GŁADYSZ Anna[1]

# Zastosowanie metod eksploracyjnej analizy tekstu w logistyce

*Logistyka, informacja, dane, przepływy informacji, text mining, worek słów, macierz częstości, model przestrzeni wektorowej, grupowanie dokumentów, grupowanie nazw towarów, towar, waga słowa, miary odległości, współczynnik Jaccarda, metryka euklidesowa, miara kosinusowa, metryka Manhattan, genetyczne algorytmy grupowania, algorytmy oparte o listy decyzyjne,*

**Streszczenie**

*Przepływ informacji, który jest w dzisiejszych czasach niezbędny do działania organizacji na rynku wspomagany powinien być technikami informatycznymi. Wiedza, która jest konieczna do pozyskania na rynku logistyki jest pochodną danych i informacji, którymi dysponuje przedsiębiorstwo logistyczne. Niezbędne jest szybkie zbieranie danych, gromadzenie ich oraz przetwarzanie i analiza nie tylko wewnątrz firmy ale również czerpanie wiedzy z zewnątrz. Pomocne w tym zakresie mogą okazać się metody bazujące na eksploracyjnej analizie danych tekstowych – text mining. Zaprezentowano możliwe przykłady wykorzystania i korzyści płynące z zastosowania metod text miningowych w logistyce. W artykule przedstawiono zagadnienie grupowania towarów/produktów, jako atrybutów, według których dokonuje się porównywania, oceny i grupowania kontrahentów. Dokonano zarówno przeglądu jak i przedstawiono propozycje metod i algorytmów grupowania w katalogu towarów.*

**THE USE OF METHODS TEXT MINING IN THE LOGISTICS**

**Abstract**

*Necessary the flow of the information which is in contemporary times to the working of the organization should be helped computer technicians. The knowledge which is essential to gaining over on the field of the logistics is the derivative data and information which the logistic company disposes. He indispensable is fast data setting, collection their and processing and analysis not only inside the firm but also draws knowledge from outside. Helpful in this field can prove methods of the text mining. Introduced the possible examples of exploitation and advantages flowing from the use of methods text mining in the logistics. In the paper introduced the question of clustering goods/ products as the attributes according to which achieved comparing, estimation and clustering contractors. It was achieved both the inspection as and the introduced proposals of methods and the algorithms of clustering in the catalog of goods.*

## 1. WSTĘP

W dzisiejszych czasach informacja na równi z kapitałem, pracą i wyposażeniem stała się podstawową wartością w przedsiębiorstwie, a zarazem warunkiem konkurencyjności firmy. Szczególne dbanie o nią, jej prawidłowy przepływ i synchronizację powinno stanowić podstawową kwestię przedsiębiorstwa. Dlatego tak duży nacisk kładzie się na rzetelność i dokładność pozyskiwanych informacji, bo dzięki temu proces podejmowania trafnych decyzji w czasie rzeczywistym w przedsiębiorstwie jest łatwiejszy. Powinna ona, z racji swojej funkcji, dostarczać zarówno odpowiedzi na pytania z zakresu funkcjonowania organizacji, a także w pełni zaspokajać potrzeby i oczekiwania obecnych jak i przyszłych klientów. Z dużym zasobem informacji mają do czynienia zarówno działy logistyki przedsiębiorstw produkcyjnych i handlowych, jak również firmy świadczące usługi logistyczne na zewnątrz. Tylko dzięki zastosowaniu nowoczesnych rozwiązań informatycznych można sprawnie i efektywnie kierować każdym przedsiębiorstwem, zwłaszcza w zakresie logistyki. Wydaje się być czymś nieodzownym, aby firma zajmująca się szeroko pojętą logistyką, posiadała narzędzia informatyczne, bo tylko wtedy jest w stanie w pełni wykorzystywać wszystkie możliwości jakie podsuwa współczesna koncepcja logistyki. W pełni efektywne wykorzystanie posiadanych informacji wpływa na zwiększoną możliwość adaptacji i elastyczne reagowanie na zachodzące zmiany w świadomości klientów.

## 2. INFORMACJA W LOGISTYCE

Zanim zostanie przedstawiona rola informacji w logistyce, warto w sposób skrótowy przyjrzeć się pojęciom ściśle związanym z tym tematem. Przedstawiona zostanie geneza i znaczenie słów: logistyka i informacja. Opisane zostanie eksploracyjna analiza danych tekstowych, wraz z podaniem obszarów jej zastosowania w logistyce.

[1] Politechnika Rzeszowska im. Ignacego Łukasiewicza, Wydział Zarządzania; 35-959 Rzeszów; al. Powstańców Warszawy 12.
Tel: + 48 48 865 10 89, E-mail: anna.gladysz@prz.edu.pl

### 2.1 Definicja logistyki

Genezę znaczenia pojęcia logistyki warto zacząć od zrozumienia znaczenia jego słowa. Jednym ze źródeł tego terminu jest greckie słowo *logistikós*, które oznacza człowieka myślącego według reguł logicznych – zarówno matematycznych jak i filozoficznych. Łaciński przymiotnik *logisticus* oznacza zaś zrozumiały, racjonalny, zdolny do logicznego myślenia. W starożytnym Rzymie *logistikas* oznaczało zaopatrzenie legionów w środki do życia, co wiązało się ściśle z planowaniem tras przemieszczania wojska. Samo więc zdefiniowanie logistyki wymaga uwzględnienia nie tylko nauki, ale musi obejmować także obszar jej stosowania. Aby więc w pełni zdefiniować logistykę niezbędne jest szersze spojrzenie.

W literaturze spotkać można szereg definicji współczesnej koncepcji logistyki. Przytoczone zostały trzy, które wzajemnie się uzupełniają:

- „Logistyka jest procesem planowania, realizacji i kontrolowania sprawności i ekonomicznej efektywności przepływu surowców, produkcji nie zakończonej i wyrobów gotowych oraz związanych z tym informacji od miejsca pochodzenia do miejsc konsumpcji w celu zaspokojenia wymagań klientów."[2]
- "Logistyka to wspomaganie procesów zarządzania w okresie życia produktu zapewniające efektywne wykorzystanie zasobów i adekwatną skuteczność elementów logistycznych podczas wszystkich faz okresu użytkowania, w ten sposób dzięki integrowaniu we właściwym czasie w system zapewnia się efektywne sterowanie zużyciem zasobów" (Society of Logistics Engineers (SOLE)).
- „Logistyka to interdyscyplinarna dziedzina wiedzy na styku techniki, informatyki i ekonomii, która integruje przepływy strumieni materiałów, informacji i kapitału w celu podniesienia produktywności i konkurencyjności przedsiębiorstwa na rynku" (Zbigniew Korzeń).

Przesłanki, które spowodowały wyodrębnienie się logistyki jako dyscypliny naukowej to:

- postępująca produkcyjność zasobów,
- duża sprawność procesów gospodarczych,
- powstanie konkurencyjnego rynku konsumenta,
- masowa komputeryzacja wszelkich sfer życia społeczno-gospodarczego.

Zachodzący w ostatnich latach dynamiczny rozwój zastosowań logistyki doprowadził do wyodrębnienia dwóch filarów logistyki, wynikających z jej aspektu naukowego oraz praktycznego: teorii i techniki. Teoria podaje warunki dla istnienia jakiegoś obiektu, technika zaś podaje środki niezbędne do wytworzenia tego obiektu. Oba, choć w pełni samodzielne, wzajemnie się uzupełniają, a pozbawione siebie tracą sens. Na uwagę zwraca również fakt, jak ważna jest odpowiednia informacja na rynku usług logistycznych.

### 2.2 Informacja

Informacja jest procesem złożonym i nie można podać jednej, powszechnie akceptowanej jej definicji. Niedzielski twierdzi, że informacja jest specyficznym dobrem niematerialnym oraz czynnikiem, który – jako swoista „metaenergia" – może przyczynić się do przeobrażenia gospodarki świata[3]. Sienkiewicz określa zaś informację, jako zbiór faktów, zdarzeń, cech itp. określonych obiektów (rzeczy, procesów, systemów) zawarty w wiadomości (komunikacie), tak ujęty i podany w takiej postaci (formie), że pozwala odbiorcy ustosunkować się do zaistniałej sytuacji, także podjąć odpowiednie działania umysłowe lub fizyczne[4].

Aby zdefiniować znaczenie informacji należy zacząć od jej konkretnej reprezentacji, czyli danych. Dane to surowe, nie poddane analizie fakty, liczby, zdarzenia, z których można opracować informacje. Czyste, nie opracowane dane nie mają większego znaczenia praktycznego. Rozwój technologii i idąca w ślad za tym komputeryzacja przedsiębiorstw znacznie ułatwiają i przyspieszają proces zarządzania danymi. Z drugiej strony stanowią pokusę do gromadzenia zbyt wielu zbędnych danych.

Dane w systemie informatycznym to reprezentacja informacji zapisana w pewnym obszarze pamięci komputera. Dane mogą reprezentować pojedynczą informację np. imię lub nazwisko albo zespół powiązanych ze sobą informacji np. adres zamieszkania. Same dane jednak bez podania kontekstu ich wykorzystania niewiele znaczą dla użytkownika systemu informatycznego. Dopiero opisanie danych (określenie ich kontekstu) poprzez metadane nadaje znaczenie tym danym.

Dane są podstawą do wykorzystania ich przy tworzeniu informacji. Przedstawiają obiektywne opisy stanów rzeczy. Dane opisują część zdarzenia, nie zawierają osądu, nie służą też jako baza działania i postępowania. Informacje to powstające w wyniku interpretacji danych wiadomości, które są rozumiane przez odbiorcę i mają dla niego znaczenie.

Informacja ustala i umiejscawia znaczenie danych w określonym kontekście i środowisku [6]. Słowo informacja pochodzi od łacińskiego *informare*, co oznacza „nadawać formę". Koncepcja informacji sformułowana została na potrzeby telekomunikacji. Znane jest twierdzenie Claude E. Shannona matematycznie definiujące możliwość przesyłania danej ilości informacji (bitów) w określonym pasmie częstotliwości.

Dla wiadomości elementarnych ilość informacji $I$ można wyrazić wzorem:

[2] Council of Logistics Management (CLM) – stowarzyszenie grupujące logistyków w USA przyjęło w 1986 roku przytoczoną definicję logistyki.
[3] Niedzielski E.: *Próba systematyzacji procesów rozwoju systemów informatycznych*, Wiadomości Statystyczne, nr 4, 1986, s. 20-23.
[4] Sienkiewicz P.: *10 wykładów*, AON, Warszawa 2005, s. 62.

$$I(x_i) = \log \frac{1}{p(x_i)} = -\log p(x_i) \tag{1}$$

gdzie: $x_i (i = \langle 1, N \rangle)$ - zdarzenie

$p(x_i)$ - prawdopodobieństwo związane z zachodzącym zdarzeniem.

Z podstaw ilościowego ujęcia informacji można wnioskować, że:

- jeżeli $x_i$ jest określone, tj. $p(x_i) = 1$, wówczas $I(x_i) = 0$,
- im mniejsze prawdopodobieństwo wystąpienia danego zdarzenia elementarnego, tym większa ilość informacji z nim związana,
- mając dwa zdarzenia niezależne $x_i$ i $x_j$ o prawdopodobieństwie $p(x_i, x_j)$ wówczas:

$$I(x_i, x_j) = p(x_i) + p(x_j). \tag{2}$$

Istotne jest również semantyczne znaczenie informacji. Informacja umożliwia zmianę sposobu postrzegania rzeczy, obiektów, wpływa na osąd i to odróżnia ją od danych. Ma tak ważną wartość, ponieważ w danej sytuacji decyzyjnej pozwala z większym prawdopodobieństwem ocenić skutki podjętych decyzji, a przez to umożliwia bardziej optymalne ich podejmowanie[5].

W logistyce wszelkim przepływom materialnym towarzyszą przepływy informacji. Pełnią one funkcję wspomagającą w stosunku do przepływów rzeczowych. Przepływy informacyjne zajmują więc kluczowe miejsce w organizowaniu efektywnych łańcuchów dostaw, będąc jednocześnie jednym z najtrudniejszych obszarów do projektowania i wdrażania [3]. Dzięki sprawnie zorganizowanemu systemowi informacyjnemu logistyki, przedsiębiorstwa mogą podejmować optymalne decyzje w zakresie zaopatrzenia, transportu i dystrybucji. Zarządzana informacja powoduje, że działania są sprawne i skuteczne, a obszary logistyki, które mogą posłużyć za potwierdzenie tego faktu to między innymi [15]:

- wielowymiarowa segmentacja klientów,
- analiza wartości i zadowolenia klientów,
- jakość produktów i usług,
- kontrola kosztów logistycznych,
- łańcuch dostaw,
- zasoby ludzkie,
- katalogi towarów/produktów,
- analiza wykorzystania kanałów elektronicznych.

Logistyczny system informacyjny to zbiór wzajemnie ze sobą powiązanych elementów: ludzi, sprzętu i procedur, zapewniający organom zarządzania logistyką przedsiębiorstwa odpowiednie informacje niezbędne do planowania, realizacji i kontrolowania działalności logistycznej [2].

### 2.3 Text mining

Informacja najczęściej zapisywana jest w postaci tekstu. Dokumenty tekstowe mogą podlegać eksploracji danych w celu ekstrakcji użytecznej wiedzy. Działalność związana z przetwarzaniem dokumentów tekstowych może być w dużym stopniu zautomatyzowana dzięki eksploracyjnej analizie dokumentów tekstowych określanej jako text mining.

Eksploracyjna analiza tekstów (ang. *text mining*) jest stosunkowo młodą multidyscyplinarną dziedziną. Wywodzi się z data mining, wyszukiwania informacji, ekstrakcji danych, kategoryzacji tekstu, modelowania probabilistycznego, wykorzystująca miedzy innymi metody statystyczne, czy maszynowe uczenie [4].

Text mining definiowany jest jako odkrywanie i wykorzystanie wiedzy zawartej w zbiorze dokumentów. To sposób znajdowania nowej wiedzy pośród olbrzymich zasobów tekstowych. Jest to zbiór metod, koncepcji, oraz algorytmów przetwarzania zasobów tekstowych zaimplementowanych w postaci programów komputerowych prowadzących do zautomatyzowania procesów przetwarzania dokumentów sporządzonych w językach naturalnych.

Text mining umożliwia analizę informacji, które nie mają tradycyjnej struktury arkusza danych i odnajduje w nich fakty ułatwiające podjęcie decyzji. Jedną z istotnych zalet text miningu jest uwzględnienie kontekstu, w którym pojawiają się szukane słowa i wyrażenia. System text miningowy ukierunkowany jest raczej na analizę i przetworzenie dużych zasobów informacyjnych, ma on dostarczać informacje pozyskane w wyniku analizy dokumentów, a nie same dokumenty.

Text mining, określany też jako text data mining, oznacza proces pooszukiwania informacji oraz zależności w nieustrukturyzowanych dokumentach tekstowych. Według Ah-Hwee Tana ponad 80% informacji przechowywanych w przedsiębiorstwach ma formę tekstu[6], dlatego też text mining posiada tak ważny potencjał komercyjny.

Obecnie większość narzędzi text miningowych nie analizuje znaczeń wyrazów czy zdań, tylko próbuje znaleźć pewne reguły i prawidłowości związane z występowaniem określonych ciągów znaków – czyli słów. Proces ten w wielkim skrócie można rozbić na trzy niezależne części: analizę tekstu, budowę macierzy częstości występowania słów

[5] Brichler U.: Bütler M.: *Information economics*, Routledge 2007, s. 32-33.

[6] Tan A.: *Text Mining: The state of the art and the challenges*, Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD, 1999, s. 65-70.

w dokumentach, a następnie użycie klasycznych metod wywodzących się z data miningu – najbardziej znane to klasteryzacja dokumentów oraz ich klasyfikowanie.

Analiza tekstu polega na przetworzeniu dokumentu tekstowego w listę słów, wyrazów, który określany jest workiem słów (ang. *bag of words*). Pomijane są informacje o kolejności słów i związków między nimi, zakłada się, że wystąpienia wyrazów są niezależne od siebie. Pierwszym etapem przetwarzania dokumentu jest jego podział na słowa. Następnie użyta zostaje tzw. stop-lista, która powoduje usunięcie informacji występujących tylko w jednym dokumencie, czy też wyrazów występujących bardzo rzadko bądź bardzo często. Stop-lista jest więc zbiorem słów, których obecność podyktowana jest zasadami gramatycznymi i koniecznością zachowania logicznej formy przekazu. Wśród słów analizowanego testu znajdują się wyrazy, które nie niosą informacji. Są nimi spójniki, rodzajniki, przyimki itp. Mogą one zostać usunięte z tekstu bez szkody dla jakości analizy, a także mogą ją poprawić powodując usunięcie szumu informacyjnego.

W reprezentacji bazującej na liście słów występujących w dokumencie strukturą reprezentującą dokument jest wektor, a jego elementy informują o liczbie wystąpienia poszczególnych słów [11]. Podejście to nosi nazwę modelu przestrzeni wektorowej (ang. *Vector Space Model, VSM*) [7]. VSM jest techniką wyszukiwania informacji, która przekształca dane tekstowe przyrównując je do algebraicznych wektorów przestrzeni. Po wykonaniu transformacji, działania z zakresu algebry liniowej są wykorzystywane do obliczania podobieństw pomiędzy oryginalnymi dokumentami. Każdy unikalny term[7] (słowo) z kolekcji analizowanych dokumentów stanowi oddzielny wymiar VSM i każdy dokument jest reprezentowany przez wektor obejmujący wszystkie te wymiary. Na przykład, jeśli wektor $v$ reprezentuje dokument $j$ w k-wymiarowej przestrzeni $\Omega$, komponent $t$ wektora $v$, gdzie $t \in 1 \ldots k$, przedstawia stopień związku pomiędzy dokumentem $j$ a termem odpowiadającym wymiarowi $t$ w przestrzeni $\Omega$. Ten stopień związku jest lepiej wyrażany jako macierz częstości $A$ (ang. *term freqeancy matrix*) o rozmiarach $t \times d$, gdzie $t$ jest numerem unikalnego termu, a $d$ jest numerem dokumentu [17]. Element $a_{ij}$ macierzy $A$ jest liczbowym przedstawieniem związku pomiędzy termem $i$ a dokumentem $j$, reprezentuje liczbę wystąpień termu $i$ w dokumencie $j$. W macierzy częstości stosuje się też łączenie słów mających bardzo podobne znaczenia, aby zmniejszyć rozmiar macierzy, a tym samym poprawić jakość danych. Kolejnym krokiem jest przekształcenie słowa do jego formy podstawowej, czyli rdzenia. Z przyczyn gramatycznych te same słowa mają różne formy, co w języku polskim jest szczególnie istotne ze względu na bardzo bogatą fleksję. Za znalezienie rdzenia słowa odpowiadają narzędzia lematyzacji i stemmingu. Lematyzacja korzysta ze słownika i analizuje kontekst słów, zwłaszcza zaś formy słów stojących obok. Wynikiem tego działania jest znalezienie form hasłowych. Stemming zaś jest prostym zbiorem reguł, operujących na poziomie liter, którego wynikiem są rdzenie słów. Po dokonaniu dalszych przekształceń prowadzących do redukcji wymiarowości macierzy częstości można przejść do grupowania dokumentów.

Grupowanie dokumentów tekstowych polega na podziale zadanego zbioru dokumentów tekstowych na pewną liczbę grup (klastrów), w których dokumenty charakteryzują się podobną treścią. Podobieństwo najczęściej jest definiowane jako podobieństwo tematyki dokumentów. Pożądane jest aby dokumenty przydzielone do tej samej grupy były do siebie wzajemnie jak najbardziej podobne, zaś dokumenty należące do różnych grup powinny się jak najbardziej różnić. Taksonomiczne metody grupowania należą do klasycznych metod analizy danych. Grupują obiekty, które są charakteryzowane za pomocą z góry przyjętego zbioru atrybutów. Jako dane wejściowe dla dokumentów tekstowych przyjmuje się zwykle macierz odległości, bądź macierz podobieństwa pomiędzy analizowanymi obiektami. Dobór danych wejściowych, ich reprezentacja, grupowanie, ocena jakości grupowania oraz interpretacja wyników stanowią zamkniętą pętlę. Posiadana przez eksperta informacja wraz z intuicją, jest często potrzebna do właściwego określenia niezbędnej kolekcji danych i jej reprezentacji. Na podstawie posiadanego zbioru danych tekstowych określa się możliwość zastosowania danych metod grupowania. Wybór odpowiedniej metody zależy od postawionego celu grupowania oraz specyfiki badanego zjawiska. Odkryte w wyniku grupowania regularności w danych mogą zostać zweryfikowane statystycznie – do tego celu służą odpowiednie mierniki taksonomiczne i testy statystyczne. Tak odkryte regularności należy sformułować w sposób zrozumiały i przyjazny dla człowieka, umożliwiając ich interpretację. Sama interpretacja wyników grupowania pozwala połączyć uzyskane rezultaty z wcześniejszą wiedzą o badanym zjawisku bądź sformułować nowe hipotezy.

Klaster zawierający kilka tysięcy dokumentów może pomóc w ujawnieniu ważnych zagadnień i kluczowych idei związanych z funkcjonowaniem przedsiębiorstwa, a zawartych w zgromadzonych w firmie dokumentach. Klasteryzację dokumentów stosuje się w analizie danych ankietowych, analizie opinii klientów lub zbiorów wiadomości e-mail.

Metody text minigowe zyskały uznanie i przynoszą oczekiwane rezultaty przy:

- wyszukiwaniu informacji – tworzone są inteligentne systemy zadawania zapytań, tak aby „rozumiały" i wykonywały polecenia użytkowników opisane językiem naturalnym,
- grupowaniu dokumentów – możliwość automatycznego budowania grup tematycznych w dużych kolekcjach dokumentów,
- ekstrakcji wyrażeń – automatyczne odnajdywanie w dokumentach słów o znaczeniu kluczowym,
- identyfikacji powiązań – wykrycie związków istniejących pomiędzy informacjami uzyskanymi z dokumentów tekstowych, bądź identyfikacja dokumentów, które są powiązane ze sobą ze względu na treści w nich zawarte,
- przeprowadzaniu rozszerzonego modelowania predykcyjnego polegającego na łączeniu danych pochodzących z różnych źródeł,
- wizualizacji danych,
- automatycznym tłumaczeniu – translacja dokumentu pomiędzy najbardziej „powszechnymi" językami świata,
- automatycznym generowaniu streszczeń – stworzenie konspektu dokumentu podsumowującego jego zawartość.

[7] Pod pojęciem termu rozumie się pojedyncze słowo, lub tez grupę słów, których znaczenie jest całkiem inne niż pojedynczych słów składających się na grupę np. izba, „Wysoka Izba".

## 3. GRUPOWANIE NAZW TOWARÓW

Grupowanie w katalogu towarów jest istotnym ogniwem w trakcie eksploracji bazy danych przedsiębiorstwa. W rzeczywistych środowiskach można spotkać katalogi liczące kilkaset lub wiele tysięcy pozycji. Szczegółowa analiza pojedynczych pozycji jest wielce nieefektywna z punktu widzenia eksploracji danych.

### 3.1 Klasyfikacja katalogu towarów

Katalogi towarów posiadają informacje o klasyfikacji ich poszczególnych pozycji. Dane te są wprowadzane ręcznie zgodnie z przyjętą w przedsiębiorstwie systematyką katalogu towarów. Taka klasyfikacja pociąga za sobą wiele niedoskonałości. Poniżej podane zostało kilka ich przykładów:

- Klasyfikacja odzwierciedla tylko podstawowe cechy towaru.
- Klasyfikacja tworzona jest ręcznie – trudno dokonywać znaczących modyfikacji.
- Wpisy o klasyfikacji towaru są uzupełniane z opóźnieniem.
- Wprowadzanie nowych rodzajów towarów winno skutkować modyfikacją istniejącej systematyki – w większości przypadków jest to proces długotrwały.
- Klasyfikacje tworzy się indywidualnie dla danego przedsiębiorstwa, w oparciu o bieżące potrzeby – realizacja przyszłych wymagań może być wręcz niewykonalna.

Alternatywą pokonującą te i inne trudności może być utworzenie automatycznego systemu klasyfikowania towarów. Proponowaną metodą jest grupowanie wykorzystujące informację zawartą w nazwach towarów. W takim podejściu można wskazać istotne zalety:

- System zapewnia sklasyfikowanie każdej pozycji katalogowej.
- Brak ograniczeń w ilości analizowanych pozycji.
- Istnieje możliwość modyfikacji parametrów klasyfikacji, co przekłada się na rezultaty.
- System opiera się na wiedzy ekspertów z danej dziedziny.
- Istnieje możliwość wprowadzania nowych informacji.

Najważniejsze przesłanki jakie stoją przed metodami grupowania to:

- zredukowanie dużej liczby danych pierwotnych do kilku podstawowych kategorii, które mogą być traktowane jako przedmioty dalszej analizy,
- odkrycie struktury analizowanych danych,
- uzyskanie podczas badania jednorodnych obiektów, ułatwiających wyodrębnienie ich zasadniczych cech,
- zmniejszenie nakładu pracy i czasu analizy mającej na celu klasyfikację obiektów.

Dokonując analizy przydatności metod grupowania do eksploracji danych zwraca się uwagę na:

- sposób określania podobieństwa i niepodobieństwa, wyznaczający zasady łączenia obiektów w klastry,
- sposób ujęcia danych,
- metody poszukiwania rozwiązań
- formę prezentacji danych.

Grupowanie jest bardzo często początkiem w procesie poznawania danego zjawiska na podstawie zbioru obserwacji. W dużej mierze bazuje na metodach algorytmicznych, jednak z powodu słabego wykorzystywania różnych aspektów wiedzy związanej z badanym zbiorem obiektów, ich atrybutów i środowiska, uzyskiwane rezultaty mogą być daleko niezadowalające.

### 3.2 Towar

Towar można zdefiniować jako obiekt podlegający transakcji sprzedaży. W bazach danych występujących w przedsiębiorstwach zarówno rejestracja przebiegów procesów biznesowych jak i faktu dokonania transakcji sprzedaży jest bardzo istotnym elementem. Przedsiębiorstwa stosują różnorodne sposoby rejestracji faktu dokonania sprzedaży. Jednak zawsze dokonywana jest rejestracja sprzedawanego obiektu, czy to towaru, usługi, czy produktu. W artykule uwzględniane są tylko takie bazy danych, w których towary zgromadzone są w postaci katalogu towarów.

Do przeprowadzenia analizy towarem będzie każdy obiekt będący przedmiotem transakcji, a więc nie tylko rzecz, lecz i usługa, licencja itp. Każdy z analizowanych towarów posiada jednoznaczny identyfikator wskazujący na daną pozycję katalogową opisaną przez nazwę i opis produktu. Każdy towar może mieć jeden lub więcej identyfikatorów. Zbiory danych, na które składają się nazwy towarów mają pewne właściwości, które powinny być uwzględnione w procesie grupowania. Należą do nich:

- nazwy składające się z maksymalnie kliku słów,
- duża ilość powtórzeń poszczególnych słów,
- ograniczona liczebność występujących słów,
- możliwość występowania wyrazów i symboli, które występują bądź nie w języku naturalnym,
- szczególne przypadki, w których nazwy różnią się tylko jednym słowem,
- nazwy towarów można traktować jak dokumenty tekstowe.

Część z nich wykorzystywana jest do polepszenia uzyskiwanych wyników, część do zwiększenia efektywności działania algorytmów grupowania.

### 3.3 Zasady grupowania nazw towarów

Przeprowadzając analizę słownego zapisu nazwy danej pozycji towarowej, można dokonać porównania nazw poszczególnych towarów, a co się z tym wiąże ich grupowania. Przyjętym założeniem jest potraktowanie nazw towarów jako bardzo krótkich tekstów, co oznacza, że do ich analizy można zastosować mechanizmy opracowane na potrzeby pozyskiwania informacji z dokumentów. Są jednak pewne znaczące różnice, które muszą zostać uwzględnione. W podejściu bazującym na eksploracyjnej analizie danych tekstowych dokumenty są reprezentowane przez wektor słów:

$$x = x_1, \ldots, x_n \tag{3}$$

gdzie: $n$ - określa liczbę wszystkich słów i termów występujących w kolekcji dokumentów,

$x_i$ - współczynnik zależny od częstości i sposobu występowanie danego słowa.

Uwzględniając te zależności, odrzucane są informacje o wzajemnym ułożeniu słów w dokumentach, pozostają jedynie informacje o częstości ich występowania. Takie uproszczenie w wielu tekstach nie wpływa znacząco na jakość otrzymywanych wyników. Na potrzeby nazw towarów może być również wykorzystane, ponieważ w przeważającej części znaczenie danego termu nie jest zależne od jego położenia w nazwie.

Nie sposób zbadać „idealnego" podobieństwa dwóch dokumentów, ponieważ zależałoby ono od każdego elementu w nich występującego. Dlatego tez dokonując pewnych uproszeń i założeń definiowane są wagi słów występujących w kolekcji dokumentów. Jednym z najczęściej występujących wskaźników jest $TF - IDF$ (ang. *Term Frequency – Inverse Document Frequency*) [10]. Jego miara jest oparta na częstotliwości występowania danego słowa (termu) w dokumencie, oraz liczbie dokumentów, w których dane słowo (term) występuje. Dla każdego słowa obliczana jest jego waga według formuły:

$$TF - IDF = TF(i) * IDF(i) \tag{4}$$

Wartości te definiowane są następująco:

$$TF(i) = \frac{N_i}{\sum_k N_k} \tag{5}$$

$$IDF(i) = \log\left(\frac{|D|}{|d_i \supset t_i|}\right) \tag{6}$$

gdzie: $N_i$ - liczba wystąpień danego wyrazu w dokumencie,

$\sum_k N_k$ - oznacza liczbę wszystkich słów w dokumencie,

$|D|$ - liczba wszystkich dokumentów w zbiorze (kolekcji),

$|d_i \supset t_i|$ - liczba dokumentów (należących do kolekcji), które zawierają wystąpienie słowa $t_i$.

Po uwzględnieniu powyższych wzorów otrzymuje się:

$$TF - IDF = \frac{N_i}{\sum_k N_k} * \log\left(\frac{|D|}{|d_i \supset t_i|}\right). \tag{7}$$

Stosowanie algorytmu TF-IDF eliminuje efekt przeszacowania wartości słowa. Uwzględnia on nie tylko częstość słowa w dokumencie, ale także obecność tego słowa w innych dokumentach. Słowo występujące w niewielkiej liczbie dokumentów ma większe znaczenie, niż słowo występujące w znacznej części kolekcji (zbiorów tekstowych). Jednak podczas badań nad katalogiem towarów dało się zauważyć, że stosowanie wskaźnika $TF - IDF$ prowadziłoby do błędnych wyników. Wynika to z faktu tworzenia nazw towarów w zupełnie odmiennym celu. Katalog towarów zawiera zbiór bardzo podobnych do siebie dokumentów tekstowych, w których sąsiednie pozycje nieznacznie różnią się między sobą. Stąd zupełnie inaczej przedstawiają się rozkłady statystyczne występowania poszczególnych słów.

Dokonując analizy nazwy poszczególnych produktów można dostrzec pewną właściwość – jeżeli nazwa towaru potraktowana zostanie jako ciąg słów to wiele par produktów posiada podobne nazwy (jedno lub więcej słów występuje w obu analizowanych nazwach produktów). Sam człowiek tworząc nazwy dla produktów stara się by były podobne do produktów pokrewnych. Zazwyczaj zmienia lub dodaje jedno lub więcej słów, aby nowa nazwa w sposób wystarczający

identyfikowała nowy produkt. Stąd zasadnym wydaje się, że w pewnym zakresie i z pewną dokładnością badając stopień podobieństwa nazw produktów otrzymuje się współczynniki, które określają podobieństwo samych produktów. Z drugiej strony nie należy pomijać też trendu do maksymalnego skracania poszczególnych nazw, jeśli nie prowadzi to do powstawania niejednoznaczności.

Porównywanie dokumentów tekstowych pod względem znaczeniowym opiera się na jednoczesnym występowaniu niektórych wyrazów w obu dokumentach. Oczekiwane rezultaty są widoczne w przypadku analizy dłuższych tekstów. W przypadku, gdy analizowane dokumenty są bardzo krótkie (nazwy towarów), znaczenia nabiera każdy wyraz ze względu na zastosowany szeroki zasób słów. W przypadku nazw towarów zasadnym wydaje się użycie słownika znaczeniowego, ukazującego relację „objaśnień" od terminów szczegółowych do bardziej ogólnych. W praktyce prowadzi to do dopisania do nazwy produktu szeregu słów, które określają pewne ogólne właściwości danego produktu, wynikające ze znaczeń poszczególnych słów. Następnie dokonuje się porównania dokumentów zawierających rozszerzone nazwy towarów, w myśl zasady, ze im bardziej produkty są do siebie podobne, tym więcej mają cech wspólnych co objawia się większą ilością występowania powtórzeń danych słów. Częstość występowania w dużej mierze zależy od stopnia szczegółowości danego termu - słowa, które opisują bardziej ogólne termy występują częściej, a te szczegółowe rzadziej. Jest to wynik sztucznego zabiegu rozwijania nazw towarów. Dany term pojawi się wszędzie tam, gdzie towar należy do grupy określanej przez ten term, a co za tym idzie stopień ogólności termu jest proporcjonalny do ilości jego wystąpień.

### 3.4 Skuteczność grupowania

Uwzględniając bazę towarów występującą w rzeczywistym przedsiębiorstwie należy wziąć pod uwagę fakt występowania wielu wtrąceń, które niekoniecznie pasują do zasadniczej części bazy. Takie sytuacje występują bardzo często, powodując występowanie szumu informacyjnego. Należy tak dobrać algorytm grupowania, aby „poradził" sobie z tego typu danymi. Aby mówić o skuteczności grupowania należy odpowiednio dobrać funkcję odległości, która powinna odzwierciedlać stopień niepodobieństwa poszczególnych obiektów. Możliwe do zbadania miary odległości przedstawia tabela poniżej (Tab. 1).

*Tab. 1. Miary odległości wraz z ich postacią algebraiczną*

| **Nazwa miary** | **Wzór** | **Dodatkowe oznaczenia** |
|---|---|---|
| Euklidesowa | $d(x,y)=\sqrt{\sum_i (x_i-y_i)^2}$ | |
| Manhattan | $d(x,y)=\sum_{i=1}^{n}\lvert x_i-y_i\rvert$ | |
| Kosinusowa | $d(x,y)=\frac{x\circ y}{\lvert x\rvert\cdot\lvert y\rvert}$ | $x\circ y=\sum_{i=1}^{n}x_i y_i$<br>$\lvert x\rvert\cdot\lvert y\rvert=\sqrt{\sum_{i=1}^{n}x_i^2}\cdot\sqrt{\sum_{i=1}^{n}y_i^2}$ |
| Wskaźnik Jaccarda | $d(x,y)=1-\frac{\sum_{i=1}^{n}x_i y_i}{\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2-\sum_{i=1}^{n}x_i y_i}$ | |

Bibliografia [7, 5, 13, 8]

Miara euklidesowa jest funkcją odległości pomiędzy obiektami w zwykłej wielowymiarowej przestrzeni. Odległość Manhattan często stosowana jest do porównywania obiektów z atrybutami zapisanymi w formacie binarnym. Miara kosinusowa jest funkcją stosowaną do porównywania dokumentów tekstowych [12]. Wskaźnik Jaccarda stosowany jest do porównywania zbiorów. Wszystkie wymienione funkcje można zastosować jako funkcje odległości w algorytmach grupowania, różnią się one charakterystyką i właściwościami.

Odnosząc się do katalogu nazw towarów przy obliczaniu funkcji odległości należy uwzględnić następujące założenia:

- Zbiór słów i termów występujących w nazwach towarów określony jest wzorem:

$$X=\{x_1,\ldots,x_p\} \tag{8}$$

gdzie: $X$ – wektor słów reprezentujący nazwę towaru,

$x_i$ - liczba wystąpień słowa o indeksie *i* w reprezentowanej nazwie.

- Każdą nazwę towaru można odwzorować w wektor $X$ o rozmiarze *p*, gdzie *p* jest liczbą słów występujących we wszystkich nazwach, zaś poszczególne współczynniki odzwierciedlają występowanie odpowiadających im słów.

Wykorzystanie w obliczeniach współczynnika Jaccarda nie przynosi spodziewanych rezultatów. Związane jest to z jednakowym traktowaniem wszystkich słów i nie rozróżnianiem ich pod względem ważności, stąd nie może on wiernie odzwierciedlać relacji zachodzących pomiędzy poszczególnymi towarami. Zasadnym wydaje się użycie wskaźnika Jaccarda uwzględniającego dodatkowo wagi ważności poszczególnych słów, a więc:

$$X = \left\{ x_1, w_{x_1}, \ldots, x_p, w_{x_p} \right\}. \tag{9}$$

Do grupowania nazw w katalogu towarów można użyć co najmniej kilku algorytmów:

- hierarchicznego deglomeracyjnego – jego działanie rozpoczyna się od pojedynczego zbioru, który zawiera wszystkie obiekty i jest poddawany sekwencyjnemu podziałowi; metoda ta nie jest w stanie odwzorować bardziej skomplikowanej struktury podziału,
- genetycznego prostego – niedociągnięciem tej metody jest fakt, iż w większości przypadków podane jest wiele parametrów algorytmu, które musza być określone i nie do końca jest jasne jakie są idealne wartości tych parametrów dla danego problemu,
- genetycznego opartego o listy decyzyjne,
- przeszukiwania heurystycznego opartego o listy decyzyjne.

W przypadku grupowania nazw towarów dobre wyniki dają metody oparte o listy decyzyjne. Metody, które korzystają z wyznaczania miar odległości są bardzo wrażliwe ma stosowaną funkcję, co przekłada się na wyniki grupowania. Algorytm genetyczny prosty nie potrafi odnaleźć optymalnego rozwiązania, zaś użycie algorytmu genetycznego opartego o listy decyzyjne nie powoduje pułapki znalezienia rozwiązania nieoptymalnego. Trudno jednak bez dalszej analizy i badań określić, które wyniki są lepsze, ponieważ nie mogą one zagwarantować dostarczenia zawsze najlepszego rozwiązania. Algorytmy przeszukiwania heurystycznego oparte o listy decyzyjne potwierdziły swoja skuteczność i przydatność w grupowania nazw towarów. Co więcej potencjalnie nadają się one także do analizy i wnioskowania zmian i trendów.

## 4. WNIOSKI

Pomimo tak wielu podobieństw do grupowania dokumentów tekstowych, grupowanie w katalogu towarów jest znacząco różne. Podstawowe różnice to:

- Wymagane jest skorzystanie z wiedzy zewnętrznej – dane zawarte w bazie nie są wystarczające do pogrupowania towarów.
- Wiele z przesłanek wykorzystywanych przy klasycznym grupowaniu dokumentów tekstowych musi być inaczej interpretowana.
- Brak jest optymalnej funkcji oceny jakości podziału.

Istnieją też cechy bardzo podobne:

- Podobieństwo dokumentów tekstowych wyznacza się na podstawie analizy statystycznej słów lub analizy strukturalnej opartej na elementach łączących znaczeniowo dokumenty. Dokonując grupowania katalogu nazw towarów, wykorzystuje się informację strukturalna zawartą w nazwach towarów, ukrytą zaś w znaczeniach poszczególnych słów.
- W procesie porównywania tekstów, w tym także nazw produktów bardzo istotne znaczenie ma przypisanie odpowiednich wag poszczególnym słowom. Jednak sam sposób wyznaczania wag zdecydowanie się różni w tych dwóch przypadkach.

Badania nad text miningowymi metodami eksploracji danych wydają się być bardzo obiecujące, gdyż pozwalają na zaoszczędzenie czasu i pieniędzy, które musiałyby zostać przeznaczone na przeczytanie i ewentualne eksplorowanie przez człowieka ogromnego repozytorium dokumentów tekstowych. Ponadto zaletą danych tekstowych jest to, iż możemy poznać przyczynę wystąpienia danego zjawiska a nie tylko skutek – jak ma to miejsce w przypadku danych ilościowych. Text Mining jest już stosowany w przedsiębiorstwach, w tym także logistycznych, a jego coraz szersze zastosowanie budzi nadzieję na dalszy rozwój.

## 5. BIBLIOGRAFIA

[1] Brichler U.: Bütler M.: *Information economics*, Routledge 2007, s. 32-33.
[2] Coyle J.J, Bardi E.J., Langley C.J. Jr.: *Zarządzanie logistyczne*, Warszawa, PWE 2002.
[3] Gołembska E.: *Kompendium wiedzy o logistyce*, Warszawa, PWN 2010.
[4] Kao A., Poteet S.: *Natural Language Processing and Text Mining*, London, Springer 2007.
[5] Konchady M.: *Text Mining Application Programming*, Boston, Massachusetts, Charles River Media 2006.
[6] Langefors B.: *Information systems theory*, Information Systems, Vol. 2(4), s. 207–219, 1977.
[7] Manning C. D., Raghavan P., Schütze H.: *Introduction to Information Retrieval*, Cambridge, Cambridge University Press 2008.
[8] Markov Z., Larose D. T.: *Eksploracja zasobów internetowych*, Warszawa, PWN 2009.

[9] Niedzielski E.: *Próba systematyzacji procesów rozwoju systemów informatycznych*, Wiadomości Statystyczne, nr 4, 1986, s. 20-23.
[10] Salton G., Buckley C.: *Term weighting approaches in automatic text retrieval*, Information Processing and Management, vol. 24(5), s. 513–523, 1988.
[11] Salton G., McGill M.J.: *Introduction to Modern Information Retrieval*, New York, s. 52-117, McGraw-Hill Book Co. 1983.
[12] Salton G.*: Automatic term class construction using relevance - a summary of word in automatic pseudoclassification*, Information Processing and Management, vol. 16(1), s. 1-15, 1980.
[13] Schenker A., Bunke H., Last M., Kandel A.: *Graph-Theoretic Techniques for Web Content Mining*, World Scientific Publishing Co. Pte. Ltd. 2005.
[14] Sienkiewicz P.: *10 wykładów*, AON, Warszawa 2005, s. 62.
[15] Szymonik A.: *Technologie informatyczne w logistyce*, Warszawa, Placet 2010.
[16] Tan A.: *Text Mining: The state of the art and the challenges*, Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD, 1999, s. 65-70.
[17] Zhang T., Oles F.: *Text categorization based on regularized linear classification methods*, Information Retreival, vol. 4, 2001.