Materiały, przykłady i ćwiczenia do przedmiotu

Komputerowe Systemy Rozpoznawania[1]

Adam Niewiadomski

21 września 2009

[1]Te materiały nie stanowią kompletnego opracowania, w szczególności **nie są skryptem** zastępującym wykład. Należy traktować je **wyłącznie** jako **zbiór notatek** ułatwiających słuchaczom udział w zajęciach.

Spis treści

Rozdział 1

Wprowadzenie

Sztuczna Inteligencja (SI, ang. *Artificial Intelligence*) – nauka informacyjna, której celem jest maszynowe (zwłaszcza komputerowe) modelowanie zachowań człowieka w sytuacjach obliczeń, decyzji, sterowania, wyboru, itp., także opis i budowa systemów, działających i "myślących" jak człowiek

Określenie SI pojawiło sie formalnie w roku 1956. Nauka ta łączy w sobie (oprócz informatyki) elementy matematyki, fizyki, psychologii, filozofii, biologii i in.

Test Turinga (1950) – zaprogramować maszynę tak, aby użytkownik myślał, że komunikuje się (słownie lub tekstowo) z żywym człowiekiem (ma te same kompetencje językowe)

Test Turinga o rozszerzeniu 1-go stopnia (1956) – jw. + kontakt wizualny, słuchowy, dotykowy, itp. z człowiekiem

Test Turinga o rozszerzeniu 2-go stopnia (1975) – jw. + kontakt wizualny, słuchowy, dotykowy z dowolnym stworzeniem ożywionym.

Sztuczna Inteligencja – dziedzina nauki zajmująca sie rozwiązywaniem zadań efektywnie niealgorytmizowalnych w oparciu o modelowanie wiedzy (M. Kasperski).

1.1 O obszarach zainteresowania SI

- Systemy ekspertowe (SE) – systemy wspomagające podejmowanie decyzji (diagnozy medyczne, analizy ekonomiczne, sterowanie, optymalizacja, grupowanie, klasyfikacja)

bazy wiedzy (knowledge bases)

bazy reguł (rule base)

automatyczne wnioskowanie, klasyczne i nieklasyczne (automated reasoning, classic and non-classic)

automatyczne dowodzenie twierdzeń, proovery, checkery (automated theorem prooving)

klasyfikacja – rozpoznawanie obiektów lub ich modeli

grupowanie (clustering) – decydowanie na ile grup podzielić dany zbiór obiektów

- Algorytmy inteligentne – rozwiązują zadania, które są zbyt kosztowne obliczeniowo dla metod klasycznych (numerycznych, statystycznych)

  sieci neuronowe (neural networks)

  zbiory przybliżone (rough sets)

  algorytmy genetyczne, obliczenia ewolucyjne (genetic algorithms, evolutionary computing)

  zbiory rozmyte i logika rozmyta (fuzzy sets, fuzzy logic)

  sztuczne życie (artificial life)

  algorytmy i systemy mrówkowe (ant algorithms and systems)

  Algorytmy inteligentne mają szerokie zastosowanie w grafice komputerowej, zwłaszcza w grach komputerowych (rozpoznawanie zachowania przeciwnika), modelowaniu wirtualnej rzeczywistości, przetwarzaniu dużych obrazów, kodowaniu i dekodowaniu animacji i multimediów (ang. *codec*)

- Przetwarzanie informacji (zorientowane na komunikację człowiek-komputer prowadzoną w języku naturalnym)

  Semantic Web

  Question-Answering systems (Q/A systems)

  Flexible querying

  Intelligent search engines

  Computing with words

  Ontologies

  Linguistic summaries of databases (lingwistyczne podsumowania baz danych)

  Data retranslation (retranslacja danych)

1.2 Rozpoznawanie danych

Rozpoznawanie – *recognition*, kojarzone z rozpoznawaniem obrazów, *image recognition*.

Rozpoznawanie – *identification*, identyfikacja, np. w systemach autoryzujących, dostęp do danych, do pomieszczeń, rozpoznawanie "swój–obcy", na podstawie hasła, PINu, cech biometrycznych (np. siatkówki oka, kształtu twarzy, kształtu małżowiny usznej, głosu, itp.)

Rozpoznawanie – *matching*, dopasowywanie wzorców, zwłaszcza tekstowych, czyli *template matching*, ścisłe – *crisp*, przybliżone, zgrubne – *rough*, rozmyte – *fuzzy*, wyszukiwanie informacji na podstawie jej fragmentów, np. *search engines*, funkcja *find* w edytorach, określanie podobieństwa (nierozróżnialności) pomiędzy rekordami/obiektami, np. w celu ich zastąpienia lub zamiany.

Rozpoznawanie – *detection*, rozpoznawanie konturów, wykrywanie krawędzi w obrazach komputerowych, *edge detection*, związek z *image recognition*.

Rozpoznawanie* – badania statystyczne, ilościowy opis cech i trendów wystepujących w pewnycm zbiorze $Z' \subseteq Z$, po to aby aproksymować te opisy dla całego Z, np. sondaże przedwyborcze nie obejmują wszystkich wyborców Z, a jedynie ich podzbiór Z'.

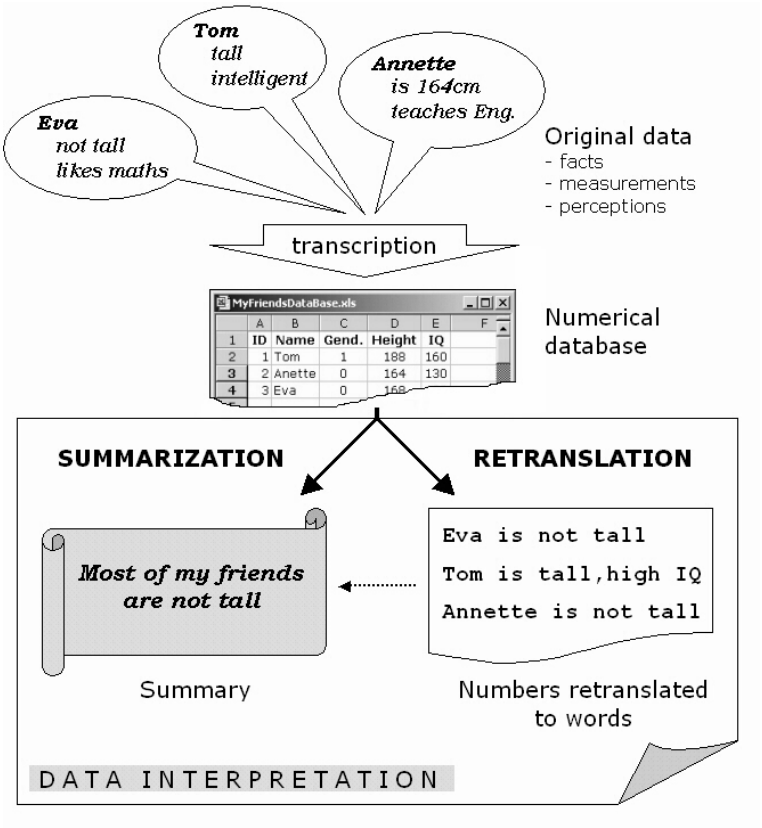
Rozpoznawanie – *interpretation*, rozpoznawanie danych, *data interpretation*. Dostęp do danych w internecie nie stanowi problemu, ale problemem jest odpowiednio szybkie i efektywne określenie przydatności tych danych, ich czytelna prezentacja, filtrowanie, określenie kontekstu, zależności od innych czynników, kolokwialnie "wytłumaczenie o co tu chodzi".

W dalszej części skupiamy się na interpretowaniu danych numerycznych i ich alternatywnym, głównie lingwistycznym, reprezentowaniu.

1.3 Inteligentna interpretacja danych

Duża ilość danych znajdujących sie w Internecie i ogólnie w źródłach elektronicznych wymaga nie tylko wyszukiwania, ale także metod ich filtrowania, selekcji, prezentacji i interpretacji.

Suche dane numeryczne są tak samo bezużyteczne jak duża ilość tekstu



”wyrzucana” przez przeglądarki.

Człowiek, aby ogarnąć ilość danych potrzebną do funkcjonowania, zwłaszcza w życiu zawodowym, najchętniej przyjmuje je z zewnątrz i wyraża w języku naturalnym. W komputerach — niestety — te dane zapisane sa w numerycznej postaci (głównie ze względu na prostotę takiego zapisu w stosunku do tekstu).

Korzystanie z dużej ilości danych pochłania percepcję człowieka w bardzo dużym stopniu, co często uniemożliwia lub znacznie spowalnia codzienne działanie.

Zadaniem *inteligentnej interpretacji danych* jest przetworzenie informacji numerycznej tak aby dostarczyć ją użytkownikowi w postaci dla niego naturalnej, czyli najczęściej: lingwistycznej.

Niektóre opracowania zaliczają do interpretacji także wnioskowanie. Jednak w tym ujęciu interpretacja polega na modelowaniu (reprezentacji) danych oraz ocenie jakości tych modeli.

Przykładowe podsumowanie zbioru danych:
Ceny paliw wzrosły o 0.73% od 23 stycznia br.

Siedem europejskich walut jest słabszych wobec dolara USA niż dwa dni temu
Niewielu pracowników poniżej 25-go roku życia dużo zarabia

Do podsumowywania danych służą metody statystyczne lub metody logiki rozmytej.

Statystyka – opisuje dane używając precyzyjnych wskaźników (średnia, mediana, centyle, odchylenie standardowe, itp.), co jest zaletą. Są one nie zawsze zrozumiałe w potocznym języku, np. w mass-mediach, zwykłej rozmowie (wada).

Logika rozmyta – opisuje dane lingwistycznie, używając potocznych ale zrozumiałych określeń, np. mało, dużo, wysoki, szybko, ok. tysiąca, między 25 a 30, itp. (zaleta). Są one "pojemne" i zrozumiałe, ale nie zawsze dokładne (wada).

Rozdział 2

Zbiory klasyczne i zbiory rozmyte

2.1 Niedoskonałość logiki dwuwartościowej

Funkcja charakterystyczna klasycznego zbioru A'

$$\xi_{A'}(x) = \begin{cases} 1, & \text{if } x \in A' \ \ \forall x \in \mathcal{X} \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

Dowolny element x należy do zbioru lub nie należy i nie ma innej możliwości, czyli *tertium non datur*, czyli prawo wyłączonego środka, czyli Law of Excluded Middle, czyli LEM.

Ćwiczenie 2.1 *Narysuj i/lub określ funkcje charakterystyczne zbiorów:*

1. $x \in [0, 10]$ *w* $\mathcal{X} = \mathbb{R}$
2. x *jest liczbą parzystą w* $\mathbb{N}$
3. $x \in \mathbb{N}$ *w* $\mathbb{R}$
4. $x < 5$ *w* $\mathbb{N}$
5. $x < 5$ *w* $\mathbb{R}$
6. $x < y$ *w* $[0, 10] \times [0, 10]$

"Paradoks łysego" Ktoś, kto ma 60 000 nie jest łysy.
Ktoś, kto ma o jeden włos mniej od osoby niełysej, nie jest łysy.

Zatem ktoś, kto ma 59 999 włosów nie jest łysy.

Powyższe rozumowanie (oparte na schemacie *modus ponens*) powtórzone 60 tys. razy prowadzi do wniosku:
Ktoś kto ma 0 włosów nie jest łysy !!!

Jest to dowód na to, że klasyczna logika i teoria zbiorów nie opisują dostatecznie dokładnie rzeczywistego świata (np. nie uwzględniają stanów pośrednich pomiędzy pojęciami skrajnymi). Przykład klasyczny: *Jutro będzie bitwa morska* – prawda czy fałsz? (Arystoteles, *Logika*).

Ćwiczenie 2.2 *Podaj inne przykłady zjawisk, określeń lub sytuacji, w których określenie przynależności do zbioru nie jest oczywiste, np. "szybka jazda".*

1918 – Jan Łukasiewicz wprowadził trzecią wartość logiczną "$\frac{1}{2}$". Jest wiele interpretacji (semantyk) tych wartości logicznych, ale *prawdopodobieństwo* do nich nie należy, choć często bywa mylone.
Powstały zbiory i logiki cztero-, pięcio-, n-wartościowe i ∞-wartościowe.

2.2 Zbiory rozmyte – podstawowe definicje i przykłady

Definicja 2.3 *Zbiór rozmyty A w niepustej przestrzeni $\mathcal{X}$ (L. A. Zadeh, 1965)*

$$A = \{\langle x, \mu_A(x)\rangle \colon x \in \mathcal{X}\} \tag{2.2}$$

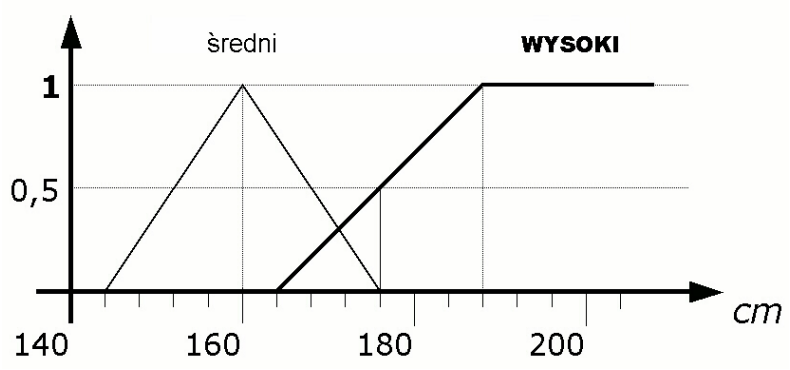
gdzie $\mu_A \colon \rightarrow [0,1]$ jest funkcją przynależności, membership function*, której istotą jest uogólnienie pojęcia funkcji charakterystycznej zbioru [?].*

Dzięki tej konstrukcji, oprócz pełnego należenia x do A, czyli $\mu_A(x) = 1$, lub nienależenia, $\mu_A(x) = 0$, możliwa jest reprezentacja *przynależności częściowej* x do A, gdzie wartość $\mu_A(x) \colon 0 < \mu_A(x) < 1$, interpretowana jest jako *stopień przynależności*, *membership level*.

Mówi się o trzech zasadniczych interpretacjach stopnia przynależności do zbioru rozmytego:

1. *Stopnia preferencji*
2. *Możliwości przynależenia*

oraz w przypadku relacji rozmytych



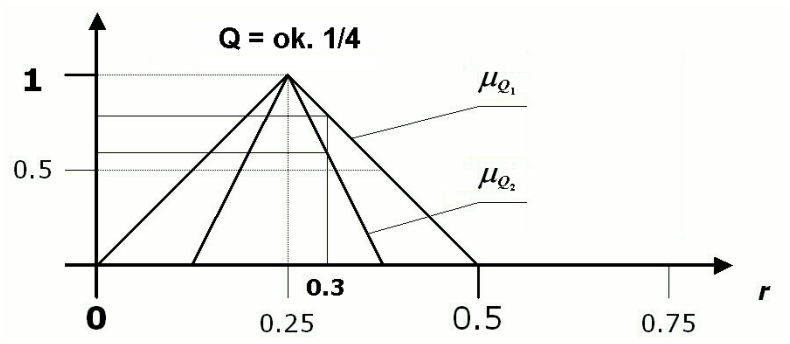
Rysunek 2.1: Funkcja przynależności zbioru rozmytego "wysoki" (człowiek)

3. *siły powiązania, stopnia podobieństwa*

W dalszych rozważaniach posługiwać będziemy się głównie interpretacją pierwszą. W rozdziale 3 o dopasowywaniu wzorców obowiązywać będzie interpretacja trzecia.

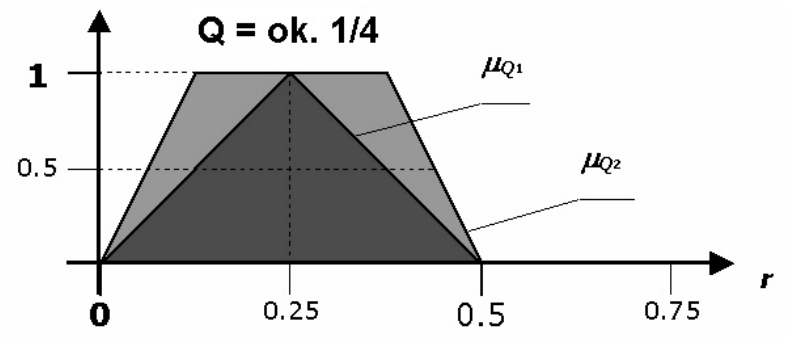
Jak widać w równaniu (2.2), zbiór rozmyty **w sensie składniowym, czyli syntaktycznym** jest szczególnym przypadkiem zbioru ostrego, gdyż jest zbiorem par uporządkowanych postaci $\langle x, \mu_A(x)\rangle$. Jednakże w sensie **znaczeniowym, czyli semantycznym** zbiory rozmyte uogólniają zbiory zwykłe, gdyż pozwalają na opisanie innych przynależności niż tylko 0 i 1.

Przykładowy zbiór rozmyty w przestrzeni skończonej pokazany jest w przykładzie 2.4. Przykładowy zbiór rozmyty w przestrzeni gęstej, ilustrujący termin "wysoki człowiek" pokazany jest na rys. 2.1.



Rysunek 2.2: Funkcje przynależności dla zbioru *około 1/4* o różnych nośnikach

Przykład 2.4 *Niech* $\mathcal{X} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.



Rysunek 2.3: Funkcje przynależności dla zbioru *około 1/4* o różnych liczbach kardynalnych

Zbiór rozmyty A w $\mathcal{X}$ — mała cyfra nieparzysta

$$\begin{aligned} A \quad = \{ \quad & \langle 0, 0.0\rangle, \langle 1, 1.0\rangle, \langle 2, 0.0\rangle, \langle 3, 1.0\rangle, \langle 4, 0.0\rangle, \\ & \langle 5, 0.7\rangle, \langle 6, 0.0\rangle, \langle 7, 0.3\rangle, \langle 8, 0.0\rangle, \langle 9, 0.0\rangle\} \end{aligned} \tag{2.3}$$

Jest to zbiór rozmyty w dyskretnej przestrzeni rozważań. Skrócona forma zapisu

$$A = \{\langle 1, 1.0\rangle, \langle 3, 1.0\rangle, \langle 5, 0.7\rangle, \langle 7, 0.3\rangle\} \tag{2.4}$$

lub

$$A = \{1.0/1 + 1.0/3 + 0.7/5 + 0.3/7\} \tag{2.5}$$

lub

$$A = \left\{ \frac{1.0}{1} + \frac{1.0}{3} + \frac{0.7}{5} + \frac{0.3}{7} \right\} \tag{2.6}$$

gdzie '+' - oznacza sumę mnogościową (nie: algebraiczną) elemntów, a '/' lub '—' – przynależność i element (nie: dzielenie).

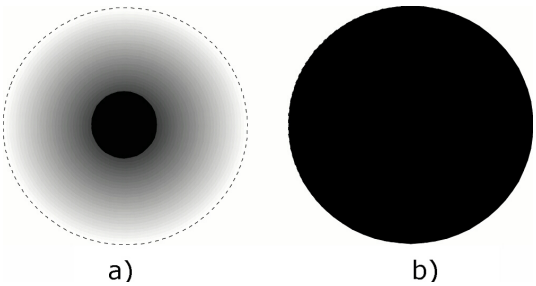
Ogólnie w przestrzeni skończonej definiujemy A jako n-elementową sumę mnogościową

$$A = \left\{ \frac{\mu_A(x_1)}{x_1} + \frac{\mu_A(x_2)}{x_2} + ... + \frac{\mu_A(x_n)}{x_n} \right\} \tag{2.7}$$

Przykład 2.5 *Niech $\mathcal{X}$ będzie kołem o promieniu r. Zbiór rozmyty B "blisko środka" ma funkcję przynależności daną jako*

$$\mu_B(x) = \min \left\{ 1, \frac{4/3 \cdot (1 - d(x))}{r} \right\} \tag{2.8}$$

Ćwiczenie 2.6 *Określ zbiór "osób o jasnych włosach" w swojej grupie.*



Rysunek 2.4: Zbiór B z przykł. 2.5 i jego nośnik

Ćwiczenie 2.7 *Zaproponuj funkcję przynależności do zbioru rozmytego "kolor bliski czerni" w* $\mathcal{X} = \{0, \ldots, 255\}$.

Ćwiczenie 2.8 *Podaj przykłady zbiorów rozmytych w niepoliczalnych przestrzeniach rozważań w* $\mathbb{R}$*:*

1. *"duża prędkość w obszarze zabudowanym", "duża prędkość poza obszarem zabudowanym", "duża prędkość na autostradzie".*
2. *"ciężki człowiek", "ciężki samochód osobowy", "ciężki statek"*

Uwaga o notacji zbiorów rozmytych w przestrzeniach niepoliczalnych

Dla $\mathcal{X}$ gęstej A zapisujemy

$$A = \int_{x \in X} \frac{\mu_A(x)}{x} = \int_{\mathcal{X}} \mu_A(x)/x \tag{2.9}$$

gdzie $\int_{\mathcal{X}}$ oznacza sumę mnogościową po całej przestrzeni $\mathcal{X}$ (nie jest to całka).

Ćwiczenie 2.9 *Zaproponuj postać funkcji przynależności dla zbioru "ciężki człowiek" wg notacji (2.9).*

Niektóre rozszerzenia teorii zbiorów rozmytych

- Intiuitionistic Fuzzy Sets, intuicjonistyczne zbiory rozmyte
- Interval-Valued Fuzzy Sets, interwałowe (przedziałowe) zbiory rozmyte
- Type-2 Fuzzy Sets, zbiory rozmyte typu 2 ("typu dwa")
- L-Intuitionistic Fuzzy Sets (LIFS)
- Interval-Valued Intuitionistic Fuzzy Sets,

- Interval-Valued $M(L)$-Fuzzy Sets,
- Rough Fuzzy Sets,
- Fuzzy Rough Sets,
- L-Fuzzy Rough Sets,
- Conceptual Fuzzy Sets,
- Φ-Flou Sets.

Rozdział 3

Dopasowywanie wzorców

Relacja binarna klasyczna (ostra) na iloczynie kartezjańskim $\mathcal{X} \times \mathcal{Y}$ to zbiór par uporządkowanych postaci $\langle x, y \rangle \colon x \in \mathcal{X}, y \in \mathcal{Y}$. Relacja jest obrazem pewnej zależności pomiędzy obiektami, dokładniej, relacja binarna opisuje zależności pomiędzy dwoma obiektami.

Przykład 3.1 *Niech $\mathcal{X} = \{0, 1, 2, 3\}$. Relacja R' na $\mathcal{X}^2$ obrazuje pojęcie "x_1 jest większe od x_2". Zatem*

$$R' = \{\langle x_1, x_2 \rangle \colon x_1 > x_2\} = \{\langle 1, 0 \rangle, \langle 2, 0 \rangle, \langle 2, 1 \rangle, \langle 3, 0 \rangle, \langle 3, 1 \rangle, \langle 3, 2 \rangle\} \tag{3.1}$$

A zatem, np. $\langle 2, 0 \rangle \in R'$, ale $\langle 2, 3 \rangle \notin R'$. Używając funkcji charakterystycznej relacji R', mamy $\xi_{R'}(2, 0) = 1$, ale $\xi_{R'}(2, 3) = 0$.

R' jest relacją ostrą – modelowana własność zachodzi dla pary obiektów lub nie zachodzi, i nie ma innej możliwości; x_1 jest większe od x_2 albo nie jest.

3.1 Relacje rozmyte: własności, zastosowania

3.1.1 Definicja i przykłady relacji rozmytej

Aby zobrazować pomiędzy obiektami powiązania występujące z różnym stopniem nasilenia, definiuje się *relację rozmytą*.

Definicja 3.2 *Relacja rozmyta R na iloczynie kartezjańskim przestrzeni rozważań $\mathcal{X} \times \mathcal{Y}$*

$$R =_{df} \{\big\langle \langle x, y \rangle, \mu_R(x, y) \big\rangle \colon x \in \mathcal{X}, y \in \mathcal{Y}\} \tag{3.2}$$

gdzie $\mu_R \colon \mathcal{X} \times \mathcal{Y} \to [0, 1]$ – funkcja przynależności do relacji rozmytej.

Wartość funkcji μ_R może być różnie interpretowana, ale najczęściej mówi się o szeroko rozumianym *stopniu powiązania* elementów x i y. Typowym przykładem jest podobieństwo x i y – jak silne jest to podobieństwo? 1 – identyczność, 0 – brak podobieństwa, wartości pośrednie – pewne podobieństwo, podobieństwo "w pewnym stopniu".

Przykład 3.3 *Niech $\mathcal{X} = \{0, \ldots, 255\}$ zawiera stopnie odcieni szarości. Ustalmy relację R opisującą podobieństwo dwóch podobnych odcieni.*

$$\mu_R(x_1, x_2) = 1 - \frac{|x_1 - x_2|}{255} \tag{3.3}$$

Ćwiczenie 3.4 *Oblicz wartość μ_R z Przykładu 3.3 dla par $\langle 30, 81\rangle$, $\langle 0, 255\rangle$, $\langle 102, 204\rangle$.*

Przykład 3.5 *Niech $\mathcal{X} = \{20, 40, 60, 80, 100\}$ i relacja S na $\mathcal{X}^2$ ma postać:*

$$\begin{array}{c} \\ 20 \\ 40 \\ 60 \\ 80 \\ 100 \end{array} \begin{array}{c} \begin{array}{ccccc} 20 & 40 & 60 & 80 & 100 \end{array} \\ \left[\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0 & 0 & 0 & 0 \\ 0.8 & 0.4 & 0 & 0 & 0 \\ 1 & 0.8 & 0.4 & 0 & 0 \\ 1 & 1 & 0.8 & 0.4 & 0 \end{array}\right] \end{array} \tag{3.4}$$

Ćwiczenie 3.6 *Zaproponuj znaczenie (semantykę) dla relacji S z przykładu 3.5.*

3.1.2 Własności relacji rozmytych

Poniżej opisane są wybrane własności relacji rozmytych. Należy zauważyć, że definicje te uogólniają analogiczne własności dla relacji ostrych.

Zwrotność

Definicja 3.7 *Mówimy, że relacja rozmyta R na $\mathcal{X}^2$ jest* zwrotna *wtw.*

$$\forall_{x \in \mathcal{X}}\ \mu_R(x, x) = 1 \tag{3.5}$$

Mówimy, że R jest lokalnie zwrotna *wtw.*

$$\forall_{x_1, x_2 \in \mathcal{X}}\ \max\{\mu_R(x_1, x_2), \mu_R(x_2, x_1)\} \leq \mu_R(x_1, x_1) \tag{3.6}$$

Wniosek 3.8 *Jeżeli R zwrotna, to jest lokalnie zwrotna.*

Przykład 3.9 *Relacja R' z przykładu 3.1 nie jest zwrotna; żaden x nie jest większy od siebie samego.*
Relacja R z przykładu 3.3 jest zwrotna; dla dowolnego x, $1 - \frac{|x-x|}{255} = 1$

Ćwiczenie 3.10 *Czy relacja S z przykładu 3.5 jest zwrotna? Czy S jest lokalnie zwrotna?*

Ćwiczenie 3.11 *Podaj przykład relacji rozmytej lokalnie zwrotnej na zbiorze ludzi.*

Symetria

Definicja 3.12 *R jest symetryczna wtw*

$$\forall_{x_1,x_2} \in \mathcal{X}\ \mu_R(x_1,x_2) = \mu_R(x_2,x_1) \tag{3.7}$$

Przykład 3.13 *Relacja R' z przykładu 3.1 nie jest symetryczna; jeśli x_1 jest większy od x_2, to x_2 nie jest większy od x_1.*
Relacja R z przykładu 3.3 jest symetryczna; dla dowolnych x_1, x_2, $|x_1 - x_2| = |x_2 - x_1|$.

Ćwiczenie 3.14 *Czy relacja S z przykładu 3.5 jest symetryczna?*

Ćwiczenie 3.15 *Podaj przykład relacji rozmytej symetrycznej na zbiorze samochodów.*

Przechodniość: aksjomat i realizacje

AKSJOMAT

$$\forall x_1,x_2,x_3 \in \mathcal{X} \langle x_1,x_2\rangle \in R\ \wedge\ \langle x_2,x_3\rangle \in R \rightarrow \langle x_1,x_3\rangle \in R \tag{3.8}$$

Przykład 3.16 *Np. relacja "być przodkiem" określona na zbiorze ludzi jest przechodnia: przodek mojego przodka jest także moim przodkiem.*

Dla relacji rozmytych aksjomat przechodniości można zrealizować na kilka sposobów.

Definicja 3.17 *t-przechodnia (t-tranzytywna)*

$$\forall x_1,x_2,x_3 \in \mathcal{X}\ \mu_R(x_1,x_2)\ t\ \mu_R(x_2,x_3) \leq \mu_R(x_1,x_3) \tag{3.9}$$

lub sup-t-*przechodnia wtw.*

$$\forall x_1,x_2,x_3 \in \mathcal{X}\ \sup_{x_2\in\mathcal{X}} \{\mu_R(x_1,x_2)\ t\ \mu_R(x_2,x_3)\} \leq \mu_R(x_1,x_3) \tag{3.10}$$

Jeśli $t = \min$, mamy relację sup-min-*przechodnią*

$$\forall x_1,x_2,x_3 \in \mathcal{X}\ \sup_{x_2\in\mathcal{X}} \min\{\mu_R(x_1,x_2), \mu_R(x_2,x_3)\} \leq \mu_R(x_1,x_3) \tag{3.11}$$

Przykład 3.18 *Relacja R' z Przykł. 3.1 jest przechodnia, bo np. jeżeli* $3 > 2$ *i* $2 > 0$ *to* $3 > 0$*, czyli wg aksjomatu (3.8)*

$$\langle 3, 2\rangle \in R' \wedge \langle 2, 0\rangle \in R' \rightarrow \langle 3, 0\rangle \in R' \tag{3.12}$$

albo wg (3.9) dla $t = \min$

$$\min\{\mu_{R'}(3, 2), \mu_{R'}(2, 0)\} \leq \mu_{R'}(3, 0) \tag{3.13}$$

Przykład 3.19 *Relacja R z Przykł. 3.3 nie jest t-przechodnia. Np.* $x_1 = 0$*,* $x_2 = 102$*,* $x_3 = 204$*:*

$$\min\{1 - \frac{|0 - 102|}{255}, 1 - \frac{|102 - 204|}{255}\} \;\; ??? \;\; 1 - \frac{|0 - 204|}{255} \tag{3.14}$$

czyli

$$\min\{1 - 0.4, 1 - 0.4\} > 1 - 0.8 \tag{3.15}$$

Inne rodzaje przechodniości relacji wymieniane w literaturze to:

- the strong transitivity (silna przechodniość)
- the moderate transitivity (umiarkowana przechodniość)
- the weak stochastic transitivity (słaba przechodniość stochastyczna)
- the α-transitivity (α-przechodniość)
- the G-transitivity (G-przechodniość)

3.1.3 Podobieństwo jako relacja rozmyta

W sekcji tej rozważamy relację podobieństwa: x_1 jest podobny do x_2, co zapisujemy jako $x_1 \sim x_2$.

Z początku uważano, że podobieństwo jest *relacją równoważności* (fuzzy equivalence), czyli:

1. zwrotna
2. symetryczna
3. przechodnia

Zwrotność podobieństwa nie budzi wątpliwości: każdy jest identyczny z samym sobą (podobny w najwyższym stopniu, czyli 1 – zwrotność), a co najmniej podobny do samego siebie bardziej niż do kogokolwiek (lokalna zwrotność). Także symetria jest oczywista ("jestem podobny do kogoś tak samo jak ten ktoś do mnie"). Jednakże przeciwko zwrotności podobieństwa świadczy tzw. paradoks Poincare'a.

$$A \sim B,\ B \sim C,\ A \not\sim C \tag{3.16}$$

Przykład 3.20 *Paradoks Poincare'a zilustrować można na przykładzie szeregu 1000 filiżanek kawy, z których w pierwszej rozpuszczono jedno ziarnko cukru, w drugiej dwa, itd. Każda porcja kawy jest dokładnie tak słodka jak ona sama (zwrotność). Także słodkość dowolnych dwóch pobliskich filiżanek jest nie do odrożnienia (symetria). Jednakże jeśli skosztować kawy z filiżanki pierwszej i z ostatniej, różnica będzie zauważalna. Relacja "podobnego stopnia zasłodzenia" ;-) nie jest zatem przechodnia.*

W rzeczywistości sprawdzają się zatem modele podobieństwa jako relacji jedynie zwrotnej i symetrycznej, ale niekoniecznie przechodniej.

Definicja 3.21 *Relację R na $\mathcal{X}$ nazywamy* relacją sąsiedztwa *(neighbourhood relation) wtw. R jest zwrotna na $\mathcal{X}$ i R jest symetryczna na $\mathcal{X}$.*

Ćwiczenie 3.22 *Określ, czy relacje z przykładów 3.1, 3.3 oraz 3.5 są relacjami sąsiedztwa.*

Inne nazwy relacji sąsiedztwa

Relację sąsiedztwa określa się także jako

- non-sup-min-transitive similarity relation (nie-sup-min-przechodnia relacja podobieństwa)
- tolerance relation (relacja tolerancji)
- proximity relation (relacja bliskości)
- partial preorder relation (relacja częściowego wstępnego uporządkowania)
- resemblance relation (relacja podobieństwa)
- approximate equality relation (relacja przybliżonej równości)

3.2 Podobieństwo zbiorów

Określanie podobieństwa zbiorów opiera się o funkcję zwaną *miarą zbiorów*:

Definicja 3.23 *Niech A, B będą dowolnymi zbiorami w przestrzeni $\mathcal{X}$. Funkcję $\mu\colon \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ nazywamy miarą zbiorów wtw.*

$$\mu(\emptyset) = 0 \tag{3.17}$$

oraz

$$\mu(A \cup B) \leq \mu(A) + \mu(B) \tag{3.18}$$

W szczególności, warunek (3.18) można zawęzić do postaci:

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \tag{3.19}$$

Przykład 3.24 *Przykładowe miary zbiorów ostrych:*

1. *Dla zbiorów skończonych – miarą jest liczba elementów, tzw.* moc zbioru, *np.* $\mu(\{1,2,3\}) = 3$, $\mu(\{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}) = 4$.
2. *Dla zbiorów nieskończonych ograniczonych w* $\mathbb{R}$ *(przedziały na prostej) – długość przedziału, np.* $\mu([10,20]) = 10$.
3. *Dla zbiorów nieskończonych ograniczonych w* $\mathbb{R}^2$ *(na płaszczyźnie) – pole powierzchni.*
4. *Miarą zbioru może byc także jego liczba kardynalna, zob. Rozdział 2, Sekcja* **??***, wzory (***??***) dla zbiorów ostrych i (***??***) dla zbiorów rozmytych.*

Inne własności miary zbiorów

$$A = B \rightarrow \mu(A) = \mu(B) \tag{3.20}$$

$$A \subseteq B \rightarrow \mu(A) \leq \mu(B) \tag{3.21}$$

Implikacje odwrotne nie muszą być prawdziwe.

3.2.1 Miara Jaccarda

Miarę Jaccarda definiujemy jako

$$J(A,B) = \frac{\mu(A \cap B)}{\mu(A \cup B)} \tag{3.22}$$

Ćwiczenie 3.25 *W przestrzeni* $\mathcal{X} = \{0, \ldots, 9\}$ *definiujemy zbiór* $A = \{0,2,4,6,8\}$, $B = \{0,1,2,3,4\}$ *oraz* $C = \{5,6,7,8,9\}$. *Określ ich sumy i iloczyny, miary tych sum i iloczynów oraz stopnie podobieństwa* $J(A,B)$, $J(A,C)$ *i* $J(B,C)$, *a także* $J(A,\mathcal{X})$, $J(B,\mathcal{X})$ *i* $J(C,\mathcal{X})$.

Ćwiczenie 3.26 *W przestrzeni* $\mathbb{R}$ *defniujemy* $D = [0,10]$, $E = [0,20]$, $F = [5,15]$. *Określ ich sumy i iloczyny, miary tych sum i iloczynów oraz stopnie podobieństwa* $J(D,E)$, $J(D,F)$ *i* $J(E,F)$.

Ćwiczenie 3.27 *Dane są dwa zbiory punktów w* $\{0, \cdots, 4\}^2$*:*
$G = \{(0,0),(0,1),(0,2),(0,3),(0,4),(1,4),(2,4),(3,4),(4,4),(4,3),(4,2),(4,1),(4,0),(3,0),(2,0),(1,0)\}$, *oraz*
$H = \{(0,0),(1,1),(2,2),(3,3),(4,4),(0,4),(1,3),(3,1),(4,0)\}$. *Narysuj te zbiory i oblicz* $J(G,H)$.

Ćwiczenie 3.28 *Dane są dwa kolory w formacie RGB:* $K_1 = (100, 50, 100)$ *i* $K_2 = (101, 51, 101)$. *Oblicz* $J(K_1, K_2)$. *Czy wynik zgadza się z intuicją?*

Wniosek 3.29 *Dla każdych* A *i* B *w* $\mathcal{X}$

$$0 \leq J(A, B) \leq 1 \tag{3.23}$$

A zatem miara Jaccarda nadaje się na funkcję przynależności relacji rozmytej.

3.2.2 Simple matching coefficient

Współczynnik prostego porównania zbiorów – rozszerzona miara Jaccarda dla zbiorów A, B w $\mathcal{X}$.

$$\hat{J}(A, B) = \frac{\mu(A \cap B) + \mu(A^c \cap B^c)}{\mu(A \cup B) + \mu(A^c \cap B^c)} \tag{3.24}$$

co można tez zapisać jako:

$$\hat{J}(A, B) = \frac{\mu(A \cap B) + \mu(A^c \cap B^c)}{\mu(\mathcal{X})} \tag{3.25}$$

gdzie A^c, B^c – dopełnienia zbiorów A, B. Znaczenie $\hat{J}$ jest następujące: na podobieństwo zbiorów mają wpływ nie tylko elementy jednocześnie należące do tych zbiorów, ale także elementy jednocześnie nie należące do tych zbiorów. Inaczej mówiąc: na podobieństwo obiektów wływają nie tylko wspólne własności, ale także wspólne braki.

Ćwiczenie 3.30 *Oblicz* $\hat{J}(A, B)$, $\hat{J}(A, C)$, $\hat{J}(B, C)$ *dla zbiorów z Ćw. 3.25.*

Ćwiczenie 3.31 *Oblicz* $\hat{J}(D, E)$, $\hat{J}(D, F)$, $\hat{J}(E, F)$ *dla zbiorów z Ćw. 3.26.*

Ćwiczenie 3.32 *Oblicz* $\hat{J}(G, H)$ *dla zbiorów z Ćw. 3.27.*

Ćwiczenie 3.33 *Czy możliwe jest obliczenie* $\hat{J}(K_1, K_2)$ *dla zbiorów z Ćw. 3.28?*

3.2.3 Współczynnik r_{card}

Miara ta, w odróżnieniu od innych miar przedstawionych w tym podrozdziale określa podobieństwo dwóch zbiorów w kontekście pewnej relacji.

$$r_{card}(A, B) =_{df} \frac{card(C)}{card(A \times B)} \tag{3.26}$$

gdzie $C = (A \times B) \cap R$ – zbiór w $\mathcal{X} \times \mathcal{Y}$, zaś R – relacja na $\mathcal{X} \times \mathcal{Y}$ opisująca pewną zależność pomiędzy elementami tych przestrzeni.

Ćwiczenie 3.34 *Oblicz współczynnik* r_{card} *dla zbiorów* A, B *z Ćw. 3.27 w kontekście relacji* $x > y$.

3.2.4 Miara Dice'a i miara zachodzenia

Miara Dice'a

$$\mu_D(A,B) = \frac{\mu(A \cap B)}{\mu(A) + \mu(B)} \tag{3.27}$$

Ćwiczenie 3.35 *Oblicz $\mu_D(A,B)$, $\mu_D(A,C)$, $\mu_D(B,C)$ dla zbiorów z Ćw. 3.25.*

Ćwiczenie 3.36 *Oblicz $\mu_D(D,E)$, $\mu_D(D,F)$, $\mu_D(E,F)$ dla zbiorów z Ćw. 3.26.*

Ćwiczenie 3.37 *Oblicz $\mu_D(G,H)$ dla zbiorów z Ćw. 3.27.*

Miara zachodzenia, *the overlapping measure*

$$\mu_O(A,B) = \frac{\mu(A \cap B)}{\min\{\mu(A), \mu(B)\}} \tag{3.28}$$

Ćwiczenie 3.38 *Oblicz $\mu_O(A,B)$, $\mu_O(A,C)$, $\mu_O(B,C)$ dla zbiorów z Ćw. 3.25.*

Ćwiczenie 3.39 *Oblicz $\mu_O(D,E)$, $\mu_O(D,F)$, $\mu_O(E,F)$ dla zbiorów z Ćw. 3.26.*

Ćwiczenie 3.40 *Oblicz $\mu_O(G,H)$ dla zbiorów z Ćw. 3.27.*

Ćwiczenie 3.41 *Czy miara Dice'a i miara zachodzenia mogą być użyte jako funkcje przynależności do relacji rozmytej?*

3.3 Podobieństwo wektorów liczbowych

Miary opisane w tej sekcji zależą nie od występowania elementu zbioru w sumie lub w części wspólnej porównywanych zbiorów, ale od jego wartości liczbowej. Niech $V_1 = \{v_{11}, v_{12}, ..., v_{1n}\}$, $V_2 = \{v_{21}, v_{22}, ..., v_{2n}\}$ – wektory (ciągi) w $\mathcal{R}^n$, $n \in \mathcal{N}$.

3.3.1 Współczynnik korelacji

$$r_{cc}(V_1,V_2) = \frac{\sum_{i=1}^{n} |v_{1,i} - av(V_1)| \cdot |v_{2,i} - av(V_2)|}{\sqrt{\sum_{i=1}^{n} (v_{1,i} - av(V_1))^2 \cdot \sum_{i=1}^{n} (v_{2,i} - av(V_2))^2}} \tag{3.29}$$

gdzie $av(V_1) = \frac{1}{n}\sum_i^n V_{1,n}$ – średnia wartość wyrazów V_1, $av(V_2)$ – analogicznie.

Ćwiczenie 3.42 *Oblicz podobieństwo $r_{cc}(A,B)$, $r_{cc}(A,C)$, $r_{cc}(B,C)$ dla zbiorów z Ćw. 3.25.*

Ćwiczenie 3.43 *Oblicz podobieństwo r_{cc} dla kolorów K_1, K_2 z Ćw. 3.28. Czy wynik jest zgodny z intuicją?*

3.3.2 Amplituda kosinusowa

$$r_{ca}(V_1, V_2) = \frac{\left| \sum_{i=1}^{n} v_{1i} \cdot v_{2i} \right|}{\sqrt{\sum_{i=1}^{n} v_{1i}^2 \cdot \sum_{i=1}^{n} v_{2i}^2}} \tag{3.30}$$

na bazie iloczynu skalarnego V_1, V_2

$$V_1 \circ V_2 = \sum_{i=1}^{n} v_{1i} \cdot v_{2i} \tag{3.31}$$

Przykład 3.44 *Przykład obliczenia amplitudy kosinusowej dla wektorów w* $\mathbb{R}^3$ *pokazuje przykład 3.50 na stronie 28.*

Ćwiczenie 3.45 *Oblicz podobieństwo* $r_{ca}(A, B)$, $r_{ca}(A, C)$, $r_{ca}(B, C)$ *dla zbiorów z Ćw. 3.25. Porównaj wyniki z otrzymanymi w Ćw. 3.42 – co zauważyłeś?.*

Ćwiczenie 3.46 *Oblicz podobieństwo* $r_{ca}(K_1, K_2)$ *dla zbiorów z Ćw. 3.28 i porównaj z wynikami z Ćw. 3.43.*

3.3.3 Minimum-maximum (min-max)

$$r_{mm}(V_1, V_2) = \frac{\sum_{i=1}^{n} \min\{v_{1i}, v_{2i}\}}{\sum_{i=1}^{n} \max\{v_{1i}, v_{2i}\}} \tag{3.32}$$

Ćwiczenie 3.47 *Oblicz* $r_{mm}(A, B)$, $r_{mm}(A, C)$, $r_{mm}(B, C)$ *dla zbiorów z Ćw. 3.25.*

Ćwiczenie 3.48 *Oblicz* $r_{mm}(K_1, K_2)$ *dla zbiorów z Ćw. 3.28.*

3.3.4 Średnia arytmetyczna-minimum

$$r_{am}(V_1, V_2) = \frac{\sum_{i=1}^{n} \min\{v_{1i}, v_{2i}\}}{\frac{1}{2} \sum_{i=1}^{n} (v_{1i} + v_{2i})} \tag{3.33}$$

Ćwiczenie 3.49 *Oblicz* $r_{am}(A, B)$, $r_{am}(A, C)$, $r_{am}(B, C)$ *dla zbiorów z Ćw. 3.25.*

Ćwiczenie 3.50 *Oblicz* $r_{am}(K_1, K_2)$ *dla zbiorów z Ćw. 3.28.*

3.3.5 Średnia geometryczna-minimum

$$r_{am}(V_1, V_2) = \frac{\sum_{i=1}^{n} \min\{v_{1i}, v_{2i}\}}{\sum_{i=1}^{n} \sqrt{v_{1i} v_{2i}}} \tag{3.34}$$

Ćwiczenie 3.51 *Oblicz $r_{gm}(A, B)$, $r_{gm}(A, C)$, $r_{gm}(B, C)$ dla zbiorów z Ćw. 3.25.*

Ćwiczenie 3.52 *Oblicz $r_{gm}(K_1, K_2)$ dla zbiorów z Ćw. 3.28.*

3.3.6 Metoda moduł-eksponent

$$r_{me}(V_1, V_2) = \exp\left(-\sum_{i=1}^{n} |v_{1i} - v_{2i}|\right) \tag{3.35}$$

Ćwiczenie 3.53 *Oblicz $r_{me}(A, B)$, $r_{me}(A, C)$, $r_{me}(B, C)$ dla zbiorów z Ćw. 3.25.*

Ćwiczenie 3.54 *Oblicz $r_{me}(K_1, K_2)$ dla zbiorów z Ćw. 3.28.*

Ćwiczenie 3.55 *Zbierz w tabelce wyniki z Ćw. 3.42–3.54 i porównaj. Jakie wnioski nasuwają się co do sposobów działania tych metod?*

3.4 Odległości zbiorów danych

3.4.1 Metryka jako funkcja

Definicja 3.56 *Funkcja $\rho\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+\{0\}$ jest* metryką *w $\mathcal{X}$ wtw*

$$\forall a \in \mathcal{X}\ \rho(a, a) = 0 \tag{3.36}$$

$$\forall a, b \in \mathcal{X}\ \rho(a, b) = \rho(b, a)\ \text{(symetria)} \tag{3.37}$$

$$\forall a, b, c \in \mathcal{X}\ \rho(a, c) + \rho(b, c) \geq \rho(a, b)\ \text{(nierówność trójkąta)} \tag{3.38}$$

Przykład 3.57 *Najczęściej używaną metryką jest odległość Euklidesa na płaszczyźnie lub w przestrzeni. Dla x_1, x_2 w $\mathbb{R}$ $\rho_E(x_1, x_2) = |x_1 - x_2|$, w $\mathbb{R}^2$ $\rho_E((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, i ogólnie w $\mathbb{R}^n$ dla $X = \{x_1, \ldots, x_n\}$ i $Y = \{y_1, \ldots, y_n\}$*

$$\rho_E(X, Y) = \sqrt{\sum\nolimits_{i=1}^{n} (x_i - y_i)^2} \tag{3.39}$$

Ćwiczenie 3.58 *Oblicz odległość Euklidesa pomiędzy kolorami K_1, K_2 z Ćw. 3.28.*

3.4.2 Metryka a podobieństwo

Odległość pomiędzy obiektami (X, Y) oraz ich podobieństwo można związać poprzez malejącą i wykładniczo się zachowującą funkcję g od $\rho(a, b)$

$$sim(X, Y) = g\Big(\rho(X, Y)\Big) \tag{3.40}$$

Jeśli istnieje $\sup \rho(X, Y) = \rho_{\max}$ w danej $\mathcal{X}$ to dowolną odległość $\rho(X, Y)$ można zamienić na wartość funkcji przynależności np. według

$$\mu_R(X, Y) = 1 - \left(\frac{\rho(X, Y)}{\rho_{\max}}\right)^r \quad r \in \mathbb{R}^+ \tag{3.41}$$

Wzory (3.35), (3.41) są przykładami funkcji g.

3.4.3 Metryka miejska

Od ang. *city metric*, zwana także "metryką taksówkową". Dla $X = \{x_1, \ldots, x_n\}$ i $Y = \{y_1, \ldots, y_n\}$

$$\rho_C(X, Y) = \sum\nolimits_{i=1}^{n} |x_i - y_i| \tag{3.42}$$

Ćwiczenie 3.59 *Oblicz metrykę miejską dla kolorów K_1, K_2 z Ćw. 3.28.*

Ćwiczenie 3.60 *Oblicz maksymalną odległość pomiędzy kolorami w przestrzeni RGB i zaproponuj na bazie (3.41) funkcję podobieństwa kolorów opartą o ρ_C.*

3.4.4 Odległość Hamminga

Odległość Hamminga definiujemy jako ilość różnic pomiędzy dwoma ciągami bitów o tej samej długości.

$$H(V_1, V_2) = \sum\nolimits_{i=1}^{n} h(i) \tag{3.43}$$

gdzie

$$h(i) = \begin{cases} 0, & \text{jeśli } v_{1i} = v_{2i} \\ 1, & \text{w przeciwnym przypadku} \end{cases} \tag{3.44}$$

Przykład 3.61 $V_1 = 1000110101$, $V_2 = 0000111001$, $n = 10$.
Zatem $H(V_1, V_2) = 3$.

Zauważmy, że miara ta oparta jest na funkcji XOR (eXclusive OR), czyli na alternatywie wyłączającej.

Późniejsze zastosowania rozszerzyły definicję dystansu Hamminga na ciągi dowolnych jednostek informacji, zwłaszcza liczbowej i tekstowej. Możemy zdefiniować tę odległość jako *ilość operacji zastępowania (replace) konieczną do zamiany jednego ciągu na drugi.*

Tablica 3.1: Obliczanie odległości Hamminga w przykł. 3.61

bit nr	1	2	3	4	5	6	7	8	9	10
wektor V_1	**1**	0	0	0	1	**0**	0	1	**1**	1
wektor V_2	**0**	0	0	0	1	**1**	0	1	**0**	1
	1					1			1	

Przykład 3.62 *Niech dane będą dwie różne wersje tej samej tabeli bazy danych. Odległością Hamminga pomiędzy nimi nazwiemy ilość operacji* `UPDATE`[1]*, które należy wykonać na jednej z nich, aby przekształcić ją w drugą.*

Ćwiczenie 3.63 *Jaka jest maksymalna odległość Hamminga pomiędzy kolorami w przestrzeni RGB?*

Poniżej podane sa przykładowe zastosowania odległości Hamminga do wyznaczenia stopnia podobeństwa pomiędzy zbiorami.

$$\mu_R(V_1, V_2) = 1 - \frac{H(V_1, V_2)}{n} \tag{3.45}$$

$$\mu_R(V_1, V_2) = 1 - \sqrt{\frac{H(V_1, V_2)}{n}} \tag{3.46}$$

$$\mu_R(V_1, V_2) = 1 - \left(\frac{H(V_1, V_2)}{n}\right)^r, \quad r \in (0, +\infty) \tag{3.47}$$

Równania te są egzemplifikacjami formuły (3.41). Zauważmy, że dla każdej z formuł (3.45)–(3.47)

$$0 \leq \frac{H(V_1, V_2)}{n} \leq 1 \tag{3.48}$$

czyli że wartość mieści się w zakresie charakterystycznym dla funkcji przynależności relacji rozmytej.

3.4.5 Edit distance (odległość edycyjna)

Podobna do odległości Hamminga, z tym że zamiast zastępowania rozważa się tu operacje **kasowania** i **wstawiania**. Istotne jest to, że pozwala na porównywanie ciągów o różnych dugościach.

Przykład 3.64 *Niech wektory V_1, V_2 – jak w przykładzie 3.61.*
Zatem $E(V_1, V_2) = 6$.

[1]Zakładamy, że pojedyncza operacja `UPDATE` aktualizuje cały rekord.

Tablica 3.2: Obliczanie odległości edycyjnej w przykładzie 3.64

bit nr	1	2	3	4	5	6	7	8	9	10
wektor V_1	**1**	0	0	0	1	**0**	0	1	**1**	1
wektor V_2	**0**	0	0	0	1	**1**	0	1	**0**	1
kasowanie	1					1			1	
wstawianie	1					1			1	

Przykład 3.65 *Niech* $V_3 = 11111 \in \{0,1\}^5$, $V_4 = 110111 \in \{0,1\}^6$.
$E(V_3, V_4) = 1$ – *należy wstawić 0 na trzeciej pozycji w* V_3 *by otrzymać* V_4.

Przykład 3.66 *Przy założeniach jak w przykładzie 3.62, odległością edycyjną pomiędzy tabelami nazwiemy ilość operacji* `DELETE` *oraz* `INSERT`*, które należy wykonać na jednej z nich, aby przekształcić ją w drugą.*

Odległość edycyjna może posłużyć do budowy funkcji podobieństwa analogicznie do formuł (3.45)–(3.47).

Ćwiczenie 3.67 *Oblicz* $E(V_1, V_2)$ *dla ciągów z Przykładu. 3.61.*

Ćwiczenie 3.68 *Jaka jest maksymalna odległość edycyjna pomiędzy kolorami w przestrzeni RGB?*

3.4.6 Loewenstein distance (odległość Loewensteina)

Jest uogólnienieniem odległości Hamminga i edycyjnej – opiera się na wszystkich trzech operacjach na elementach ciągu, czyli **zastępowania, kasowania** i **wstawiania**. Podobnie jak *edit distance* pozwala na porównywanie ciągów o różnych dugościach.

W przecwieństwie do dwóch poprzednich odległości, "Lewensztajna" nie da się określić jednoznacznie. Widoczne jest to w szczególności, gdy mamy do wyboru zastąpienie albo skasowanie+wstawienie jakiegoś elementu. Przyjmuje się zatem, że odległością Loewensteina pomiędzy ciągami elementów jest **najmniejsza ilość** operacji zastąpienia, kasowania i wstawiania, jaką należy wykonać, aby jeden z ciągów przekształcić w drugi.

Inne odległości pomiędzy zbiorami

1. Wagner–Fischer distance (pomiędzy łańcuchami tekstowymi)
2. Canberra metric
3. metryka Minkowskiego

4. r-metryka (r-metric)
5. Bhattacharyya distance
6. Hausdorff distance
7. dissemblance index

3.5 Podobieństwo wzorców tekstowych

Dlaczego tekst wymaga osobnych miar podobieństwa? Ze względu na dodatkową informację zawartą w kodach liter. Numeryczna interpretacja tej informacji nie jest wystarczająca, np. A (65) i B (66) to dwie różne jednostki informacji, choć ich kody są bardzo bliskie w sensie liczbowym. Przeciwnie, A (65) i a (97) to właściwie ta sama litera, choć ich kody są różne i odległe.

Przykład 3.69 *Niech* S_1=ACGT, S_2=BDHU. *W ASCII* $S_1 = \{65,67,71,84\}$, $S_2 = \{66,68,72,85\}$. *Współczynnik korelacji (3.29) tych dwóch ciągów wynosi* $r_{cc}(S_1, S_2) = 1$, *zaś ich podobienstwo wg miary Jaccarda (3.22)* $J(S_1, S_2) = 0$ *(pusta część wspólna ciągów).*

3.5.1 Term frequency matrix

Czyli "macierz częstości występowania terminów". Określa podobieństwo dokumentów d_1 i d_2 ze względu na wybrany zbiór terminów, np. słów kluczowych $\{t_1, t_2,..., t_n\}$.

$$\begin{array}{c} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_n \end{array} \\ \left[\begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \end{array} \right] \end{array} \tag{3.49}$$

Podobieństwo otrzymanych wektorów oblicza się przy pomocy amplitudy kosinusowej (3.30) lub, w ogólności, dowolnej innej miary podobieństwa wektorów liczbowych, np. (3.29), (3.32), (3.33) lub (3.34), patrz sekcja 3.3. Zauważmy, iż wynik nie zależy od długości porównywanych dokumentów.

Przykład 3.70 *Niech* d_1, d_2 *będą tekstami, których wzajemne podobieństwo mamy określić. Niech* t_1="rower", t_2="kolarz", t_3="góry" *będą słowami kluczowymi. Niech np.*

$$\begin{array}{c} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{ccc} \textit{rower} & \textit{kolarz} & \textit{góry} \end{array} \\ \left[\begin{array}{ccc} 2 & 0 & 2 \\ 1 & 2 & 3 \end{array} \right] \end{array} \tag{3.50}$$

co oznacza, że "rower" występuje dwa razy w d_1 i raz w d_2, itd. Zatem:

$$\begin{aligned} r_{ca}(d_1, d_2) &= \frac{2 \cdot 1 + 0 \cdot 2 + 2 \cdot 3}{\sqrt{(2^2 + 0^2 + 2^2) \cdot (1^2 + 2^2 + 3^2)}} = \\ &= \frac{2 + 6}{\sqrt{(4 + 4) \cdot (1 + 4 + 9)}} = \frac{8}{10.58} = 0.76 \end{aligned} \tag{3.51}$$

Oczywiste jest, że dane dwa dokumenty moga wykazywać bardzo bliskie podobieństwo lub jego zupełny brak, w zależności od wybranego zbioru słów kluczowych.

3.5.2 Metoda n-gramów

Metoda ta określa podobieństwo łańcuchów tekstowych s_1, s_2 w oparciu o **ilość wspólnych podciągów** n-elementowych, czyli n**-gramów**.

$$sim_n(s_1, s_2) = \frac{1}{N - n + 1} \sum_{i=1}^{N-n+1} h(i) \tag{3.52}$$

gdzie

$h(i) = 1$ jeśli n-elementowy podciąg zaczynający się od i-tej pozycji w s_1 występuje przynajmniej raz w s_2 (w przeciwnym przypadku $h(i) = 0$);

$N - n + 1$ – ilość możliwych n-elementowych podciągów w s_1

Metoda n-gramów wykorzystywana jest najczęściej dla $n = 3$. Wówczas zwana jest metodą *trigramów*, zaś formuła (3.52) przybiera postać

$$sim_3(s_1, s_2) = \frac{1}{N - 2} \sum_{i=1}^{N-2} h(i) \tag{3.53}$$

Przykład 3.71 *Porównanie słów* s_1=SUMMARY, s_2=SUMMARIZATION.
$N(s_1) = 7$, $N(s_2) = 13$, $N = \max\{N(s_1), N(s_2)\} = 13$.

$$sim_3(s_1, s_2) = \frac{1}{11} \sum_{i=1}^{11} h(i) = \frac{4}{11} \simeq 0.29 \tag{3.54}$$

ponieważ 4 trigramy w SUMMARY*:* SUM, UMM, MMA, *and* MAR *występują w* SUMMARIZATION.

3.5.3 Uogólniona miara n-gramów (Miara Niewiadomskiego)

Oryginalnie n-gramy badają podobieństwo słów w oparciu o podciągi jednej tylko długości. Dokładniejsze badanie podobieństwa wymaga sprawdzenia podciągów różnych długości, generalnie od jedno- do N-elementowych, gdzie N jest długością słowa:

$$\mu_N(s_1, s_2) = \frac{2}{N^2 + N} \sum_{i=1}^{N(s_1)} \sum_{j=1}^{N(s_1)-i+1} h(i, j) \tag{3.55}$$

gdzie

$h(i, j) = 1$ jeśli i-elementowy podciąg w słowie s_1 zaczynający się od j-tej pozycji w słowie s_1 pojawia się przynajmniej raz w słowie s_2 (inaczej $h(i, j) = 0$);

$N(s_1), N(s_2)$ – ilość liter w słowach s_1 i s_2;

$N = \max\{N(s_1), N(s_2)\}$;

$\frac{N^2+N}{2}$ – ilość możliwych podciągów od 1-elementowych do N-elementowych w słowie o dlugości N

Przykład 3.72 *Niech* s_1=PROGRAMMER, s_2=PROGRAMMING, *zatem* $N(s_1) = 10$, $N(s_2) = 11$, $\max\{N(s_1), N(s_2)\} = 11$. *Obliczamy:*

$$\begin{aligned} \mu_N(s_1, s_2) &= \frac{2}{121 + 11} \sum_{i=1}^{11} \sum_{j=1}^{11-i+1} h(i, j) = \\ &= \frac{9 + 7 + 6 + 5 + 4 + 3 + 2 + 1}{66} \simeq 0.561 \end{aligned} \tag{3.56}$$

ponieważ w s_2 *występują następujące podciągi z* s_1*:*

9 1-elementowych P, R, O, G, R, A, M, M, R*;*

7 2-elementowych PR, RO, OG, GR, RA, AM, MM*;*

6 3-elementowych PRO, ROG, OGR, GRA, RAM, AMM*;*

...

1 8-elementowy PROGRAMM.

Zaproponowana relacja jest zwrotna – każde słowo jest identyczne z sobą samym, $\mu_N(s_1, s_1) = 1$. Relacja natomiast nie jest symetryczna. Dla zapewnienia własności symetrii obliczamy podobienstwo słów jako:

$$\mu_{Nsym}(s_1, s_2) = \min\{\mu_N(s_1, s_2), \mu(s_2, s_1)\} \tag{3.57}$$

Relacja opisana funkcją przynależności μ_{Nsym} jest relacją sąsiedztwa.

3.5.4 Uogólniona miara n-gramów z ograniczeniami

Miara podobieństwa słów zaproponowana w sekcji 3.5.3 pomimo swojej dokładności, jest bardzo kosztowna obliczeniowo. Aby uniknąć zbędnego porównywania podciągów zbyt krótkich, np. 1-literowych, które zaciemniają wynik podobieństwa[2], a także zbyt długich, np. dłuższych niż jedno z porównywanych słów, wprowadza się górne i dolne ograniczenie długości podciągów których występowanie w porównywanych słowach jest sprawdzane.

$$\mu_N(s_1, s_2) = f(N, n_1, n_2) \sum_{i=n_1}^{n_2} \sum_{j=1}^{N(s_1)-i+1} h(i, j) \tag{3.58}$$

gdzie

$$f(N, n_1, n_2) = \frac{2}{(N - n_1 + 1)(N - n_1 + 2) - (N - n_2 + 1)(N - n_2)} \tag{3.59}$$

wyraża ilość możliwych podciągów o długościach od n_1 do n_2, $1 \leq n_1 \leq n_2 \leq N$, zaś pozostale symbole – jak w (3.55).

Ćwiczenie 3.73 *Oblicz podobieństwo słów podanych w Przykładzie 3.72 dla* $n_1 = 2$ *i* $n_2 = 3$.

Ćwiczenie 3.74 *Sprawdź czy miara (3.58) dla* $n_1 = n_2 = n$ *odpowiada metodzie n-gramów opisanej w sekcji 3.5.2.*

3.5.5 Miara podobieństwa zdań

Na tej podstawie określić można także formułę podobieństwa dla zdań, traktowanych jako zbiory (ale nie ciągi) wyrazów.

$$\mu_{N_Z}(z_1, z_2) = \frac{1}{N} \sum_{i=1}^{N(z_1)} \max_{j=1,\ldots,N(z_1)} \mu_N(s_{1j}, s_{2i}) \tag{3.60}$$

gdzie:

s_{1i} – i-ty wyraz w zdaniu z_1;

s_{2j} – j-ty wyraz w zdaniu z_2;

$\mu_N(s_{1i}, s_{2j})$ – wartość funkcji (3.55) dla (s_{1i}, s_{2j});

$N(z_1), N(z_2)$ – liczba słów w zdaniach s_1, s_2;

[2]

$N = \max\{N(z_1), N(z_2)\}$

Przykład 3.75 *Niech* s_1=JOHN WALKS, s_2=JOHN IS WALKING, $N(s_1) = 2$, $N(s_2) = N = 3$. *Zatem poprzez (3.60) mamy*

$$\mu_{N_Z}(z_1, z_2) = \frac{1}{3}\sum_{i=1}^{3}\max_{j=1,2}\{\mu_N(s_{1i}, s_{2j})\} \tag{3.61}$$

Stąd

$$\mu_{N_Z}(z_1, z_2) = \frac{0.067 + 1 + 0.357}{3} \simeq 0.457 \tag{3.62}$$

gdzie

0.067 $= \max\{\mu_{N_{sym}}(\text{JOHN}, \text{IS}), \mu_{N_{sym}}(\text{WALKS}, \text{IS})\}$;

1.0 $= \max\{\mu_{N_{sym}}(\text{JOHN}, \text{JOHN}), \mu_{N_{sym}}(\text{WALKS}, \text{JOHN})\}$;

0.357 $= \max\{\mu_{N_{sym}}(\text{JOHN}, \text{WALKING}), \mu_{N_{sym}}(\text{WALKS}, \text{WALKING})\}$.

3.5.6 Koszt obliczeniowy n-gramów. Współczynnik P2P

Za jednostkę kosztu przyjmujemy pojedyncze porównanie dwóch liter, które daje rezultat "takie same" bądź "różne".

Przykład 3.76 *Porównanie uogólnioną metodą n-gramów ciągów "AB" i "CD" to koszt 6-ciu operacji: (A,C), (A, D), (B,C), (B,D) – koszty pojedyncze oraz (AB, CD) – koszt podwójny.*

Ogólnie koszt porównania dwóch słów wyraża się wzorem

$$cost(s_1, s_2) = \sum\nolimits_{i=1}^{N}(N - i + 1)^2 \cdot i \tag{3.63}$$

gdzie N – ilość liter w dłuższym z s_1, s_2. Jeśli przyjąć ograniczenia długości badanych podciągów n_1 i n_2, równanie (3.63) przybiera postać

$$cost(s_1, s_2) = \sum_{i=n_1}^{n_2}(N(s_1) - i + 1) \cdot (N(s_2) - i + 1) \cdot i \tag{3.64}$$

Na tej podstawie określić można współczynnik P2P – "performance to price", opisujący stopień podobieństwa słów s_1, s_2 w stosunku do poniesionego kosztu obliczeniowego.

$$P2P = \frac{\text{similarity}}{\text{cost}} = \frac{\mu_N(s_1, s_2)}{cost(s_1, s_2)} \tag{3.65}$$

Tablica 3.3: Podobieństwa i koszty dla s_1=CASE i s_2=CASUALTY dla wybranych wartości n_1 i n_2

Sim.	0.233	0.230	0.167	0.091	0.286	0.23	0.16	0.33	0.286
Cost	540	130	98	56	110	78	36	74	42
P2P	.0004	.0017	.0017	.0016	.0026	.0029	.0046	.0045	.007
n_1	1	1	2	3	1	2	3	1	2
n_2	max	4	4	4	3	3	3	2	2

Inne metody porównywania łańcuchów tekstowych

1. Knutt–Morris–Pratt alg.,
2. Brute force similarity – algorytm brutalnej siły
3. Matching with finite automata – porównywanie metodą automatów skończonych
4. Boyer–Moore alg.
5. Karp–Rabin alg.
6. Wagner–Fischer distance
7. String kernels (jądra łańcuchowe)