

WYKORZYSTANIE DANYCH OKREŚLONYCH LINGWISTYCZNIE W SYSTEMACH POZYSKIWANIA WIEDZY.

Dariusz Mazur

Wydział organizacji i zarządzania. Politechnika Śląska w Gliwicach
dmazur@polsl.gliwice.pl

1. Streszczenie/Abstract

Artykuł poświęcono problematyce pozyskiwania wiedzy z baz danych, odnosząc się zarówno do danych w postaci numerycznej jak i postaci lingwistycznej. Przedstawiono pojęcie danych lingwistycznych i ich zastosowanie w bazach danych. Omówiono przykładowe techniki pozyskiwania wiedzy oraz przedstawiono pewne modyfikacje modeli przetwarzania danych numerycznych w celu zastosowania ich do danych w postaci lingwistycznej.

Słowa kluczowe w języku referatu: Pozyskiwanie wiedzy, lingwistyka komputerowa, grupowanie.

This paper is data mining. linguistic form or categorical attributes Presented examples of gathering knowledge and some of changes method computing numerical data to use in computing symbolic and linguistic data.

Data mining, linguistic, computing with words, clustering.

2. Wstęp

Pozyskiwanie wiedzy na podstawie danych zgromadzonych w bazach danych, powszechnie znane jako termin angielski *data mining*, jest stosowane w wielu zagadnieniach. Najczęstsze zastosowania systemów pozyskiwania wiedzy spotyka się w sferze biznesu, ale również innych dziedzinach np.: meteorologia, systemy przetwarzania obrazów, systemy przetwarzania dokumentów (np. bazy publikacji naukowych, wyszukiwarki internetowe). Obecna technologia informatyczna pozwala gromadzić olbrzymie zasoby danych, ale ich analiza jest bardzo wolna i kosztowna. *Pozyskiwanie wiedzy* stało się znaczącym obszarem badań naukowych w ostatniej dekadzie XX wieku, czego dowodem są liczne publikacje naukowe w tym zakresie.

Szczególnie wiedza o klientach dla wielu organizacji biznesowych staje się krytyczna dla ich dalszego istnienia. W bazach danych gromadzi się olbrzymie ilości faktów, zdarzeń i innych informacji ale właściwa wiedza pozostaje ukryta i nie uchwycona. Z drugiej strony zaostarzająca się konkurencja wymaga, aby organizacja stale dopasowywała swoją ofertę do indywidualnych preferencji klienta. To wzmacnia rozwój narzędzi pozyskiwania wiedzy, szczególnie w kontekście wspomaganie podejmowania decyzji.

Celem pracy jest przedstawienie zagadnienia pozyskiwania wiedzy ze szczególnym uwzględnieniem przetwarzania danych postaci lingwistycznej. Przedstawione zostaną pewne modyfikacje modeli numerycznych w celu zastosowania ich dla danych lingwistycznych.

Niniejsze opracowanie składa się z 3 części. W pierwszej omówiono podstawowe metody pozyskiwania wiedzy. W drugiej przedstawiono zagadnienie reprezentacji danych w formie lingwistycznej oraz wynikające z tego aspekty związane z pozyskiwaniem wiedzy. W trzeciej przedstawiono przykładowe zagadnienia związane z przetwarzaniem danych w postaci lingwistycznej, wskazano pewne formalne reguły, które są podstawą tworzenia rozwiązań dla przetwarzania danych lingwistycznych.

3. Metody pozyskiwania wiedzy

Wyróżnia się szereg metod stosowanych przy pozyskiwaniu wiedzy. Najbardziej powszechne opierają się na poszukiwaniu skojarzonych reguł, uogólnianiu (streszczaniu, podsumowaniu), klasyfikowaniu oraz grupowaniu. Poszczególne metody nie tworzą wykluczających się nawzajem przepisów na uzyskanie rozwiązania, jakim jest odkrycie wiedzy zgromadzonej w bazie danych. Często zdarza się jednocześnie użytkowanie poszczególnych metod, tworzone są algorytmy, w których zaimplementowane są różne idee, a wyniki uzyskane jedną metodą są przekazywane do kolejnej w celu uzyskania optymalnych rezultatów.

3.1. Schemat procesu pozyskiwania wiedzy

Bazy, będące źródłem dla systemów pozyskiwania wiedzy, zwykle są bardzo obszerne, w zależności od źródła mogą liczyć nawet wiele milionów rekordów. Wielkość ta zależy zarówno od badanego czasookresu jak i rodzaju zastosowania. Jak wiadomo koszt obliczeń silnie zależy od wielkości tej bazy, tym nie mniej spodziewane zyski z odkrycia istniejących zależności są znacznie większe tam, gdzie skala prowadzonych działań jest większa.

Proces pozyskania wiedzy z dużej bazy danych jest zadaniem wykonywanym iteracyjnie i interaktywnie. Dla lepszego zrozumienia tego procesu można podzielić go na pewne etapy. Schemat zaproponowany w [Mann97] składa się z następujących etapów:

1. Zrozumienie i przyswojenie dziedziny.
2. Przygotowanie zbioru danych.
3. Poszukiwanie szablonów, zależności i inne techniki analizy stosowane w pozyskiwaniu wiedzy.
4. Przetwarzanie uzyskanych szablonów (z pkt.3) w formę oczekiwaną przez użytkownika.
5. Zastosowanie rezultatów.

Przedstawiony schemat może być zapętłony, tzn. w razie potrzeby (gdy rezultaty danego etapu na to wskazują) może nastąpić powrót do punktu powyżej i powtórzenia wykorzystując zgromadzoną wiedzę.

3.2. Poszukiwanie szablonów

Poszukiwanie szablonów polega na odkrywaniu pewnych zależności, których występowanie ma stosunkowo wysoką częstotliwość oraz szablon taki jest prostszy niż zwyczajne wyliczenie poszczególnych przypadków ich występowania. Zagadnienie to, zaprezentowane w [AIS93], stanowi użyteczny mechanizm poszukiwania korelacji ukrytej w bazie danych.

Problem poszukiwania reguł polega na poszukiwaniu zależności występujących przy określonych minimach występowania i pewności. W tym celu zostało zaproponowanych szereg algorytmów [AS94, BAG99, BMUT97].

Formalnie można to zapisać następująco: jeżeli $I = \{i_1, \dots, i_m\}$ jest zbiorem m różnych elementów, D jest zbiorem transakcji, gdzie każda transakcja T jest zbiorem elementów takich że $T \subseteq D$.

Szablonem nazywamy implikację: $P(X, Y) = \{X \Rightarrow Y \mid X \subset I, Y \subset I, X \cap Y = \emptyset\}$

Źródłowe zadanie poszukiwania szablonów może być zapisane następująco:

$$PI(D, P) = \{p \in P \mid p \text{ występuje wystarczająco często w } D \text{ i } p \text{ jest interesujące}\}.$$

3.3. Uogólnianie

Uogólnianie danych to pewna klasa narzędzi i metod określanych również jako metody dokonywania podsumowania, streszczenia z bazy danych. W literaturze anglojęzycznej spotyka się następujące terminy związane z tym zagadnieniem: *data generalization*, *summarization tools*, *on-line analytical processing (OLAP)*, *data abstraction*, *charakterization*. Celem stosowania uogólniania jest prezentacja w pewien mniej szczegółowy sposób wyspecyfikowanego zbioru danych zgromadzonych w bazie danych. Może to zostać zrealizowane poprzez stosowanie różnych poziomów abstrakcji, na których dane są przetwarzane, stosowanie różnych przekrojów i kierunków analizy. Przykładowo można

analizować dane elementarne, powstałe podczas zapisu poszczególnych transakcji w bazie transakcyjnej albo dane zagregowane według poszczególnych atrybutów (np. data transakcji, obiekt podlegający transakcji).

3.4. Taksonometria

Klasyfikacja (taksonometria) opiera się o hierarchiczny model decyzyjny, dzięki któremu otrzymujemy podział danych wejściowych na poszczególne, uprzednio zdefiniowane, klasy. Definiowanie klas można dokonać na różne sposoby. Może to wynikać z powszechnie dostępnej wiedzy na przykład:

- klasyfikacja na podstawie atrybutu jakim jest adres stosowane w poczcie klasycznej, przy czym *adres = (kraj, miejscowość, ulica, nr budynku, nr lokalu)*,
- wiedzy, którą można pozyskać od eksperta w danej dziedzinie,
- klasyfikacji otrzymanej w wyniku analizy specjalnie dobranej bazy przykładowej, traktowanej jako zbiór treningowy i na tej podstawie stworzenia odpowiedniego schematu decyzyjnego klasyfikowania obiektów.

3.5. Grupowanie

Grupowanie polega na łączeniu ze sobą poszczególnych obiektów w zbiory, w których znajdują się obiekty podobne do siebie pod pewnym określonym względem, natomiast obiekty niepodobne przydzielane są do różnych zbiorów. Do pewnego stopnia jest to technika podobna do wymienionej poprzednio, różni ją sposób pozyskiwania schematu decyzyjnego podziału (reguł podziału), który jest tworzony automatycznie w trakcie procesu grupowania. Grupowanie jest techniką osadzającą się na maksymie rzymskiej „dziel i rządź”, dzięki której duże zbiory danych dekomponowane są na mniejsze fragmenty, jednak bez utraty pewnych, ustalanych w procesie grupowania cech i właściwości. Mniejsze zbiory danych łatwiej poddają się analizie, koszt przetwarzania się zmniejsza. Algorytmy grupowanie, identyfikujące regiony (klastry), opierają się o **funkcję odległości lub podobieństwa** pomiędzy dwoma obiektami. Technika grupowania jest również stosowana jako element pozostałych metod pozyskiwania wiedzy, przykładowo w [LLL01] zaproponowano algorytm poszukiwania szablonów, w który zostało wbudowane grupowanie w celu polepszenia jego efektywności.

4. Dane określone lingwistycznie

Bazy danych, mogące stać się źródłem dla systemów pozyskiwania wiedzy, mają różnorodną postać. Podstawowe rozróżnienie wynika z podziału na dane o postaci strukturalnej oraz o postaci niestrukturalnej. Dane o postaci strukturalnej obejmują dane przechowywane w ściśle określonej strukturze, charakteryzującą się sformalizowaną metodologią ich pozyskiwania i wprowadzania do bazy. Dzięki wprowadzeniu pewnych ram i ograniczeń otrzymuje się strukturę łatwiejszą do dalszego przetwarzania.

Przykładem danych strukturalnych są dane zawarte w bazach transakcyjnych i operacyjnych przedsiębiorstw, szczególnie przy stosowaniu systemów relacyjnych baz danych. Natomiast dane o postaci niestrukturalnej to pozostała część danych gromadzonych w systemach komputerowych, najczęściej mają one postać tekstu zapisanego w języku naturalnym. Dane o postaci niestrukturalnej można spotkać w poczcie elektronicznej, edytorach tekstowych. Również w bazach danych strukturalnych można znaleźć elementy (np. atrybuty w postaci opisowej), które mają taką niezdefiniowaną postać. Elementy takie zwane są danymi określonymi lingwistycznie [GKCR94] oraz danymi symbolicznymi [Gatn98]. Przetwarzanie informacji zgromadzonej w takiej postaci jest łatwe dla człowieka (forma zapisu jest bliska człowiekowi, gdyż wyrażona w jego języku naturalnym), dla komputera jest to bariera, której przekroczenie jest niesłychanie trudne i udaje się w bardzo ograniczonym zakresie.

4.1. Przykłady danych określonych lingwistycznie

Do każdego obiektu, sytuacji, zjawiska można przypisać pewne wielkości je charakteryzujące- dane. Podstawowa klasyfikacja danych wynika z podziału na dane:

- jakościowe (symboliczne, opisowe),
- ilościowe (numeryczne).

Dane ilościowe pojawiają się tam, gdzie znana jest metoda pomiaru. Dotyczy to przede wszystkim wszelkich wielkości fizycznych jak długość, waga, czas. W wyniku pomiaru otrzymujemy konkretną liczbę, która w ustalony sposób reprezentuje mierzoną wielkość. Zupełnie inne cechy wiążą dane jakościowe. Po pierwsze znacznie poszerzony zostaje zakres stosowania: dotyczy zarówno wielkości mierzalnych, czyli tych związanych z reprezentacją ilościową, jak i niemierzalnych.

Dane jakościowe wynikają przede wszystkim z bezpośredniego postrzegania świata przez człowieka. Dzięki swojej percepcji człowiek zdolny jest postrzegać zarówno wielkości fizyczne, jak i duchowe. Wynika to z możliwości postrzegania odległości, wielkości, czasu, koloru, prędkości, kierunku, siły, zapachu, ilości, podobieństwa, prawdziwości, jakości. Istnieje również wiele innych pojęć, które trudno nazwać wielkościami, ale z pewnością charakteryzują obiekty z nimi związane. Wśród nich można wymienić piękno, brzydotę, przestępstwo, wiara, patriotyzm, kara. Cechą wspólną danych jakościowych jest ich reprezentacja w postaci słów analogicznie do stosowanych w języku naturalnym.

4.2. Przetwarzanie danych lingwistycznych

Jeżeli przytoczymy definicję przetwarzania jako pewne określone manipulacje na liczbach lub symbolach to wynika, że przetwarzanie danych lingwistycznych to manipulacje na słowach w znaczeniu takim jak w języku naturalnym.

Przetwarzanie słów (*ang. computing with words* [Zade99]) jest wzorowane na szczególnej ludzkiej zdolności rozwiązywania zadań bez jakiegokolwiek pomiaru jak również obliczeń. Zdolności te grają kluczowe role w umiejętności rozpoznawania i wnioskowania. Najbardziej znanym przykładem może tu być umiejętność napisania streszczenia z danego tekstu.

Podstawową różnicą pomiędzy ludzką percepcją a pomiarem jest to, że pomiar ma charakter punktowy natomiast to co człowiek postrzega ma charakter rozmyty, nieprecyzyjny. Z drugiej jednak strony określenia słowne mimo swej nieprecyzyjności, w odróżnieniu od danych liczbowych, oddają istotę sprawy. Pojęcie wiek może być określone numerycznie w latach. Można stwierdzić, że ktoś ma 21 lat i będzie to fakt absolutnie prawdziwy, dla każdego i w każdej sytuacji, natomiast bez innych informacji nie można dokonać żadnego wnioskowania z tym związanego. Korzystając ze słów ktoś może powiedzieć, że dana osoba jest młoda. I tu następuje szereg niejasności, wątpliwości a nawet sprzeczności. Słowo „młody” można przypisać do każdej osoby będącej w określonym przedziale wiekowym. Niestety nie ma jednolitej definicji tego przedziału. Co więcej, może się zdarzyć że osoba „młoda” dla jednych, dla innych jest już „stara”. Ale jednocześnie można wysnuć następujące wnioski z określenia danej osoby jako „młodej” przez kogoś innego. Po pierwsze określenie „młody” nie odnosi się wyłącznie do bezwzględnego wieku danej osoby mierzonego jednostką czasu, może również określać doświadczenie, siły witalne, nastawienie do życia. Taka wieloznaczeniowość jest powszechna w języku naturalnym. Po drugie można wnioskować o istniejących relacjach pomiędzy osobą określającą a określaną, można również wysnuć pewne wnioski na temat osoby określającej. Po trzecie i chyba najważniejsze: określenie słowne bardzo często zawiera pewien pierwiastek oceny. Zazwyczaj opisu obiektu (sprawy, itp.) dokonuje się w celu późniejszej oceny, dokonania wnioskowania czy też podjęcia działań. W podanym przykładzie, jeżeli znany jest charakter zadania, do którego dobierane są osoby, to podanie określenia, że dany kandydat jest „młody” implikuje jego nadawanie się bądź nie.

W świecie nauki, odwrotnie niż w życiu codziennym, najczęściej posługujemy się danymi numerycznymi. Istnieją jednak dwa podstawowe wskazania, które mogą zdecydować o użyciu przetwarzania danych lingwistycznych w miejsce numerycznych. Po pierwsze jest to konieczne tam, gdzie dostępna informacja jest zbyt nieprecyzyjna aby ją zapisać w postaci numerycznej. Po drugie tam,

gdzie w wyniku przetwarzania danych lingwistycznych otrzymamy bardziej optymalne rozwiązania, lepiej korespondujące z rzeczywistością, przy niższym koszcie obliczeniowym.

4.3. Zasady grupowania lingwistycznego

Grupowanie danych określonych lingwistycznie bazuje na upodobnieniu do sposobu myślenia przez człowieka, przede wszystkim wykorzystuje informację zawartą w termach lingwistycznych. Na drodze budowy narzędzi wzorujących się na człowieku ważne jest opracowanie właściwej metody grupowania wykorzystującej granulację jako właściwość informacji zapisanej lingwistycznie. W trakcie tworzenia odpowiedniego algorytmu grupowania należy rozwiązać szereg problemów. Pierwszy wynika ze sposobu określenia brzegów przedziałów, które wyznaczają poszczególne grupy. Klasyczne teorie zbiorów oparte są o ostre krawędzie. Algorytmy na tym oparte wprowadzają znaczący błąd w przetwarzaniu spowodowany zupełnie różnym traktowaniem elementów bliskich sobie lecz ułożonych po przeciwnych stronach granicy rozdzielających przedziały. Sposób ten również nie odpowiada intuicyjnemu podejściu do grupowania odpowiedniemu dla ludzkiej percepcji.

Drugi problem grupowania wynika z aspektu częściowej przynależności do zbioru. Teorie klasyczne nie przewidują takiego przypadku. Element może należeć do zbioru bądź nie należeć. Świat realny wymaga jednak bardziej elastycznego podejścia. Typowe są przykłady określania wzrostu człowieka: kogoś określa się mianem „trochę niski”, „prawie wysoki” itd. Idąc dalej można oczekiwać jednoczesnego przynależenia do kilku zbiorów. W oparciu o poprzedni przykład oba cytowane określenia mogą dotyczyć tej samej osoby, a z tego wynika że przynależy ona jednocześnie do zbioru wysokich jak i do zbioru niskich.

Zadanie to może być rozwiązane poprzez jawne wprowadzenie takich funkcji przez użytkownika bądź eksperta lub też poprzez zastosowanie technik grupowania dokonanej na wskazanej bazie danych. Stosowanie funkcji wprowadzanych jawnie ma pewne ograniczenia. Wynikają one z kilku czynników. Aby dobrze skonstruować funkcję przynależności należy dobrze znać charakter opisywanych obiektów. Warunek ten jest spełniony dla obiektów z bezpośredniego otoczenia człowieka, takich, które powszechnie są opisywane i grupowane. Wielkości takie jak wzrost człowieka, temperatura otoczenia przy podaniu dodatkowych parametrów (np. obszar czy środowisko) stosunkowo łatwo poddają się podziałowi na grupy określone lingwistycznie, a reguły tym rządzące są dla większości oczywiste. Analogicznie sprawa się przedstawia dla danych, których charakter znany jest grupie ekspertów. Takie środowisko dysponuje również własnymi terminami lingwistycznymi, dla których skonstruowanie funkcji przynależności jest możliwe. Natomiast w stosunku do pozostałych obiektów sprawa się komplikuje. Oczywiście możliwe jest aby użytkownik przeanalizował bazę danych, odpowiednio pogrupował obiekty i określił funkcje przynależności. Podejście to jest jednak pracochłonne. Można to próbować dokonać automatycznie, korzystając z technik grupowania.

5. Model matematyczny

5.1. Pojęcie grupowania

Formalnie grupowanie można zapisać w następujący sposób:

Dany jest zbiór elementów określony w wielowymiarowej przestrzeni ciągłej

$$x_i \in R^s, i = 1, \dots, n$$

Problem grupowania polega na podzieleniu zbioru X na c podzbiorów, tak aby elementy w danym podzbiorze były do siebie bardziej podobne niż do elementów w innych podzbiorach. W tym celu odnajduje się wektor wzorcowy (odniesienia):

$$v_j \in R^s, j = 1, \dots, c \leq n$$

Jest to zbiór elementów możliwie najlepiej charakteryzujących członków poszczególnych grup.

Formalnie zadanie grupowania można potraktować jako minimalizację błędu kwantyzacji zdefiniowanego następująco:

$$E_{rec} = \frac{1}{n} \sum_{k=1}^c \sum_{i=1}^n w_{ik} \|x_i - v_k\|^2, w_{ik} \in [0,1]$$

gdzie

- $\|\cdot\|$ - oznacza identyfikator odległości,
- n - liczbę elementów podlegających grupowaniu
- c - liczbę grup, do których zostaną przydzielone elementy
- w_{ik} - współczynnik przynależności elementu i do wzorca k

5.2. Definicja grupy w grupowaniu elementów określonych atrybutami opisowymi

Intuicyjnie grupa (ang. *cluster*) jest zbiorem numerycznych atrybutów identyfikujący „skupiony region” w przestrzeni atrybutów. Taki region zawiera znacząco więcej elementów określonego rodzaju niż pozostała przestrzeń. Można to przedstawić, korzystając z analogii do wielowymiarowej przestrzeni metrycznej, jako hiper-sześcian. Przykładowo w przestrzeni trójwymiarowej można zdefiniować pewien region następująco $[1,2] \times [2,4] \times [3,5]$ co będzie korespondowało z obszarem o postaci sześcianu, którego brzegi określone są danymi atrybutami numerycznymi. Uogólniając, klasę hiper-sześcianów można wyrazić jako iloczyn kartezjański przedziałów określonych dla poszczególnych wymiarów. Niestety w przestrzeni atrybutów określonych opisowo nie istnieje relacja porządku, więc pojęcie przedziału nie ma sensu. Ale można zastosować bezpośrednie uogólnienie pojęcia przedziału do przestrzeni atrybutów opisowych poprzez zastosowanie zbiorów wartości tych atrybutów. Korzystając z tej analogii, można przyjąć że istnieją hiper-sześcienne regiony określone zbiorami wartości poszczególnych atrybutów.

Formalny zapis jest następujący:

Niech A_1, \dots, A_s jest zbiorem atrybutów opisowych z dziedzinami odpowiednio D_1, \dots, D_s .

Niech zbiór danych jest zbiorem rekordów, gdzie każdy rekord $t: t \in D_1 \times \dots \times D_s$ opisany atrybutami $t = \langle t.A_1, \dots, t.A_s \rangle \in D^s$

Nazywamy $S = S_1 \times \dots \times S_s$ grupą jeżeli $\forall i \in \{1, \dots, s\}, S_i \subseteq D_i$

Rekord t należy do grupy $S: t \in S \Leftrightarrow \forall i \in \{1, \dots, s\}, t.A_i \in S_i$.

5.3. Miara odległości dla danych postaci lingwistycznej

Większość algorytmów grupowania bazuje na miarze odległości. Odległością nazywamy funkcję przekształcającą dwa dane obiekty w $R: D: V \times V \rightarrow \{r \in R^1: r \geq 0\}$. Dla części danych numerycznych można stosować miary odległości w przestrzeni metrycznej, np. euklidesowa miara odległości $d^2(x, y) = (x - y)^T (x - y)$, $\forall x, y \in V$. Dla danych postaci lingwistycznej formuła ta się nie sprawdza. W tym wypadku należy się odwołać do istoty informacji ukrytej w słowach (termach) lingwistycznych. Na przykład jeżeli dany obiekt zawiera adres administracyjny, to można skonstruować algorytm, korzystający z odpowiedniego słownika, który obliczy odległość geograficzną. Przykładowo dla bazy z trzema transakcjami $I = \{i_1 = \text{'KATOWICE'}, i_2 = \text{'KRAKÓW'}, i_3 = \text{'WARSZAWA'}\}$ można przyjąć $d^l(i_1, i_2) = 70\text{km}$, $d^l(i_1, i_3) = 300\text{km}$ (d^l jako oznaczenie funkcji liczącej odległość geograficzną pomiędzy miejscowościami).

Obok funkcji odległości pomiędzy obiektami do grupowania można również zastosować funkcje określające podobieństwo, które w prosty sposób można zaadaptować do algorytmów grupowania. Przykładem funkcji podobieństwa jest współczynnik Jaccarda [JD88] opierający się na następującej zależności:

$$h^J(x, y) = \frac{|x \cap y|}{|x \cup y|}, \quad \forall x, y \in V,$$

Aby jednak właściwie obliczyć współczynnik Jaccarda należy odpowiednio przekształcić badane obiekty tak, aby móc zdefiniować funkcje sumy i iloczynu. Przykładowo podczas porównywania

tekstów w języku naturalnym dokonuje się przekształcenia tekstu na postać zbioru, którego elementami są poszczególne słowa. Dalsze przetwarzanie jest już trywialne, natomiast u podstawy takiego traktowania tekstu leży dogłębna analiza danych o takiej postaci. Zakłada się, że podobieństwo dwóch tekstów w języku naturalnym głównie zależy od jednoczesnego występowania pewnych słów w obu tekstach. Takie założenie dość dobrze sprawdza się w przypadku przetwarzania dokumentów typu publikacje naukowe, prasowe, katalogowanie stron internetowych. Jednak w wielu zastosowaniach stosuje się inne, bardziej zaawansowane funkcje obliczania podobieństwa uwzględniające takie czynniki jak fleksja, wieloznaczność, semantyka i syntaktyka języka.

6. Zastosowanie zbiorów rozmytych do przetwarzania danych lingwistycznych

Założenia przetwarzania danych lingwistycznych są znane od dawna. Jednak przekształcenie się w rzeczywistą metodologię nastąpiło dopiero w wyniku zastosowania logiki rozmytej (fuzzy logic) [Zade77]. Kluczowe stało się zastosowanie zmiennych rozmytych do reprezentacji słów języka naturalnego. Należy sobie odpowiedzieć dlaczego teoria zbiorów rozmytych tak dobrze pasuje do przetwarzania danych lingwistycznych, czy też języka naturalnego.

Aby właściwie wytłumaczyć to zagadnienie należy wyjaśnić sposób w jaki człowiek postrzega świat, swoje otoczenie i siebie. Człowiek posiada szereg zmysłów, dzięki którym otrzymuje sygnały z otoczenia, dzięki nim może postrzegać takie wielkości jak odległość, kolor, prędkość itd. Znana jest również dokładność, a właściwie niedokładność tych zmysłów, wszelkie otrzymanywane wyniki są przybliżone, sygnały które są ich nośnikami są podatne na wiele zakłóceń. I właśnie na takich danych, danych rozmytych, mózg ludzki dokonuje przetwarzania.

Język naturalny, którego zadaniem jest opisywanie świata, takim jak go człowiek postrzega, ma odcisnięte piętno możliwości ludzkich zmysłów. Jest tak samo niedokładny, nieprecyzyjny, poszczególne słowa posiadają wiele znaczeń jak również znaczenia poszczególnych słów pokrywają się. Podczas opisywania danego obiektu człowiek często posługuje się słowami określającymi pewne właściwości obiektu. Obiekty ze względu na daną właściwość mogą być nierozróżnialne, podobne. Różne wartości jakie przybierają dane właściwości są podstawą do grupowania obiektów. Grupy wyznaczane są poprzez wskazanie przedziałów wartości, które stanowią o przynależności do danej grupy. W zależności od przyjętej metody przedziały te mogą mieć punktowo wyznaczone granice, bądź nieostre, rozmyte, tak że przedziały nakładają się na siebie, a dany obiekt może należeć do wielu grup na raz. Zalety stosowania zbiorów rozmytych:

1. Podobieństwo logiki zbiorów rozmytych do stosowanej w języku naturalnym.
2. Możliwość przetwarzania wiedzy niepełnej, niepewnej.
3. Uniwersalne metodologia wnioskowania i przetwarzania.

7. Wnioski

Obecnie dostępna technologia baz danych pozwala na gromadzenie olbrzymich ilości danych. W znacznej części są to dane w postaci lingwistycznej, gdyż używanie takiej postaci jest naturalne dla człowieka. Język naturalny, kształtowany przez tysiące lat stanowi dobre narzędzie opisu otaczającego świata i dostosowany jest do możliwości percepcji człowieka. Natomiast dla komputera stanowi barierę, której pokonanie wymaga stosowania zaawansowanych modeli przetwarzania. Podczas poszukiwania rozwiązań należy się wzorować na samym człowieku i jego umyśle oraz doszukiwać się pewnych analogii z przetwarzaniem danych numerycznych. Istotnym aspektem związanym z pozyskiwaniem wiedzy jest etap określany mianem *rozumienie i przyswojenie dziedziny* która podlega analizie. Zastosowanie metod analizy danych określonych lingwistycznie jest użyteczne w eksploracji baz danych; istotnym zagadnieniem jest zastosowanie odpowiedniej metodologii przetwarzania.

8. Literatura

- [AIS93] Agrawal R., Imielinski T., Swami A.: Mining Association Rules between Sets of Items in Large Databases. Proceedings of the ACM SIGMOD,1993
- [AS94] Agrawal R., Srikant R.: Fast Algorithms for Mining Association Rules in Large Databases, International Conference of Very Large Data Bases, 1994
- [BAG99] Bayardo R.J., Agrawal R., Gunopulos D.: Constraint-Based Rule Mining in Large, Dense Databases, International Conference on data Engineering,1999
- [BEZ81] Bezdek J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press NY 1981.
- [BMUT97] Brin S., Motwani R., Ullman J., Tsur S.: Dynamic Item set Counting and Implication Rules for Market Basket Data, ACM SIGMOD 1997
- [CHY99] Chen M.-S., Han J., Yu S.P.: Data Mining: An Overview from Database Perspective, 1999.
- [CCZL01] Chen N., Chen A. Zhou L., Lu L.: A graph-based clustering algorithm in large transaction databases. IOS Press 2001.
- [Gatn98] Gatnar E.: Symboliczne metody klasyfikacji danych, PWN 1998.
- [GKCR94] Groenemans R., Kerre E., Cooman G., Ranst E.: The Use of Linguistic Terms in Database Models
- [JaDu88] Jain A.K., Dubes R.C.: Algorithms for Clustering data, Prentice Hall,1988.
- [Kacp01] Kacprzyk J.: Wieloetapowe sterowanie rozmyte, WNT 2001.
- [LLL01] Liu F., Lu Z., Lu S.: Mining association rules using clustering, IOS Press 2001.
- [Mann97] Mannila H.: Methods and problems in data mining. International Conference on Database Theory,1997
- [SSTW01] Shaw M.J., Subramaniam C., Tan G.W., Welge M.E.: Knowledge management and data mining for marketing, Decision Support Systems, 2001
- [Zade77] Zadeh, L.A.: Fuzzy sets and their applications to classification and clustering. Academic Press,1977
- [Zade99] Zadeh L.A.: From Computing with Numbers to Computing with Words- From Manipulation of Measurements to Manipulations of Perception. IEEE,1999