad = \emptyset \qquad (3)$$

gdzie:

A – skończony, niepusty zbiór atrybutów,
Aw – zbiór atrybutów warunkowych,
ad – atrybut decyzyjny, którego wartości wyznaczają klasy decyzyjne – zbiór decyzji oznaczony jako $\{d\}$,
f – funkcja decyzyjna,
U – skończony, niepusty zbiór obiektów,
V – zbiór wartości atrybutów ze zbioru A,
$\emptyset$ – zbiór pusty.

12.5. Klasyfikacja danych w procesie decyzyjnym

Po etapie opracowania reprezentacji danych numerycznych, na bazie tablicy decyzyjnej *TD* oraz współczynników wynikających z metody Teorii Zbiorów Przybliżonych,

obliczana jest istotność danych uwzględnionych w systemie informacyjnym *SI*, która obliczana jest zgodnie z etapami:

1. obliczanie ogólnej istotności danych,
2. obliczanie istotności poszczególnych atrybutów.

Do obliczenia ogólnej istotności danych ze względu na jakość wiedzy zawierającej się w systemie informacyjnym *SI* wykorzystywane są współczynniki jakości i dokładności rodziny konceptów decyzyjnych, zdefiniowanych za pomocą wzorów (4) oraz (5), natomiast do obliczenia istotności poszczególnych atrybutów, ze szczególnym uwzględnieniem atrybutu, który wyraża wynik eksploracji danych tekstowych, używany jest współczynnik istotności, określony za pomocą wzoru [20] (6).

$$\gamma_B(X) = \frac{card(POS_B(X))}{card(U)} \tag{4}$$

gdzie:

X – aproksymowany zbiór obiektów,
U – uniwersum, skończony, niepusty zbiór obiektów $u_1, u_2 \ldots u_{mu}$,

mu –liczba obiektów z uniwersum,

card(POS$_B$(X)) – liczba obiektów (obiektów) w pozytywnym obszarze zbioru *X*,
card(U) – liczba obiektów w uniwersum *U*.

$$\beta_B(X) = \frac{card(POS_B(X))}{card(\overline{BX})} = \frac{card(\underline{B}X)}{card(\overline{B}X)} \tag{5}$$

gdzie:

X – aproksymowany zbiór obiektów,
card(POS$_B$(X)) – liczba obiektów (obiektów) w pozytywnym obszarze zbioru *X*,
$card(\underline{B}X)$ – liczba przypadków należących do dolnego przybliżenia zbioru *X*,
$card(\overline{B}X)$ – liczba przypadków należących do górnego przybliżenia, znajdujących się w pozytywnym oraz brzegowym obszarze zbioru *X*.

$$\sigma_{Aw}(a_{ia}) = 1 - \frac{\gamma_{Aw}(X) - \gamma_{Aw-\{a_{ja}\}}(X)}{\gamma_{Aw}(X)} \text{ , o ile } 1 < ia, ja < card(Aw), ia \neq ja \tag{6}$$

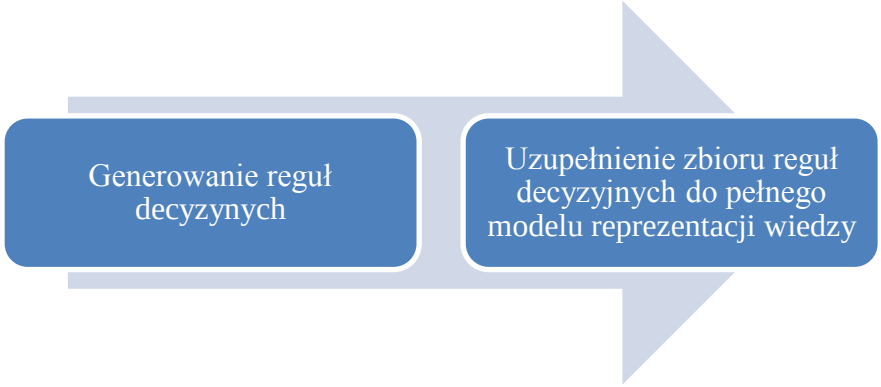
gdzie:

$\gamma_{\mathrm{Aw}}(\mathrm{X})$ – współczynnik jakości atrybutów ze zbioru Aw,
card(Aw) – liczba atrybutów warunkowych *Aw*,
a – atrybut warunkowy, jeśli $a \in Aw$,
X – aproksymowany zbiór obiektów,
ia, ja – indeksy atrybutów.

Dobór atrybutów reprezentujących dane numeryczne w celu uzyskania systemie informacyjnym *SI* wiedzy wykorzystywanej do podejmowania decyzji w procesie *PD* o jak najwyższej jakości możliwy jest na podstawie powyższych współczynników.

Po ostatecznym opracowaniu systemu informacyjnego *SI* przy użyciu metody TZP w etapie 7 procedury z rysunku 1 generowane są reguły decyzyjne. Zdefiniowanie zbioru

reguł decyzyjnych pokrywających pełną dziedzinę analizowanych zależności odbywa się zgodnie ze schematem przedstawionym na rysunku 5.



Rysunek 4. Etapy definiowania pełnego zbioru reguł decyzyjnych
Źródło: opracowanie własne

W pierwszym etapie przedstawionym na rysunku 5, generowane, a następnie upraszczane są reguły decyzyjne, które wykorzystywane są w dalszej eksploracji danych. W przypadku generowania reguł z ograniczonej ilości danych eksperymentalnych istnieje ryzyko braku pokrycia pełnej dziedziny analizowanych zależności. Dlatego w celu uzupełnienia zbioru reguł dla obiektów, dla których nie istnieje właściwa reguła wyekstrahowana z danych eksperymentalnych zostaje utworzona reguła wyznaczona przez eksperta dziedzinowego.

Wynikiem operacji przeprowadzonych w etapie 7 procedury z rysunku 1 jest wiedza zapisana w postaci reguł decyzyjnych, na podstawie której możliwe jest podejmowanie ostatecznych decyzji w procesie *PD*.

Ostatecznie w etapie 8 procedury z rysunku 1, za pomocą reguł wygenerowanych przy użyciu Teorii Zbiorów Przybliżonych, obiekty zostają poddane klasyfikacji.

13. Badania testowe

Badania testowe miały na celu potwierdzenie hipotezy badawczej postawionej w pracy. W związku z tym w ramach badań testowych zostały porównane cztery warianty eksploracji, tj.:

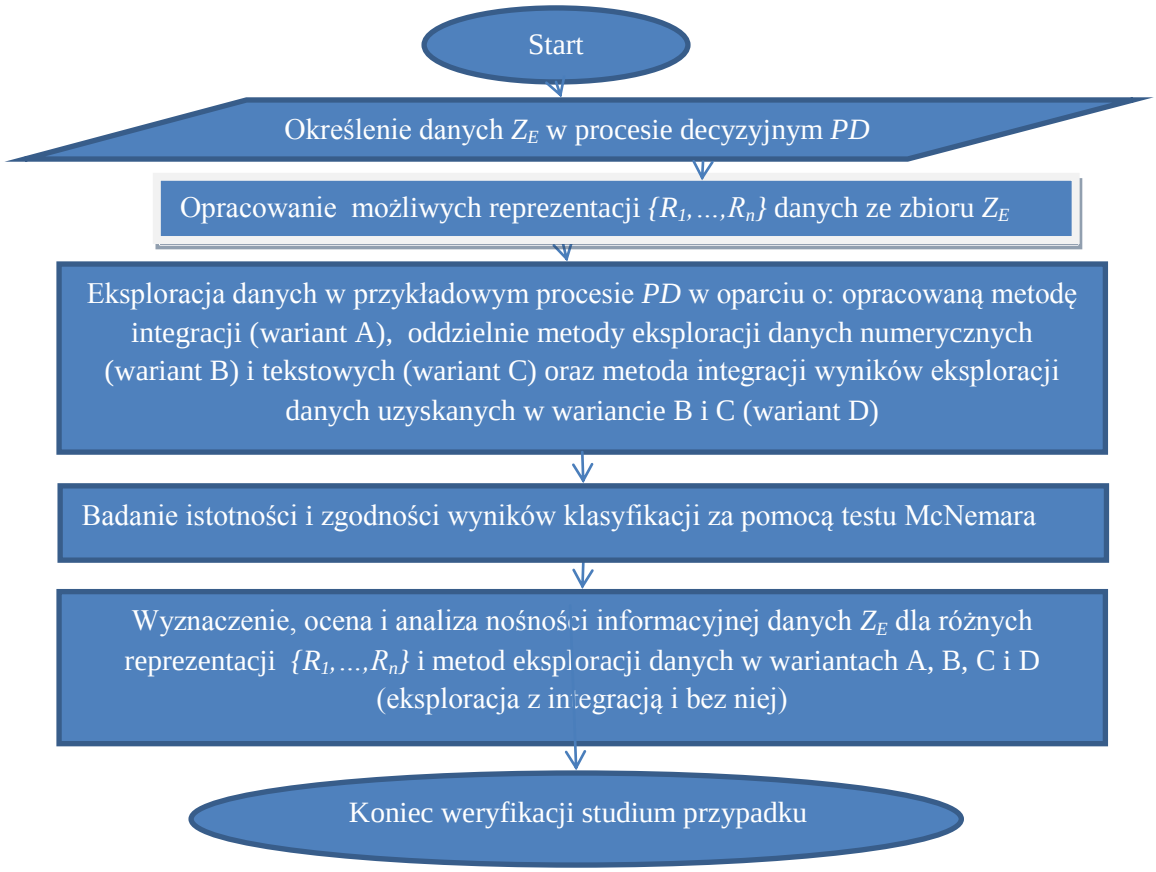
Wariant A. Zintegrowana eksploracja danych tekstowych i numerycznych,
Wariant B. Eksploracja wyłącznie danych numerycznych,
Wariant C. Eksploracja wyłącznie danych tekstowych.
Wariant D. Zintegrowany wynik eksploracji z wariantów B i C.

W przypadku obliczania Wariantu D integracja odnosi się do procedury scalania wyników z Wariantów B i C, który polega na porównaniu wyniku eksploracji danych numerycznych z wynikami eksploracji danych tekstowych na poszczególnych poziomach zwrotu (ang. recall), a następnie na wyborze korzystniejszego wyniku (wynik Wariantu B lub wynik Wariantu C dla danego poziomu zwrotu) w przypadku 11 kolejnych porównań (11 poziomów zwrotu). Ostatecznie 11 wybranych wyników jest zsumowanych i uśrednionych.

W wariantach A, C oraz D eksploracji, w których użyto eksploracji danych tekstowych wykorzystano trzy różne reprezentacje danych tekstowych, tj.:

- Reprezentację unigramową - uwzględniającą pojedyncze wyrazy,
- Reprezentację n-gramową (bigramową) – uwzględniającą sekwencje dwóch wyrazów,
- Reprezentację γ-gramową - uwzględniającą sekwencje wyrazów o zmiennej długości, opracowaną za pomocą wzorców informacyjnych definiowanych przez eksperta dziedzinowego oraz metod analizy fleksyjnej tekstu.

Ogólny algorytm weryfikacji hipotezy został zaprezentowany na rysunku 6.



Rysunek 6. Ogólny algorytm weryfikacji hipotezy dla każdego studium przypadków (przypadki I, II i III)
Źródło: opracowanie własne

W badaniach wykorzystano trzy różne przykładowe procesy podejmowania decyzji (trzy przypadki użycia), w których możliwe było zastosowanie zintegrowanej eksploracji danych tekstowych i numerycznych, tj.:

- przykład I. proces decyzyjny dotyczący wyboru rentownych zamówień publicznych spośród zbioru takich zamówień,
- przykład II. proces decyzyjny dotyczący sposobu inwestowania na Giełdzie Papierów Wartościowych,
- przykład III. proces decyzyjny dotyczący wyszukiwania atrakcyjnych ofert pracy.

Pierwsze zadanie decyzyjne (przykład I) polegało na wytypowaniu zamówień publicznych z bazy Biuletynu Zamówień Publicznych (BZP) [33], w których opis przedmiotu zamówienia jest tożsamy z zakresem prac wykonywanych przez podmiot. W rozważanym przykładzie

proces *PD* jest realizowany przez firmę, która wykonuje usługi w zakresie mechanicznego koszenia traw, głównie koszenia w pasach drogowych przy czym usługi takie firma wykonuje za pomocą kosiarek bijakowych lub rotacyjnych doczepianych do ciągników. W procesie wyboru zamówień dostępnych w BZP dużą trudnością jest odróżnienie opisów przedmiotu zamówienia dotyczących ręcznego koszenia traw od opisów dotyczących mechanicznego koszenia trawy wyłącznie z ręcznym obkaszaniem słupków, barier i pozostałych tego typu elementów. Dodatkową trudnością jest wybór takich zamówień z BZP, które według wybranych kryteriów numerycznych tj. obszaru do koszenia, odległości od siedziby firmy oraz wartości zamówienia, wskazują na rentowność. Badanie testowo przeprowadzano na zbiorze 200 testowych przypadków zamówień publicznych oraz z wykorzystaniem łącznie 22 przypadków treningowych (11 przypadków dla kategorii rentownych i 11 przypadków dla kategorii pozostałych zamówień).

Drugie zadanie decyzyjne (przykład II) polegało na wytypowaniu komunikatów wybranych spółek giełdowych, w których podano informacje o tym, że osoba blisko związana z zarządem dokonała transakcji na akcjach spółki. W tym przypadku jedną z większych trudności jest odróżnienie komunikatów o dokonaniu transakcji na akcjach oraz na kontraktach terminowych, jak również dokonaniu tych transakcji przez inne podmioty. Dodatkowo ocenie podlegały dane numeryczne tj. wielkości sprzedaży akcji oraz cena sprzedaży akcji opublikowane w komunikacie giełdowym. Atrybutem decyzyjnym jest w tym przypadku cena akcji w dniu emisji komunikatu giełdowego. Na podstawie eksploracji danych tekstowych oraz numerycznych możliwe jest np. odkrycie reguł stwierdzających, że w przypadku znaczącej ilości sprzedaży akcji po cenie niższej od zakupu przez osobę blisko związaną z zarządem spółki rynek zazwyczaj będzie reagował negatywie, czyli ceny akcje spółki po emisji takiego komunikatu będą spadały. Badanie testowo przeprowadzano na zbiorze 200 testowych przypadków transakcji z wykorzystaniem łącznie 22 przypadków treningowych (11 dla kategorii oznaczającej spadek kursu akcji oraz 11 dla kategorii oznaczającej utrzymanie lub wzrost cen akcji).

Ostatnie zadanie decyzyjne (przykład III) wykorzystane w badaniach testowych dotyczyło wyszukiwania atrakcyjnych ofert pracy umieszczonych w serwisach www.pracuj.pl i www.praca.pl. Zadanie poległo na wytypowaniu ofert pracy, w których opis jest zgodny z charakterystyką pracy pracownika ds. rekrutacji. O atrakcyjności oferty oprócz zgodności opisu tekstowego świadczyły dane numeryczne w postaci liczby mieszkańców miejscowości, w której znajduje się miejsce pracy oraz poziomu hierarchii stanowiska (*stażysta, specjalista, kierownik*). Biorąc pod uwagę wartości wymienionych kryteriów badana jest atrakcyjność pracy względem historycznych wskazań kandydata. Badanie testowo przeprowadzano na zbiorze 200 testowych przypadków ofert pracy oraz łącznie 22 przypadków treningowych (11 przypadków treningowych dla kategorii oznaczającej atrakcyjne oferty oraz 11 dla kategorii z pozostałymi przypadkami).

Zintegrowana eksploracja danych tekstowych i numerycznych (Wariant A) została przeprowadzona w 8 etapach, zgodnych ze schematem procedury integracji przedstawionym na rysunku 1.

Do zbadania istotności i wiarygodności wyników różnych wariantów eksploracji (Warianty: A, B, C, D), wykorzystano test statystyczny MyNemara. Jest to test nieparametryczny, który pozwala na badanie prób zależnych z uwzględnieniem zmiennych

dychotomicznych tj. zmiennych posiadających tylko dwie kategorie. Analiza statystyczna z wykorzystaniem testu MyNemara wykazała, że dla wszystkich przypadków użycia w większości porównań różnice pomiędzy wynikami poszczególnych wariantów są istotne.

Po określeniu czy wyniki dla porównywanych wariantów eksploracji różnią się w sposób istotny statystycznie, została przeprowadzana weryfikacja hipotezy badawczej polegająca na ocenie nośności informacyjnej danych, zdefiniowanej za pomocą wzoru (7).

$$Ninf(Z_E, M) = g(W_M, R, \zeta) \tag{7}$$

gdzie:

Ninf – nośność informacyjna danych Z_E,
Z_E – zbiór dostępnych danych (zbiór danych w procesie *PD*, wyrażonych za pomocą danych tekstowych lub numerycznych),
R – reprezentacja danych Z_E,
W_M – wynik procesu *PD* określony za pomocą miar jakości decyzji w tym procesie,
M – zbiór wybranych miar jakości decyzji w procesie *PD*,
ζ – niezidentyfikowany zbiór informacji mających wpływ na nośność informacyjną danych *Ninf*,
g – funkcja, której wartość określa nośność informacyjną *Ninf*.

Wyznaczenie nośności informacyjnej danych w procesie decyzyjnym *PD* ze wzoru *(7)* jest trudne ze względu na brak znajomości funkcji *g* oraz zbioru ζ. Dlatego do szacowania nośności informacyjnej danych dla różnych reprezentacji tych danych stosuje się metody porównawcze. Dla przykładu nośność informacyjna danych Z_E o reprezentacji R_1 jest wyższa od nośności informacyjnej danych Z_E o reprezentacji R_2 w przypadku osiągnięcia w procesie podejmowania decyzji *PD* korzystniejszego wyniku W_M dla danych reprezentowanych przez R_1 w stosunku do R_2. Taką zasadę pomiaru nośności informacyjnej przyjęto w niniejszej pracy.

Wynik W_M ze wzoru *(4)* w eksploracji danych jest określany przez zbiór wybranych miar jakości decyzji *M* [22, s. 206]. Przyjęte miary jakości decyzji uzależnione są od rozpatrywanego problemu decyzyjnego i przyjętej metody oceny jakości wyniku procesu *PD* (oceny jakości decyzji) [27, s. 88][11]. W pracy do oceny nośności informacyjnej danych wykorzystano współczynnik całkowitej dokładności – ACC oraz współczynnik całkowitego poziomu błędu – ERR, zdefiniowane za pomocą wzorów *(8)* i *(9)*, które są najczęściej wykorzystywanymi miarami jakości klasyfikacji wynikającymi z macierzy pomyłek.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$ERR = \frac{FP+FN}{FP+FN+TP+TN} \tag{9}$$

gdzie:

TP – współczynnik prawdziwy pozytywny,
FP – współczynnik fałszywy pozytywny,
TN – współczynnik prawdziwy negatywny,
FN – współczynnik fałszywy negatywny.

Współczynniki TP, FP, TN oraz FN są wskazaniami systemu klasyfikacji , którego specyfikę przedstawiono w tabeli 2.

Tabela 2. Specyfika system klasyfikacji z wykorzystaniem wielu miar jakości

Kategoria	Stan rzeczywisty: TAK	Stan rzeczywisty: NIE
Klasyfikacja systemowa: TAK	TP	FP
Klasyfikacja systemowa: NIE	FN	TN

Źródło: [18, ss. 21–22]

W ramach weryfikacji hipotezy zostały porównane wyniki badań testowych dotyczące przykładowych procesów *PD* (przykłady: I, II oraz III) w postaci poszczególnych średnich wartości miar jakości decyzji (*ACC, ERR*) osiągniętych dla czterech wariantów eksploracji tj.:

1. zintegrowanej eksploracji danych tekstowych i numerycznych (Wariant A),
2. eksploracji wyłącznie danych numerycznych (Wariant B),
3. eksploracja wyłącznie danych tekstowych (Wariant C),
4. metody integracji wyników eksploracji danych uzyskanych w wariancie B i C (Wariant D),

z uwzględnieniem trzech różnych reprezentacji danych tekstowych (reprezentacji unigramowej, bigramowej i γ-gramowej).

Procentową różnicę pomiędzy wartościami miar jakości decyzji *ACC* oraz *ERR* osiągniętymi dla wariantu A, a wartościami tych miar jakości decyzji uzyskanymi dla pozostałych wariantów eksploracji (warianty B, C oraz D przedstawiono w tabelach 3, 4 oraz 5. W porównaniu wykorzystano współczynniki Δ_{ACC} oraz Δ_{ERR} wyrażone procentowo, zgodne ze wzorami *(10)* oraz *(11)*.

$$\Delta_{ACC} = ACC_{WN} - ACC_{WM} = 1 - \Delta_{ERR} \tag{10}$$

lub

$$\Delta_{ERR} = ERR_{WM} - ERR_{WN} = 1 - \Delta_{ACC} \tag{11}$$

gdzie:

ACC_{WN} , ACC_{WM} – współczynnik całkowitej dokładności dla wybranego wariantu *WN* oraz *WM* (wariant A, B, C, D),

ERR_{WM} , ACC_{WN} – współczynnik całkowitego poziomu błędu dla wybranego wariantu *WM* oraz *WN* (wariant A, B, C, D).

Tabela 3. Porównanie wartości miar jakości decyzji osiągniętych w 1 przykładzie procesu *PD* w wariancie A eksploracji z wartościami miar jakości decyzji uzyskanymi w wariantach B, C, D

Reprezentacja (dotyczy wariantów, w których wykorzystywane są dane tekstowe)	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**B**) [%]	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**C**) [%]	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**D**) [%]
unigramowa	15	19	15
n-gramowa	14	16	14

γ-gramowa	20	16	16

Źródło: opracowanie własne

Tabela 4. Porównanie wartości miar jakości decyzji osiągniętych w 2 przykładzie procesu *PD* w wariancie A eksploracji z wartościami miar jakości decyzji uzyskanymi w wariantach B, C, D

Reprezentacja (dotyczy wariantów, w których wykorzystywane są dane tekstowe)	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**B**) [%]	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**C**) [%]	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**D**) [%]
unigramowa	3	23	3
n-gramowa	12	36	12
γ-gramowa	16	25	16

Źródło: opracowanie własne

Tabela 5. Porównanie wartości miar jakości decyzji osiągniętych w 3 przykładzie procesu *PD* w wariancie A z wartościami miar jakości decyzji uzyskanymi w wariantach B, C, D

Reprezentacja (dotyczy wariantów, w których wykorzystywane są dane tekstowe)	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**B**) [%]	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**C**) [%]	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**D**) [%]
unigramowa	5	42	5
n-gramowa	3	35	3
γ-gramowa	12	34	12

Źródło: opracowanie własne

Porównanie z uwzględnieniem współczynnika Δ pomiędzy wartościami miar jakości decyzji dla reprezentacji γ-gramowej w stosunku do pozostałych reprezentacji danych tekstowych w wariancie zintegrowanej eksploracji danych tekstowych i numerycznych (wariant A) i wariancie eksploracji wyłącznie danych tekstowych (wariant C), zaprezentowano w tabelach 6, 7 oraz 8.

Tabela 6. Porównanie wartości miar jakości decyzji osiągniętych w 1 przypadku procesu *PD* z wykorzystaniem reprezentacji γ-gramowa w stosunku do wartości miar jakości decyzji uzyskanymi dla reprezentacji unigramowej i n-gramowej w wariantach A i C (eksploracji tylko danych tekstowych)

Reprezentacja danych tekstowych	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**A**) [%]	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**C**, WM=**C**) [%]
γ-gramowa dla WN i **unigramowa** dla WM	5	8
γ-gramowa dla WN	6	6

i **n-gramowa** dla WM		

Źródło: opracowanie własne

Tabela 7. Porównanie wartości miar jakości decyzji osiągniętych w 2 przypadku procesu *PD* z wykorzystaniem reprezentacji γ-gramowa w stosunku do wartości miar jakości decyzji uzyskanymi dla reprezentacji unigramowej i n-gramowej w wariantach A i C (eksploracji tylko danych tekstowych)

Reprezentacja danych tekstowych	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**A**) [%]	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**C**, WM=**C**) [%]
γ-gramowa dla WN i **unigramowa** dla WM	13	11
γ-gramowa dla WN i **n-gramowa** dla WM	4	15

Źródło: opracowanie własne

Tabela 8. Porównanie wartości miar jakości decyzji osiągniętych w 3 przypadku procesu *PD* z wykorzystaniem reprezentacji γ-gramowa w stosunku do wartości miar jakości decyzji uzyskanymi dla reprezentacji unigramowej i n-gramowej w wariantach A i C (eksploracji tylko danych tekstowych)

Reprezentacja danych tekstowych	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**A**, WM=**A**) [%]	$\Delta_{ACC}/\Delta_{ERR}$ (WN=**C**, WM=**C**) [%]
	ACC/ERR	ACC/ERR
γ-gramowa dla WN i **unigramowa** dla WM	7	15
γ-gramowa dla WN i **n-gramowa** dla WM	9	10

Źródło: opracowanie własne

Ostatecznie dokonano porównania najniższych wartości miar jakości decyzji (osiągniętych w przykładowych procesach *PD*: I, II oraz III) w wariancie eksploracji danych tekstowych i numerycznych (wariant A) w stosunku do najwyższych wartości miar jakości osiągniętych w pozostałych wariantach (warianty B, C oraz D). Procentowe porównanie miar jakości decyzji zaprezentowano w tabeli 9.

Tabela 9. Porównanie wartości miar jakości decyzji osiągni
ętych w wariancie A (procedura integracyjna metod eksploracji danych tekstowych i numerycznych) z wartościami miar jakości decyzji uzyskanymi w wariancie eksploracji B (eksploracja tylko danych tekstowych), C (eksploracja tylko danych numerycznych) i D (integracja wyników uzyskanych w wariancie B i C)

Przypadek procesu *PD*	$\Delta_{ACC}/\Delta_{ERR}$ **(WN=A, WM=B) [%]**	$\Delta_{ACC}/\Delta_{ERR}$ **(WN=A, WM=C) [%]**	$\Delta_{ACC}/\Delta_{ERR}$ **(WN=A, WM=D) [%]**
I	14	10	10
II	3	12	3
III	3	25	3

Źródło: opracowanie własne

Mając na uwadze powyższe wyniki badań testowych można uznać, że hipoteza postawiona w pracy dla rozważanych przypadków użycia (przykładowe procesy *PD*: I, II oraz III) została zweryfikowana, ponieważ:

- miary jakości decyzji (*ACC, ERR*) eksploracji danych dla przykładowych procesów podejmowania decyzji *PD* w wariancie eksploracji z wykorzystaniem integracja metod eksploracji danych tekstowych i numerycznych (wariant A) są wyższe w przypadku miary *ACC* oraz niższe w przypadku miary *ERR* od takich miar w pozostałych wariantach B (eksploracja oparta tylko na danych numerycznych w procesie *PD*) , C (eksploracja oparta tylko na danych tekstowych) oraz wariancie D integrującym wyniki wariantu B, C, co potwierdzają zestawienia zawarte w tabelach od 3 do 5,
- nośność informacyjna mierzona miarą *ACC* oraz *ERR* eksploracji danych w przypadku reprezentacji γ-gramowej, opracowanej za pomocą wzorców informacyjnych oraz analizy fleksyjnej języka polskiego jest wyższa w porównaniu do nośności informacyjnej danych osiągniętej dla reprezentacji unigramowej i bigramowej, co potwierdzają zestawienia danych w tabelach od 6 do 8,
- nawet w przypadku najmniej korzystnego wyniku (wyrażonego miarami *ACC* i *ERR*) dla wariantu eksploracji danych tekstowych i numerycznych (wariant A), we wszystkich przypadkach procesów *PD*, jakość wyniku jest wyższa w stosunku do jakości najkorzystniejszych wyników eksploracji danych pozostałych wariantów (warianty B, C i D), co potwierdza zestawienie zawarte w tabeli 9.

14. Spis publikacji własnych

[1] Gibert M., Śmiałkowska B.: Wycena domen internetowych z wykorzystaniem teorii zbiorów przybliżonych, Metody Informatyki Stosowanej nr 4/2010 (25), Polska Akademia Nauk Oddział w Gdańsku, Komisja Informatyki, Szczecin, 2010, s. 11-24

[2] Gibert M., Śmiałkowska B.: Kryteria i parametry procesu pozycjonowania stron WWW, Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedzą Nr 28, Polskie Stowarzyszenie Zarządzania Wiedzą, Bydgoszcz, 2010, s. 58-70

[3] Gibert M., Śmiałkowska B.: Method for making decisions on investing on the internet domain market with use of the fuzzy sets theory, Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedzą Nr 572010, Polskie Stowarzyszenie Zarządzania Wiedzą, Bydgoszcz, 2011, s. 372-383

[4] Śmiałkowska B., Gibert M.: The classification of text documents in Polish language by using Latent Semantic Analysis for extracted information, Internet in the Information Society Computer Systems Architecture and Security, Wydawnictwo Wyższej Szkoły Biznesu w Dąbrowie Górniczej, Gliwice, 2013, s.57-68

[5] Śmiałkowska B., Gibert M.: The classification of text documents in Polish language by using Latent Semantic Analysis for extracted information, Theoretical and Applied Informatics (wersja rozszerzona), Instytut Informatyki Teoretycznej i Stosowanej Polskiej Akademii Nauk w Gliwicach, Gliwice 2013,

[6] Śmiałkowska B., Gibert M.: The base of knowledge in the internet domain name rating system by using extraction and inflectional classification of words from domain names, Ekonomiczne Problemy Usług, Zeszytach Naukowych Uniwersytetu Szczecińskiego, Szczecin, 2013, s. 345-358

[7] Śmialkowska B., Jankowski J., Gibert M.: The Classification of Text Documents by Using Latent Semantic Analysis for Extracted Information, Monografia anglojęzyczna Data analysis. Selected Problems, SMI 2013, Szczecin, 2013,

[8] Gibert M.: Improving quality of text mining by using ontology for terms selection in the Vector Space Model, 10th International Conference Internet in the Information Society 2015, s.143-152, Dąbrowa Górnicza, 2015

[9] Gibert M.: The Influence of Data Information-carrying Capacity on Quality of Text Mining, Advances in Data Mining/15th Industrial Conference on Data Mining, ICDM 2015, s.35-43, Hamburg, 2015

[10] Gibert M.: Improving Information-Carrying Data Capacity in Text Mining, Lecture Notes in Artificial Intelligence/7th International Conference Computational Collective Intelligence 2015, Tom 9330, s.648-657, Springer, 2015

15. Zakończenie

W pracy ujęto obszerny przegląd aktualnego stanu badań wraz z przeglądem poszczególnych metod i technik dotyczących eksploracji bazującej na klasyfikacji danych tekstowych oraz numerycznych wykorzystywanych w procesach podejmowania decyzji. W opracowaniu skoncentrowano się na klasyfikacji ze względu na jej znaczącą rolę w zbiorze metod eksploracji danych wspierających procesy podejmowania decyzji *PD*. W analizie szczególną uwagę poświęcono procesowi opracowania reprezentacji eksplorowanych danych, w którym również stosowane są metody tzw. wstępnej eksploracji danych. Za pomocą wstępnej eksploracji, w której wykorzystywane jest zarówno wiedza eksperta dziedzinowego jak
i metody uczenia maszynowego możliwe jest opracowanie reprezentacji danych tekstowych dostosowanej (w sensie wykorzystania jako elementów reprezentacji jedynie informacji, które mają istotny wpływ na podejmowaną decyzję w procesie *PD*) do rozpatrywanego problemu decyzyjnego. W pracy scharakteryzowano najczęściej wykorzystywane reprezentacje danych tekstowych:

- unigramową,
- n-gramową,
- γ-gramową.

Podkreślono podział metod eksploracji danych tekstowych na dwie główne grupy, z których jedna bazuje na uczeniu maszynowym, a druga na wiedzy eksperta. Następnie szczegółowo opisano eksplorację danych numerycznych bazującą na Teorii Zbiorów

Przybliżonych. W pracy skoncentrowano się na metodach opracowania reprezentacji danych numerycznych w systemie informacyjnym, ze szczególnym uwzględnieniem technik eliminujących szum informacyjnych.

W ramach rozprawy opracowano autorską procedurę integracji metod klasyfikacji danych tekstowych i numerycznych, która zwiększa nośność informacyjną danych w procesie podejmowania decyzji. Metoda ta została oparta na wiedzy eksperta, analizie fleksyjnej danych tekstowych dostępnych w procesie decyzyjnym *PD* oraz eksploracji danych numerycznych.

Badania testowy przeprowadzone w ramach niniejszej pracy, z wykorzystaniem trzech przykładowych procesów podejmowania decyzji, potwierdziły przedstawioną hipotezę tj. integracja metod analizy fleksyjnej tekstu oraz metod eksploracji danych numerycznych zwiększa nośność informacyjną danych w procesie podejmowania decyzji. Ponadto nośność informacyjna danych mierzona za pomocą miar jakości decyzji (*ACC, ERR*) jest wyższa w przypadku zastosowania γ-gramowej reprezentacji danych tekstowych, która jest opracowana z wykorzystaniem zarówno wiedzy eksperta (zdefiniowany przez eksperta wzorców informacyjnych) jak i metod uczenia maszynowego (ekstrakcja rzeczowych informacji oraz ich weryfikacja przy użyciu analizy fleksyjnej) w stosunku do pozostałych badanych reprezentacji (unigramowej, n-gramowej - bigramowej). Wykazano również, że dzięki użyciu analizy fleksyjnej tekstu możliwe jest wyodrębnienie poprawnych rzeczowych informacji, które mają istotny wpływ na podejmowane decyzji w procesie *PD*.

W kontekście rozwiązywanego w pracy problemu badawczego interesujące jest podjęcie dalszych badań nad opracowaniem systemu wspomagającego proces decyzyjny *PD*, opartego na opracowanej w pracy procedurze uwzględniającej γ-gramową reprezentację danych tekstowych oraz na automatycznym wyszukiwaniu informacji tekstowych a także numerycznych związanych z procesem *PD* i dostępnych w treściach stron internetowych tworzonych w ramach tzw. semantycznego Internetu z użyciem języka OWL.

16. Bibliografia

[1] Adhikari A., Adhikari J., *Advances in Knowledge Discovery in Databases*. Springer, 2015

[2] Aggarwal C.C., Zhai C.: *Mining Text Data.* Springer, 2012

[3] Bechhofer S., Harmelen F., Hendler J., Horrocks I., McGuinness L., Patel-Schneider P.F., Stein A. L. *OWL*. [on line: 16.09.2016]. Dostępny w Internecie: www.w3.org/TR/owl-ref

[4] Berry M., Linoff G.: *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley & Sons, New York, 2000

[5] Dong X. L., Gabrilovich E., Murphy K., Dang V., Horn W., Lugaresi C., Sun S., Zhang W.: *Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources*. Cornell University Library, 2015

[6] Gabrys B., Howlett R. J., Jain L. C.: *Analysis of Stock Price Retudn Using Textual Data and Numerical Data Through Text mining*. Springer, Berlin, 2006

[7] Gawrysiak P.: *Automatyczna kategoryzacja dokumentów*. Uniwersytet Warszawki, 2001

[8] Hand D. J., Mannila H., Padhraic S.: *Principles of Data Mining*. Massachusetts Institute of Technology Press, 2001

[9] Jackson P., Moulinier I.: *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins Publishing, Amsterdam, 2007

[10] Janakiraman V. S., Sarukesi K.: *Decision Support Systems*. PHI Learning Pvt. Ltd., New Delphi, 2008
[11] Katu V., Deshpande B.: *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Elsevier, Waltham, 2015
[12] Langford G. O.: *Engineering Systems Integration: Theory, Metrics, and Methods*. CRC Press, Boca Raton, 2012
[13] Libal U.: *Algorytmy rozpoznawania obrazów - Praktyczna ocena jakości klasyfikacji*. Politechnika Wrocławska, 2015
[14] Lubaszewski W.: *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu*. Wydawnictwo AGH, Kraków, 2009
[15] Miner G.: *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press, Waltham, 2012
[16] Ming Fai Wang F., Liu Z., Chiang M.: *Stock Market Prediction from WSJ: Text Mining via Sparse Matrix Factorization*. Cornell University Library, 2014
[17] Nettleton D.: *Commercial Data Mining: Processing, Analysis and Modeling for Predictive Analytics Projects*. Elsevier, Waltham , 2014
[18] Neustein A.: *Text Mining of Web-Based Medical Content*. Walter de Gruyter GmbH & Co KG, 2014
[19] Nowak A.: *Teoretyczne podstawy zbiorów przybliżonych*. [on line: 16.09.2016]. http://zsi.tech.us.edu.pl/~nowak/se/konspektTD.pdf
[20] Piegat A.: *Materiały z wykładów Teorii Zbiorów Przybliżonych*. Zachodniopomorski Uniwersytet Technologiczny w Szczecinie, 2010
[21] Power D. J., *Decision Support Systems: Concepts and Resources for Managers*. Greenwood Publishing Group, Westport, 2002
[22] Raghavan V., Bollmann P., Jung G. S.: *A critical investigation of recall and precision as measures of retrieval system performance*. ACM Transactions on Information Systems (TOIS), tom 7, nr 3, New York, 1989
[23] Schumaker R. P., Chen H.: *Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System*, ACM Transactions on Information Systems (TOIS), tom 27, nr 2, New York, 2009
[24] Sołdacki P.: *Zastosowanie metody płytkiej analizy tekstu do przetwarzania dokumentów w języku polskim*. Politechnika Warszawska, 2006
[25] Śmiałkowska B., Gibert M.: „The classification of text documents by using Latent Semantic Analysis for extracted information", Ekonomiczne Problemy Usług, tom 6, 2013
[26] Thomas J. D., Sycara K.: *Integration Genetic Algorithms and Text learning for Financial Prediction*, GECCO-2000 Workshop on Data Mining with Evolutionary Algorithms, 1999
[27] Tittel E., Noble J.: *HTML, XHTML and CSS For Dummies*. Wiley Publishing Inc., Indiana, 2008
[28] Wang R. Y., Strong D. M.: *Beyond Accuracy: What data quality means to data consumers*. Journal of Management Information Systems, tom 12, nr 4, 1996
[29] Weiss S. M., Indurkhya N., Zhang T.: *Fundamentals of Predictive Text Mining*. Springer, London, 2010

[30] Wuthrich B., Permunetilleke D., Leung S., Cho V., Zhang J., Lam W.: *Daily Prediction of Major Stock Indices from textual WWW Data*, KDD-98 Proceedings, AAAI, New York, 1998
[31] Yin Y., Kaku I., Tang J., Zhu J., *Data Mining: Concepts, Methods and Applications in Management and Engineering Design.* Springer, London, 2011
[32] Yuan J.: *Image and Video Data Mining*, Northwestern University, Illinois, 2009
[33] *Biuletyn Zamówień Publicznych*. [on line: 16.09.2016]. Dostępny w Internecie: http://bzp1.portal.uzp.gov.pl
[34] Słownik Języka Polskiego. [on line: 16.09.2016]. Dostępne w Internecie: www.sjp.pl