Emilia Stańczyk-Wołowiec

Język naturalny jako pomost między danymi cyfrowymi maszyny a rozumieniem człowieka

Wstęp

Współczesna rzeczywistość wymaga gromadzenia olbrzymich ilości danych. Lawinowo rosną archiwa informacji dotyczących najróżniejszych dziedzin (bankowych, personalnych, pogodowych itp.). Jednocześnie istnieje potrzeba szybkiego przetwarzania danych, a następnie na ich podstawie raportowania, podejmowania decyzji czy oceny ryzyka procesu opisywanego przez bazę danych.

Dla człowieka ogarnięcie dużych ilości danych liczbowych oraz operowanie na nich jest trudne i niewygodne, a czasami po prostu niemożliwe. Oczekujemy informacji w języku naturalnym, który może nie jest szczególnie precyzyjny, ale za to powszechnie zrozumiały. Zwłaszcza kiedy chodzi jedynie o szybką ocenę sytuacji, przedkładamy określenia takie jak „bardzo dużo", „średnio", „prawie nikt" nad dane procentowe. Dlatego też w pewnych dziedzinach język naturalny powinien być pomostem między maszyną przetwarzającą dane a człowiekiem, dla którego są one przeznaczone. (Np. kiedy interesujemy się pogodą w Zakopanem, wystarczy wiedzieć, że od lipca do września opady są tam niewielkie, a od grudnia do lipca pokrywa śniegu jest duża. Niepotrzebne są dane liczbowe opadów z każdego miesiąca z ostatnich dziesięciu lat).

Lingwistyczne streszczenia baz danych

Ostatecznym podmiotem oceniającym sytuację lub proces jest zawsze człowiek. Na podstawie metod statystycznych i inteligentnego szacowania dokonuje on podsumowania wiedzy zawartej w bazie danych. Następnie formułuje ogólnozrozumiałe wnioski w języku naturalnym, np. w formie raportu, notatek czy prezentacji. Jednakże maszyna może wspomóc człowieka na tym etapie. Publikacje ostatnich lat [Kacprzyk, Yager 2001; Kacprzyk, Yager 2006; Niewiadomski 2006a, 2006b] opisują możliwość zastosowania zbiorów rozmytych Zadeha [1965] do tworzenia lingwistycznych podsumowań relacyjnych baz danych. Innymi słowy, jest to praktyczne zastosowanie logiki rozmytej w relacyjnych bazach danych.

Automaty podsumowujące (czyli programy komputerowe) na podstawie baz danych (często olbrzymich) tworzą krótkie wiadomości tekstowe, streszczenia i podsumowania w języku naturalnym. Wiadomości te są wykorzystywane jako new-

sy, notatki prasowe, wnioski raportów itp. Działanie automatów jest wystarczające w przypadku większości zastosowań, kiedy chodzi o oszacowanie lub podsumowanie dużej ilości danych.

Ogólna budowa algorytmu

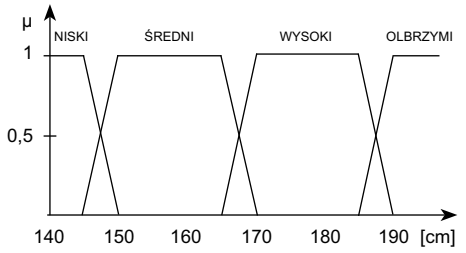
W praktycznym zastosowaniu jest dziś wiele aplikacji zawierających algorytmy podsumowujące. Ogólnie rzecz biorąc, w algorytmie tego rodzaju możemy wyróżnić cztery zasadnicze elementy:

- bazę danych – przy czym obszar roboczy nie musi obejmować całej bazy, a np. konkretne widoki (dane o pracownikach, dane o dochodach firmy, dane o warunkach pogodowych w wybranym miejscu na świecie),
- moduł atrybutów,
- moduł funkcji charakterystycznych,
- właściwy generator.

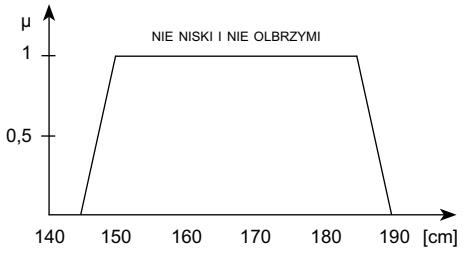
Tabela 1. Przykładowa tabela w bazie danych

ID	Nazwisko	Wiek	Staż	Zarobki
1	A.	35	10	1200
2	B.	20	1	1000
3	C.	49	6	5500
4	D.	52	25	2000

Relacyjna baza danych składa się z tabel, a te z rzędów i kolumn. Kolumna opisuje konkretny atrybut (w tabeli 1. wiek, staż, nazwisko, zarobki), a wiersz – konkretny przypadek (w tabeli 1. konkretnego pracownika). W tym miejscu należy zaznaczyć, że stosowanie opisywanych automatów dla baz zawierających mniej niż 10 000 rekordów jest mało zasadne.

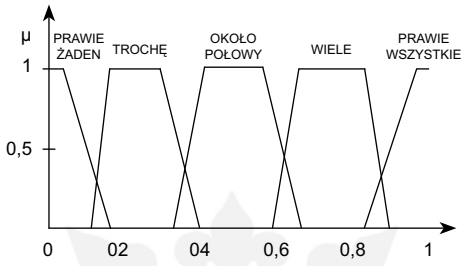


Rycina 1. Przykładowa funkcja charakterystyczna (przynależności)



Rycina 2. Przykładowa funkcja charakterystyczna (przynależności)

Moduł atrybutów jest częścią uwidocznioną w interfejsie. Służy użytkownikowi do wyboru atrybutów, które go interesują (np. wiek, zarobki), oraz kwantyfikatorów (np. „mało", „ponad połowa"). Kwantyfikatory opisują liczbę – dzielimy je na bezwzględne (np. „mniej niż 30", „około 100", „ponad 1000") i relatywne (np. „około 1/4", „prawie żaden", „kilka"). Efektem końcowym pracy tego modułu jest sumaryzator, dla którego jest szukane podsumowanie (np. „niska pensja", „średni wiek i krótki staż pracy").



Rycina 3. Przykładowe funkcje przynależności dla kwantyfikatorów rozmytych

Funkcja charakterystyczna jest matematycznym odwzorowaniem między wielkością liczbową a określeniem w języku naturalnym.

Generator korzysta z trzech opisanych wyżej modułów. Jego zadaniem jest dopowiedzieć na jedno z pytań:

A) „Jak wiele rekordów spełnia warunek S?" (np. „Jak wielu pracowników ma wysoką pensję?").

B) „Jak wiele rekordów, które spełniają warunek S_1, spełnia również warunek S_2?" (np. „Jak wielu pracowników, którzy są w młodym wieku, ma wysokie zarobki?").

Pytanie A zostało sformalizowane przez Zadeha (965) następująco:

Q P jest/ma S [T],

gdzie: Q jest lingwistycznym określeniem liczności, czyli kwantyfikatorem rozmytym, np. „około połowy", „dużo", określającym, ile rekordów z podsumowywanej bazy spełnia określenie cechy S; P jest podmiotem podsumowania i oznacza rzeczywiste obiekty, których dane są przechowywane w rekordach w bazie; S (sumaryzator) jest określeniem cechy, ze względu na którą jest dokonywane podsumowanie, np. „młody wiek", „bardzo duża prędkość" (jest to zbiór rozmyty w przestrzeni X); T jest wskaźnikiem trafności podsumowania określającym, w jakim stopniu kwantyfikator Q poprawnie oddaje ilość rekordów w pełni lub częściowo charakteryzowanych przez S.

Podsumowanie B ma postać:

Q P, które ma cechę S_1, posiada (również) cechę S_2 [T].

Generator odnajduje w bazie danych atrybut, który interesuje użytkownika. Tworzy sumaryzator, a następnie sprawdza, jak wiele rekordów spełnia warunek sumaryzatora. Oblicza ich liczebność i odnosi do określenia lingwistycznego („mało", „dużo", „średnio" itp.).

Ostatnim krokiem generatora jest ocena jakości podsumowania, wyrażana wskaźnikiem trafności z przedziału [0,1]. W praktyce interesują nas podsumowania, których wskaźnik trafności jest bliski 1.

Praktyczne realizacje

Poniżej przedstawiono kilka praktycznych realizacji podsumowań przeprowadzonych na rzeczywistych bazach danych. W nawiasach podano wskaźniki trafności.

1. **Baza danych:** dane pogodowe z norweskiej stacji meteorologicznej w Linnedalen na Antarktydzie (ok. 16 000 rekordów). **Atrybuty:** opady [mm], prędkość wiatru [m/s], temperatura powietrza [ºC], ciśnienie [hPa]…
   - Około połowy pomiarów ma małą prędkość wiatru (0,82).
   - Prawie wcale pomiarów ma wysokie ciśnienie i średnie opady (0,80).
   - Około połowy pomiarów ma średnie opady i małą prędkość wiatru (0,69).

2. **Baza danych:** dane GUS-u dotyczące liczby podmiotów gospodarczych w poszczególnych powiatach w Polsce. **Atrybuty:** osoby fizyczne prowadzące działalność, spółki handlowe, fundacje, spółdzielnie…
   - W I półroczu 2006 roku bardzo mała liczba powiatów miała zarejestrowaną dużą liczbę osób fizycznych prowadzących działalność gospodarczą (0,53).
   - W I półroczu 2005 roku w więcej niż połowie powiatów była zarejestrowana mała liczba fundacji i duża liczba stowarzyszeń (0,73).

3. **Baza danych:** dane dotyczące różnych gałęzi przemysłu w Stanach Zjednoczonych. **Atrybuty:** zyski firmy, liczba zatrudnionych pracowników, roboczo-godziny, płace, zużycie energii…
   - Bardzo wiele firm, które mają bardzo duże wydatki na pensje, ma wielki kapitał zakładowy i ogromne zużycie energii (0,91).
   - Prawie w ogóle firm ma wielkie inwestycje (0,94).

4. **Baza danych:** dane dotyczące wypadków samochodowych na terenie Stanów Zjednoczonych. **Atrybuty:** wiek kierowcy, prędkość pojazdu, trzeźwość, godzina, miesiąc wypadku, liczba ofiar, rok produkcji pojazdu…
   - Wiele wypadków przy dużej prędkości miało miejsce pod koniec roku (0,66).
   - Prawie wszystkie wypadki spowodowane przez kierowców w średnim wieku miały małą liczbę ofiar (0,74).
   - Prawie wszystkie wypadki przy bardzo dużej prędkości mieli kierowcy w młodym wieku (0,50).
   - Około połowy wypadków miało miejsce w godzinach wieczornych (0,71).

5. **Baza danych:** dane dotyczące trzęsień ziemi na terenie Stanów Zjednoczonych. **Atrybuty:** rok, siła trzęsienia, prognoza…
   - Około 1/10 trzęsień ziemi w USA w 2003 roku miała dużą siłę (0,60).

- Około 1/2 trzęsień ziemi w USA w 2003 roku miała małą siłę (0,73).
- Około 3/4 trzęsień ziemi w USA w 2004 roku miało małą siłę i zostało wcześniej przewidzianych (0,96).

Jak widać, zdania wygenerowane przez automat wymagają korekty gramatyczno-składniowej.

Podsumowanie

Automaty podsumowań lingwistycznych w wielu branżach z powodzeniem mogą wspomóc człowieka w procesie obróbki danych z dużych baz danych. Potrafią generować komunikaty w języku naturalnym, wychodząc naprzeciw oczekiwaniom człowieka. Język naturalny jest niezrozumiały dla maszyny, jednakże wykorzystując zaawansowany aparat matematyczny, możemy nauczyć ją pewnych reguł tworzenia zdań i odpowiadania na stawiane przez nas pytania. Podsumowania lingwistyczne są obecnie wykorzystywane jako źródło wiadomości sygnalizujących i szacujących pewne zjawiska (procesy), i w wielu przypadkach są wystarczające dla przeciętnego użytkownika.

BIBLIOGRAFIA

Kacprzyk J., Yager R. (2001). *Linguistic summaries of data Rusing fuzzy logic*. „Int. J. of General Systems" 30, s. 133–154.

Kacprzyk J., Yager R., Zadrożny S. (2000). *A fuzzy logic based approach to linguistic summaries of databases*. „Int. J. of Appl. Math. And Comp. Sci." 10.

Niewiadomski A. (2006b). *A type-2 fuzzy approach to linguistic summarization of data*. IEEE Trans. On Fuzzy Systems.

Niewiadomski A. (2006b). *News Generating via Fuzzy Summarization of Databases*. „Lecture Notes In Computer Science" 3831, s. 419–429.

Niewiadomski A., Szczepaniak P. (2006). *News generating based on Interval Tye-2 Linguistic Summaries of Databases. Proceedings of IPMU 2006 Conference*, pp. 1324–1331. Paris.

Zadeh L.A. (1965). *Fuzzy sets*. „Inf. and Control" 8.