Izabela Superson
Adam Niewiadomski
Politechnika Łódzka

POZYSKIWANIE WIEDZY Z RELACYJNYCH BAZ DANYCH: WIELOPODMIOTOWE PODSUMOWANIA LINGWISTYCZNE

Wprowadzenie

Obecnie trudno wyobrazić sobie jakąkolwiek instytucję, przedsiębiorstwo lub placówkę społeczną, która byłaby w stanie sprawnie funkcjonować bez obszernej bazy skomputeryzowanych danych. Niesie to za sobą konieczność analizy oraz przetwarzania dużych zbiorów danych. Ponadto oczekujemy szybkich wyników, przedstawionych w maksymalnie skompresowany, zwięzły sposób, zrozumiały dla szerokiego grona odbiorców. Z pomocą przychodzą *podsumowania lingwistyczne relacyjnych baz danych*, czyli zdania w języku naturalnym opisujące znaczenie wybranych danych, np. *Około połowa [badanych] dzieci to chłopcy. Większość dziewczynek w wieku wczesnoszkolnym, w odniesieniu do chłopców, jest wysokiego wzrostu.* Zastosowanie podsumowań lingwistycznych może znacząco usprawnić proces zarządzania wiedzą. Za pomocą odpowiednich algorytmów można pozyskać wiedzę na temat zbioru danych w postaci intuicyjnego komunikatu w języku naturalnym. Dzięki zastosowaniu nieprecyzyjnych wyrażeń liczbowych, takich jak *większość, wysoki wzrost, wiek wczesnoszkolny*, otrzymuje się komunikat bardziej przyjazny i zrozumiały, bez konieczności posiadania dodatkowej wiedzy na temat analizowanych danych, ponieważ jest wyrażony językiem naturalnym, a nie liczbami, przez co staje się on komunikatywny i czytelny dla statystycznego odbiorcy. Jeżeli w powyższym przykładzie użytoby precyzyjnych liczb, np. *5679 dziewczynek w wieku od 7 do 12 lat, w odniesieniu do chłopców, jest wzrostu od 153 do 165 cm*, komunikat mógłby stać się niejasny dla osoby nieposiadającej wiedzy na temat analizowanych danych. Jak łatwo zauważyć, taki komunikat nie daje żadnej praktycznej wiedzy na temat danych użytkownikowi, który nie wie ile jest dziewczynek w analizowanej bazie (nie można określić jaką część zbioru dziewczynek stanowi liczba 5679), jaką część

stanowią dzieci w wieku od 7 do 12 lat oraz czy przedział 153-165 cm traktować jako *średni* wzrost, czy może już *wysoki*. Niewymagane są żadne operacje wstępne, takie jak np. sprawdzenie podstawowej wiedzy na temat danych, a zatem z metody tej mogą korzystać wszyscy użytkownicy, również nieposiadający wiedzy z zakresu informatyki, a jedynie umiejętność obsługi komputera. Zastosowanie tej metody skutkuje wzrostem jakości otrzymywanej wiedzy, ponieważ eliminuje błędy ludzkie oraz wynikające z zastosowania bardzo skomplikowanych i złożonych algorytmów, przy jednoczesnym skróceniu czasu, jaki byłby potrzebny na analizę obszernej bazy danych innymi metodami. Jest to krok w kierunku interfejsów naturalnych, przyjaznych użytkownikowi, gdyż opartych na języku naturalnym.

Celem pracy jest przedstawienie możliwości analizy dużych zbiorów danych za pomocą logiki rozmytej oraz zaprezentowanie wyników tej analizy za pomocą języka naturalnego.

1. Podsumowania lingwistyczne relacyjnych baz danych: przegląd literatury

Ponad trzydzieści lat temu R.R. Yager zaproponował koncepcję podsumowań lingwistycznych (relacyjnych) baz danych[1], np. p*onad połowa koszykarzy jest bardzo wysoka.* Była to odpowiedź na potrzebę szybkiego interpretowania informacji i pozyskiwania wiedzy z dużych zbiorów danych. Głównym atutem tej metody jest to, że pozyskana wiedza jest zaprezentowana w formie przyjaznej dla statystycznego użytkownika systemów komputerowych. Nie odnosi się ona do żadnej ze statystycznych metod agregacji danych (średnia, wariancja, odchylenie standardowe itp.), lecz – zamiast tego – opiera się na rozmytych modelach wyrażeń w języku naturalnym. Nawet jeżeli takie wyrażenia są mniej precyzyjne niż liczby, np. *ponad połowa obiektów* zamiast *55,6% obiektów* lub *bardzo wysoki chłopiec* zamiast *chłopiec o wzroście 195 cm*, to są one popularnie stosowane i dostarczają prostą w odbiorze wiedzę na temat podsumowywanych danych.

Koncepcja lingwistycznych podsumowań baz danych opiera się na rachunku Zadeha dotyczącym wyrażeń kwantyfikowanych lingwistycznie. Istnieją dwie, podstawowe formy podsumowań lingwistycznych (opierające się na dwóch formach wyrażeń kwantyfikowanych lingwistycznie), zaprezentowano je m.in. w pracach[2]:

---

[1] R.R. Yager, *A New Approach to the Summarization of Data*, „Information Science" 1982, 28, s. 69-86.

[2] J. Kacprzyk, R.R. Yager, *Linguistic Summaries of Data Using Fuzzy Logic*, „International Journal of General Systems" 2001, 30, s. 133-154; J. Kacprzyk, R.R. Yager, S. Zadrożny, *A Fuzzy Logic Based Approach to Linguistic Summaries of Databases*, „International Journal of Applied Mathematics and Computer Science" 2000, 10, s. 813-834; J. Kacprzyk, R.R. Yager,

$$Q\ P\ jest/są\ S\ [T] \quad (1)$$

Np. *Wielu chłopców jest wysokich [0.83],* i

$$Q\ P\ będących\ W\ jest/są\ S\ [T] \quad (2)$$

Np. *Wielu chłopców będących nastolatkami, jest wysokich [0.63]*.

W obu formach (1) i (2) *Q* jest kwantyfikatorem lingwistycznym, np. *Znacznie więcej niż 900*, reprezentowanym przez operator agregacji, np. kwantyfikator rozmyty (zbiór rozmyty o odpowiednich właściwościach) lub operator OWA[3], *P* jest podmiotem podsumowania, np. mężczyźni, samochody lub jakiekolwiek inne obiekty opisane w podsumowywanej bazie danych, *S* jest sumaryzatorem – wyrażeniem lingwistycznym dotyczącym właściwości obiektów, reprezentowanym przez zbiór rozmyty. Symbol *W*, pojawiający się jedynie w formie (2), jest kwalifikatorem, reprezentowanym przez zbiór rozmyty, który reprezentuje dodatkowe właściwości obiektów biorących udział w podsumowaniu. *T* [0, 1] jest stopnień prawdziwości i wyznacza prawdziwość podsumowania (jak bardzo jest bliskie prawdzie). Wartości *T* są wyznaczane na podstawie rachunku Zadeha dotyczącego wyrażeń kwantyfikowanych lingwistycznie oraz innych metod opisanych w pracach[4]. Niniejsza praca jest za krótka żeby móc opisać wszystkie z istniejących metod oraz aplikacji dotyczących podsumowań lingwistycznych relacyjnych baz danych, przykłady takich metod lub aplikacji można znaleźć np. w pracach[5]. Ponadto, nie jesteśmy w stanie zaprezentować wszystkich koncepcji dotyczących podsumowań danych opartych na zbiorach rozmytych, ale przyjmujących inne założenia niż Yager, np.[6].

S. Zadrożny, *Fuzzy Linguistic Summaries of Databases for an Efficient Business Data Analysis and Decision Support* [w:] *Knowledge Discovery for Business Information Systems*, eds. W. Abramowicz, J. Zurada, Kluwer Academic Publisher, Boston 2001, s. 129-152; J. Kacprzyk, S. Zadrożny, *Flexible Querying Using Fuzzy Logic: An Implementation for Microsoft Access* [w:] *Flexible Query Answering Systems*, eds. T. Andreasen, H. Christiansen, H.L. Larsen, Kluwer, Boston 1997, s. 247-275.

[3] R.R. Yager, *On Ordered Weighted Averaging Operators in Multicriteria Decision Making*, „IEEE Transactions on Systems, Man, and Cybernetics" 1988, 18, s. 183-190.

[4] J. Kacprzyk , R.R. Yager, *Linguistic Summaries of Data Using Fuzzy Logic*, op. cit.; A. Niewiadomski, *News Generating via Fuzzy Summarization of Databases*, „Lecture Notes in Computer Science" 2006, 3831, s. 419-429.

[5] *Flexible Query Answering System*, eds. T. Andreasen, H. Christiansen, H.L. Larsen, Kluwer, Boston 1997, s. 247-275; A. Niewiadomski, *News Generating*…, op. cit.

[6] P. Bosc, O. Pivert, *Fuzzy Querying in Conventional* Databases [w:] *Fuzzy Logic for the Management of Uncertainty*, eds. L.A. Zadeh, J. Kacprzyk, Wiley, New York 1992, s. 645-671; A. Niewiadomski, *Six New Informativeness Indices of Data Linguistic Summaries* [w:] *Advances in Intelligent Web Mastering*, eds. P.S. Szczepaniak, K. Węgrzyn-Wolska, Springer-Verlag, 2007, s. 254-259; G. Raschia, N. Mouaddib, *SAINTETIQ: A Fuzzy Set-Based Approach to Database Summarization*, „Fuzzy Sets and Systems" 2002, 129, s.137-162; D. Rasmussen, R.R. Yager, *A fuzzy SQL Summary Language for Data Discovery* [w:] *Fuzzy Information Engineering: A Guided Tour of Application's*, eds. D. Dubois, H. Prade, R.R. Yager, Wiley, New York 1997, s. 253-264.

Najważniejszym elementem zaprezentowanym w niniejszej pracy są wielopodmiotowe podsumowania lingwistyczne relacyjnych baz danych, np. *Większość chłopców w odniesieniu do dziewczynek jest wysokiego wzrostu.* Oznacza to, że podsumowania będą dotyczyć więcej niż jednego podmiotu P_1, np. P_1 i P_2, natomiast modele lingwistycznych wyrażeń nieprecyzyjnych (dla sumaryzatorów, kwantyfikatorów itp.) są utworzone za pomocą zbiorów rozmytych. Jest to istotne rozszerzenie istniejących koncepcji podsumowywania baz danych, które – jak dotąd – umożliwiały podsumowywanie danych na podstawie tylko jedego podmiotu.

Dalsza część pracy jest zorganizowana następująco: w Sekcji 2 został przedstawiony pomysł wielopodmiotowych podsumowań lingwistycznych relacyjnych baz danych. Skonstruowano i wyznaczono podsumowania odnoszące się do więcej niż jednego podmiotu P reprezentowanego przez krotki w podsumowywanej bazie danych D, np. P_1 *i* P_2 lub P_1 w odniesieniu do P_2. Podmioty te są reprezentowane przez zbiory krotek zgromadzone w oddzielnych tabelach w bazie D lub zbiory wyznaczone za pomocą selekcji, filtrowania krotek itp., z uwagi na pewien atrybut, np. kobiety i mężczyźni. Sekcja 3 zawiera opis eksperymentu z użyciem stworzonej w tym celu aplikacji, co pomoże nam zaprezentować oraz wyznaczyć użyteczność oraz wydajność wielopodmiotowych podsumowań lingwistycznych relacyjnych baz danych. Zaprezentujemy przykładowy wynik działania aplikacji dla użytej bazy danych oraz w jaki sposób niezaawansowani technicznie użytkownicy mogą korzystać i wpływać na podsumowania generowane przez program.

2. Wielopodmiotowość w relacyjnych bazach danych oraz wielopodmiotowe podsumowania lingwistyczne

2.1. Relacyjne bazy danych oraz wielopodmiotowość

Niniejszy podpunkt systematyzuje oznaczenia i przybliża pojęcia dotyczące relacyjnych baz danych, opartych na podejściu klasycznym, czyli Codda (1970). Wprowadza oznaczenia, jakie będą stosowane w dalszej części pracy. Założono, że baza zawierająca dane, które mają zostać podsumowane składa się z tabel, będących zbiorami krotek (zwykle nazywanych rekordami), a jedna krotka reprezentuje dokładnie jeden obiekt (np. dziecko, osobę, samochód itp.). Taki zbiór oznaczono jako $Y = \{y_1, \ldots, y_m\}$. Tabela D' istniejąca w bazie danych $\in$ składa się z krotek d_i, $i = 1, 2, \ldots, m$, które stanowią wiersze tabeli: $D' = \{d_1, \ldots, d_m\}^T$, $m \in N$ jest liczbą krotek w tabeli D'. Każda krotka d_i składa się z n N wartości atrybutów $V_1,\ldots, V_n$ posiadające odpowiednio dziedziny $X_1,\ldots, X_n$. Wartości atrybutów

wyrażają właściwości obiektu, np. wzrost, wypłatę, koszt itp. oraz są traktowane jako kolumny tabeli. Dziedziny atrybutów są zbiorami wartości, jakie może przyjąć dany atrybut, np. zbiór X_1 = [50, 200] może być dziedziną V_1 = „wzrost osoby w centymetrach". Wartość atrybutu V_j dla obiektu y_i jest oznaczona jako $V_j(y_i)$ X_j, $i \in \{1, 2, \ldots, m\}$, $j \in \{1, 2, \ldots, n\}$. Baza danych D zawierająca informacje o elementach z $Y = \{y1, \ldots, ym\}$ posiada następującą postać:

$$D = \begin{Bmatrix} \langle V_1(y_1), & \cdots, & V_n(y_1)\rangle \\ \vdots & & \vdots \\ \langle V_1(y_m), & \cdots, & V_n(y_m)\rangle \end{Bmatrix} = \begin{Bmatrix} d_1 \\ \vdots \\ d_m \end{Bmatrix} \tag{3}$$

Przykładowa baza danych w formie (3) jest przedstawiona w tab. 1. Jest to część większej bazy danych, biorąca udział w podsumowaniu w eksperymencie opisanym w punkcie 2. Tabela prezentuje również możliwość ekstrakcji dwóch zbiorów podmiotów dla podsumowań wielopodmiotowych. W tym przypadku atrybutem służącym do ekstrakcji jest atrybut „Płeć", który umożliwia podział zbioru danych na dwa podzbiory przedstawione w tab. 2 i 3.

Tabela 1

Przykładowa baza danych *D* zawierająca dane dotyczące dzieci w wieku szkolnym

ID	Płeć	Wiek	Wzrost
1.	Dziewczynka	7	130
2.	Chłopiec	8	120
3.	Chłopiec	13	150
4.	Dziewczynka	8	140
5.	Dziewczynka	18	160

Tabela 2

Podzbiór bazy danych *D* zaprezentowanej w tab. 1, utworzony przez wybór krotek reprezentujących chłopców

ID	Płeć	Wiek	Wzrost
2.	Chłopiec	8	120
3.	Chłopiec	13	150

Tabela 3

Podzbiór bazy danych *D* zaprezentowanej w tab. 1, utworzony przez wybór krotek reprezentujących dziewczynki

ID	Płeć	Wiek	Wzrost
1.	Dziewczynka	7	130
4.	Dziewczynka	8	140
5.	Dziewczynka	18	160

Należy podkreślić, że tab. 2 i 3 nie reprezentują rzeczywistych tabel (w ujęciu technicznym jako rozdzielonych zbiorów rekordów) zawartych w systemie zarządzania bazą danych. Utworzenie oraz przechowywanie tych tabel w systemie mogłoby okazać się nieefektywne, zwłaszcza w odniesieniu do tzw. postaci normalnych tabel w relacyjnej bazie danych, stanowiących popularne kryterium optymalizacji baz danych. Przedstawione tabele są jedynie rezultatem operacji filtrowania wykonanej na tabeli *D'* (reprezentowanej przez tab. 1) przy użyciu wybranego atrybutu, w tym przypadku będzie to atrybut „Płeć", który umożliwi podział obiektów na „chłopców" i „dziewczynki".

Szczególnie istotne z punktu widzenia niniejszej pracy jest wyznaczenie dwóch, oddzielnych podzbiorów obiektów, przechowywanych wcześniej jako jeden w bazie *D*. Podzbiory reprezentują różne podmioty P_1 oraz P_2, użyte w wielopodmiotowych podsumowaniach lingwistycznych zaprezentowanych w kolejnym podpunkcie.

2.2. Nowe formy podsumowań: wielopodmiotowe podsumowania lingwistyczne

Pierwsza z zaproponowanych form podsumowań ma postać:

$$Q\ P_1\ w\ odniesieniu\ do\ P_2\ jest\ S_1 \tag{4}$$

gdzie Q jest kwantyfikatorem rozmytym, P_1 i P_2 są podmiotami podsumowania i S_1 jest sumaryzatorem reprezentowanym przez zbiór rozmyty. Stopień prawdziwości podsumowania w formie (4) jest obliczany za pomocą wzoru (5):

$$T(Q\ P_1\ w\ odniesieniu\ do\ P_2\ jest\ S_1) = \mu_Q\left(\frac{\frac{1}{M_{P_1}}\Sigma - count(S_{1P_1})}{\frac{1}{M_{P_1}}\Sigma - count(S_{1P_1}) + \frac{1}{M_{P_2}}\Sigma - count(S_{1P_2})}\right) \tag{5}$$

gdzie:

$$\Sigma - count\left(S_{1P_1}\right) = \sum_{i=1}^{m}\{u_{S_1}(d_i): d_i \in^* P_1\} \tag{6}$$

Analogicznie:

$\Sigma - count\left(S_{1P_2}\right)$. Notacja $d_i \in^* P_1$ oznacza, że krotka d_i jest obiektem reprezentującym podmiot P_1. M_{P_1} oraz M_{P_2} są liczbami krotek reprezentujących odpowiednio podmioty P_1 oraz P_2:

$$M_{P_1} = \sum_{i=1}^{m} t_i \tag{7}$$

gdzie t_i:

$$t_{i_{P_1}} = \begin{cases} 1, & if\ d_i \in^* P_1 \\ 0, & w\ innym\ przypadku \end{cases} \tag{8}$$

Przykład:

$$t_{i_{chłopcy}} = \begin{cases} 1, & if\ V_j(d_i) = "chłopiec" \\ 0, & if\ V_j(d_i) = "dziewczynka" \end{cases} \tag{9}$$

V_j = *Płeć*. Przykład podsumowania w formie (4):

Większość chłopców w odniesieniu do dziewczynek jest

wysokiego wzrostu [0.456] (10)

gdzie Q = *Większość*, P_1 = chłopcy, P_2 = dziewczynki, S_1 = *wysoki wzrost.*

Druga forma wielopodmiotowych podsumowań lingwistycznych ma następującą postać:

$$Q\ P_1\ w\ odniesieniu\ do\ P_2\ będących\ S_2\ jest\ S_1, \tag{11}$$

gdzie Q jest kwantyfikatorem relatywnym, P_1 i P_2 są podmiotami podsumowania, S_2 jest kwalifikatorem odnoszącym się do obu podmiotów P_1 i P_2, natomiast S_1 jest sumaryzatorem. Stopień prawdziwości podsumowania jest obliczany za pomocą wzoru (12):

$$T(Q\ P_1\ w\ odniesieniu\ do\ P_2\ będących\ S_2\ jest\ S_1) = \mu_Q\left(\frac{\frac{1}{M_{P_1}}\Sigma - count(S_{1P_1} \cap S_{2P_1})}{\frac{1}{M_{P_1}}\Sigma - count(S_{2P_1}) + \frac{1}{M_{P_2}}\Sigma - count(S_{1P_2})}\right) \tag{12}$$

gdzie:

$$\sum -count(S_{1_{P_1}} \cap S_{2_{P_1}}) = \sum_{i=1}^{m} \min\{\mu_{S_1}(d_i), \mu_{S_2}(d_i)\}, d_i \in^* P_1 \quad (13)$$

Wzory oraz oznaczenia $\sum -count(S_{2_{P_1}})$, $\sum -count(S_{2_{P_2}})$, $d_i \in^* P_1$ są analogiczne jak dla formy (4). Przykład podsumowania w formie (11):

Około dwóch trzecich chłopców w odniesieniu do dziewczynek będących nastolatkami, jest wysokiego wzrostu [0.39] (14)

gdzie Q = *około dwóch trzecich*, P_1 = *chłopcy*, P_2 = *dziewczynki*, S_1 = *wysoki wzrost*, S_2 = *nastoletni wiek.*

Podsumowania w formie (11) umożliwiają otrzymywanie informacji dotyczących wybranych cech S_1 podmiotów, w zależności od warunków, jakie oba podmioty muszą spełniać (cechy, które muszą posiadać oba podmioty). W tym przypadku krotki, które będą brane pod uwagę podczas analizy muszą reprezentować chłopców i dziewczynki w wieku nastoletnim, o czym decyduje kwalifikator S_2.

Trzecia z zaproponowanych form ma następującą postać:

$$Q\ P_1\ \textit{będących}\ S_2\ \textit{w odniesieniu do}\ P_2\ \textit{jest}\ S_1 \quad (15)$$

Stopień prawdziwości formy (15) jest podany wzorem:

$$T(Q\ P_1\ \textit{będących}\ S_2\ \textit{w odniesieniu do}\ P_2\ \textit{jest}\ S_1) = \mu_Q\left(\frac{\frac{1}{M_{P_1}}\sum - count(S_{1_{P_1}} \cap S_{2_{P_1}})}{\frac{1}{M_{P_1}}\sum - count(S_{1_{P_1}}) + \frac{1}{M_{P_2}}\sum - count(S_{1_{P_2}})}\right), \quad (16)$$

gdzie S_2 jest kwalifikatorem odnoszącym się jedynie do podmiotu P_1.

Przykład takiego podsumowania:

Około połowa chłopców będących nastolatkami

w odniesieniu do dziewczynek, jest wysokiego wzrostu [0.256], (17)

gdzie Q = *około połowa,* P_1 = *chłopcy,* P_2 =*dziewczynki,* S_1 = *wysoki wzrost,* S_2 = *nastoletni wiek.*

Podsumowania w formie (15) umożliwiają generowanie informacji dotyczących wybranych cech podmiotów w zależności od posiadanych właściwości podmiotu P_1. Oznacza to, że krotki biorące udział w podsumowaniu reprezentują podmioty P_1 *i* P_2, ale jedynie podmiot P_1 musi posiadać cechy narzucone przez kwalifikator.

Czwartą formą spośród zaproponowanych form podsumowań wielopodmiotowych jest:

$$Więcej\ P_1\ niż\ P_2\ jest\ S_1 \tag{18}$$

Jak można zauważyć, forma (18) nie wymaga zastosowania dodatkowego kwantyfikatora. Stopień prawdziwości jest podany za pomocą wzoru (19):

$$T(Więcej\ P_1\ niż\ P_2\ \ jest\ S_1) = \frac{\sum_{i=1}^{m}\mu_{S_1}(d_{iP_1})}{\sum_{i=1}^{m}\mu_{S_1}(d_{iP_1}) + \sum_{i=1}^{m}\mu_{S_1}(d_{iP_2})}, \tag{19}$$

gdzie P_1 i P_2 są podmiotami podsumowania, M_{P_1} i M_{P_2} są liczbą krotek reprezentujących odpowiednio podmioty P_1 i P_2,

$d_{iP_1}: d_i \in^{*} P_1 \wedge d_{i2}: d_i \in^{*} P_2$. Przykład podsumowania w formie (19):

Więcej chłopców niż dziewczynek jest wysokiego wzrostu [0.756] (20)

gdzie P_1 = *chłopcy,* P_2 = *dziewczynki,* S_1 = *wysoki wzrost.*

Podsumowania w formie (18) umożliwiają użytkownikowi porównywanie dwóch podmiotów bez konieczności użycia dodatkowych miar lub modeli rozmytych, np. kwantyfikatorów. Takie podejście umożliwia szybkie generowanie podsumowań, których treść jest bardzo intuicyjna.

2.3. Różnice pomiędzy klasycznymi formami podsumowań a formami wielopodmiotowymi

Należy zauważyć, że żadna z klasycznych form podsumowań nie pozwala na porównywanie dwóch, różnych podmiotów, pod względem posiadanych przez nie cech, np. chłopcy i dziewczynki i ich wzrost, wiek itp. Z drugiej strony, takie relacje mogą być przedstawione w łatwy i czytelny sposób za pomocą podsumowań wielopodmiotowych. Dla klasycznych podsumowań jedyną możliwością jest zastosowanie jako kwalifikator W wyodrębnionego podzbioru obiektów, np. chłopcy lub dziewczynki (wzór (2)), np. *Około połowa CHŁOPCÓW jest wysokiego wzrostu*, gdzie W = *CHŁOPCY*.

3. Lingwistyczne opisywanie oraz podsumowywanie baz danych za pomocą podsumowań wielopodmiotowych: przykład zastosowania

3.1. Cele oraz metody aplikacji

Aplikacja utworzona w celu testowania nowych form podsumowań, została napisana z użyciem języka Java w wersji 1.7. Baza danych użyta w eksperymencie zawiera dane dotyczące dzieci w wieku od 7 do 18 roku życia. Dane zawierają m.in. wzrost, wagę, datę urodzenia, warunki w jakich żyją, takie jak liczba pomieszczeń w mieszkaniu, liczba osób w rodzinie, sytuacja finansowa itp. Baza zawiera dane dotyczące 13 956 dzieci, w tym 6 991 chłopców oraz 6 965 dziewczynek.

Podsumowania generowane w ramach eksperymentu pokazują zależność wzrostu od wieku i płci dziecka. Podmioty biorące udział w podsumowaniach to chłopcy i dziewczynki. Proces logicznego podziału danych na dwa podzbiory jest widoczny w tab. 1-3. Kwantyfikatory relatywne zastosowane w podsumowaniach to: *większość, około dwóch trzecich, około połowy*. Propozycja funkcji przynależności zastosowanych dla kwantyfikatorów *większość* oraz *około dwóch trzecich* zaprezentowano na rys. 1-2.

Wygenerowane podsumowania opierają się na kwalifikatorach i sumaryzatorach reprezentowanych przez zbiory rozmyte. Przykładowe sumaryzatory użyte w podsumowaniach:

- wysoki (wzrost)
- niski (wzrost)
- wczesnoszkolny (wiek)
- nastoletni (wiek)

Etykieta *wysoki (wzrost)* jest reprezentowana przez zbiór rozmyty *WYSOKI_WZROST:*

$$WYSOKI\ WZROST = \{(x, \mu_{WYSOKI\ WZROST}(x)) : x \in [150, 195], \mu_{WYSOKI\ WZROST}(x) \in [0,1]\}, \quad (21)$$

gdzie:

$$\mu_{WYSOKI\ WZROST}(x) = \begin{cases} \frac{2(x-150)}{45}, & if\ 150 \leq x \leq \frac{150+195}{2} \\ \frac{2(195-x)}{45}, & if\ \frac{150+195}{2} \leq x \leq 195 \\ 0, & if\ x \leq 150\ lub\ x \geq 195 \end{cases} \quad (22)$$

Etykieta *niski (wzrost)* jest reprezentowana przez zbiór rozmyty *NISKI WZROST:*

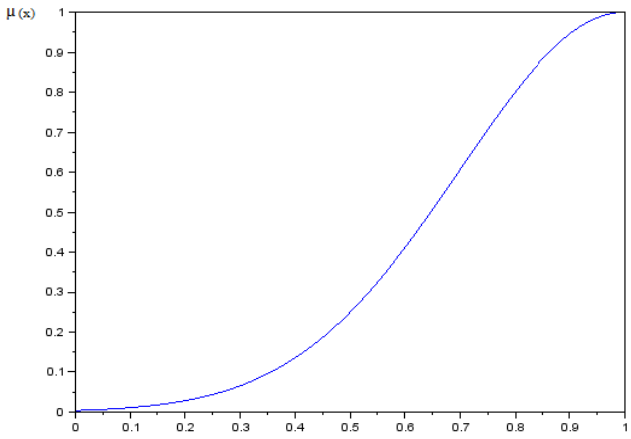
$$NISKI\ WZROST = \{(x, \mu_{NISKI\ WZROST}(x)) : x \in [103, 150], \mu_{NISKI\ WZROST}(x) \in [0,1]\}, \quad (23)$$

gdzie:

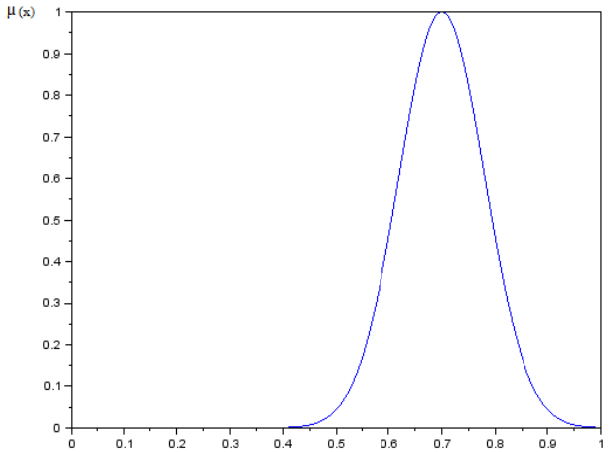
$$\mu_{NISKI\ WZROST}(x) = \begin{cases} \frac{2(x-103)}{47}, & if\ 103 \leq x \leq \frac{103+150}{2} \\ \frac{2(150-x)}{47}, & if\ \frac{103+150}{2} \leq x \leq 150 \\ 0, & if\ x \leq 103\ lub\ x \geq 150 \end{cases} \quad (24)$$

Analogicznie, etykieta *nastoletni (wiek)* jest reprezentowana przez zbiór:

$$NASTOLETNI\ WIEK = \{(x, \mu_{NASTOLETNI\ WIEK}(x)) : x \in [13, 18], \mu_{NASTOLETNI\ WIEK}(x) \in [0,1]\}, \quad (25)$$



Rys. 1. Funkcja przynależności kwantyfikatora *WIĘKSZOŚĆ*



Rys. 2. Funkcja przynależności kwantyfikatora *OKOŁO DWÓCH TRZECI*

gdzie:

$$\mu_{NASTOLETNI\ WIEK}(x) = \begin{cases} \frac{2(x-13)}{5}, & if\ 13 \leq x \leq \frac{13+18}{2} \\ \frac{2(18-x)}{5}, & if\ \frac{13+18}{2} \leq x \leq 18 \\ 0, & if\ x \leq 13\ lub\ x \geq 18 \end{cases} \quad (26)$$

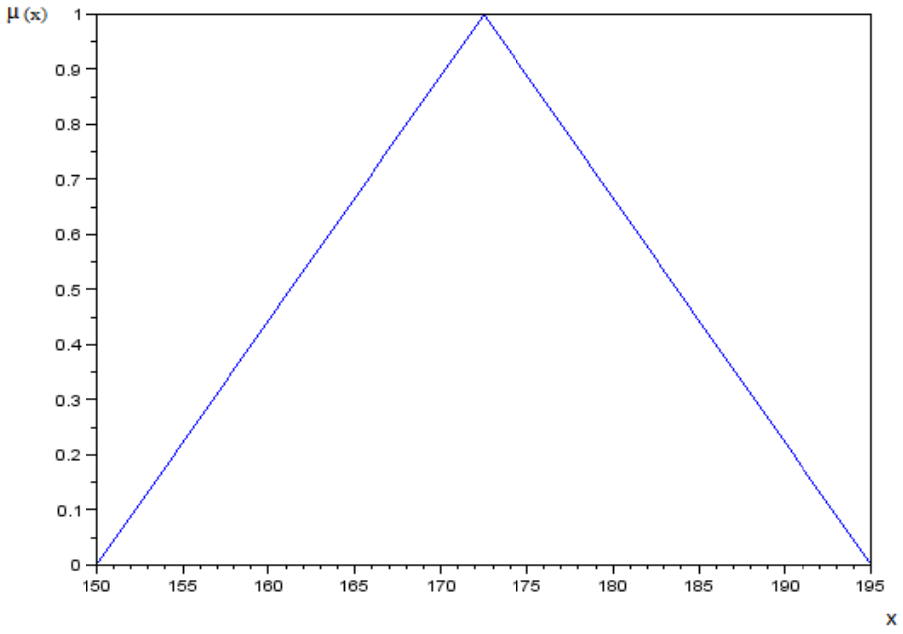
Etykieta *wczesnoszkolny (wiek)* jest reprezentowana przez zbiór:

$$WCZESNOSZKOLNY\ WIEK = \{(x, \mu_{WCZESNOSZKOLNY\ WIEK}(x)) : x \in [7,12], \mu_{WCZESNOSZKOLNY\ WIEK}(x) \in [0,1]\}, \quad (27)$$

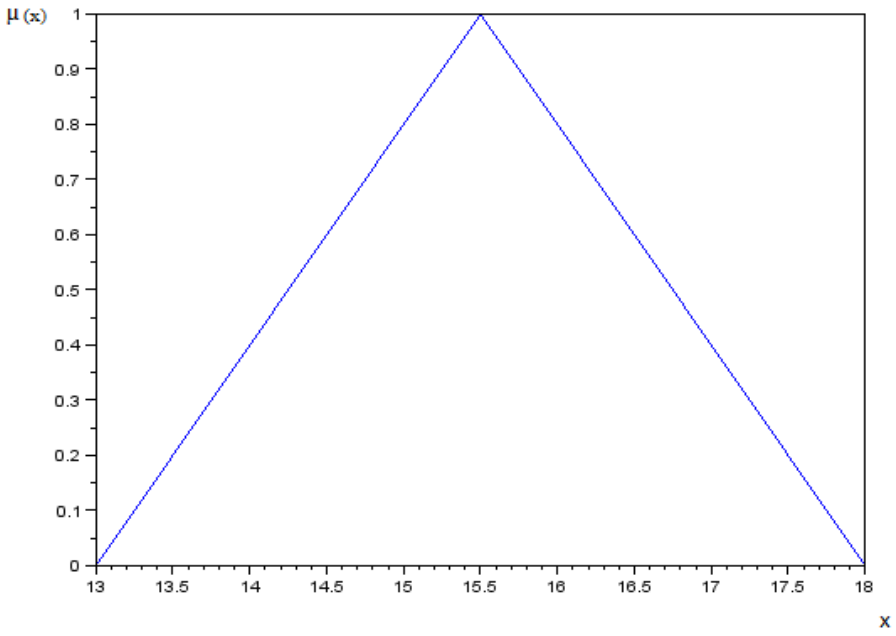
gdzie:

$$\mu_{WCZESNOSZKOLNY\ WIEK}(x) = \begin{cases} \frac{2(x-7)}{5}, & if\ 7 \leq x \leq \frac{7+12}{2} \\ \frac{2(12-x)}{5}, & if\ \frac{7+12}{2} \leq x \leq 12 \\ 0, & if\ x \leq 7\ lub\ x \geq 12 \end{cases} \quad (28)$$

Wykresy funkcji przynależności dla zbiorów *WYSOKI WZROST* oraz *NASTOLETNI WIEK* są widoczne na rys. 3-4.



Rys. 3. Funkcja przynależności etykiety *WYSOKI WZROST*



Rys. 4. Funkcja przynależności etykiety *NASTOLETNI WIEK*

3.2. Otrzymane rezultaty oraz ich interpretacja

Wynik działania aplikacji został zaprezentowany w tab. 4. Dla każdego podsumowania obliczono i podano jego stopień prawdziwości (kolumna *T*) oraz zastosowaną formę podsumowania (kolumna „*Forma podsumowania*"), odnoszącą się do zaprezentowanych wzorów (4), (11), (15), (18) dotyczących pierwszej, drugiej, trzeciej oraz czwartej formy podsumowań wielopodmiotowych oraz (1) i (2) odnoszących się do klasycznych form podsumowań.

Zgodnie z opinią eksperta, wyniki są intuicyjnie poprawne. Pierwszych 8 podsumowań zostało zbudowanych za pomocą formy pierwszej dla nowych podsumowań wielopodmiotowych (wzór (4)). Analizując ich stopnie prawdziwości, można dojść do wniosku, że nie istnieją znaczne dysproporcje pomiędzy wielkością zbiorów chłopców i dziewcząt.

Kolejne podsumowania, od 9 do 16, świadczą o tym, że jest więcej wysokich dziewczynek niż chłopców w grupie dzieci w wieku wczesnoszkolnym, np. podsumowanie 9 zawiera przeciwne założenie (większość wysokich chłopców) i posiada bardzo niski stopień prawdziwości. Sytuacja wygląda inaczej wśród nastolatków: grupa wysokich chłopców jest większa niż wysokich dziewczynek (podsumowanie 10). Jednakże nie można powiedzieć, że w porównaniu do chłopców, większość nastoletnich dziewczynek jest niska, co jest zgodne z prawdą, ponieważ taka sytuacja oznaczałaby, że jest dużo nastoletnich dziew-

cząt o wzroście z przedziału od 103 cm do 150 cm (czytelnik musi mieć na uwadze, że dzieci w bazie danych posiadają wzrost z przedziału od 103 cm do 195 cm, zatem niskie dziecko w tym przypadku posiada wzrost z przedziału od 103 do 150).

Nr.	Podsumowanie	*T*	Forma podsumowania
1	2	3	4
1.	Większość dziewczynek w odniesieniu do chłopców jest w wieku wczesnoszkolnym	0.495	(4)
2.	Większość chłopców w odniesieniu do dziewczynek jest wieku wczesnoszkolnym	0.505	
3.	Większość dziewczynek w odniesieniu do chłopców jest w wieku nastoletnim	0.511	
4.	Około połowa chłopców w odniesieniu do dziewczynek jest w wieku nastoletnim	0.994	
5.	Większość dziewczynek w odniesieniu do chłopców jest wysokiego wzrostu	0.206	
6.	Większość chłopców w odniesieniu do dziewczynek jest wysokiego wzrostu	0.298	
7.	Większość dziewczynek w odniesieniu do chłopców jest niskiego wzrostu	0.249	
8.	Około dwie trzecie chłopców w odniesieniu do dziewczynek jest niskiego wzrostu	0.043	
9.	Większość chłopców w odniesieniu do dziewczynek, będących w wieku wczesnoszkolnym jest wysokiego wzrostu	0.004	(11)
10.	Większość chłopców w odniesieniu do dziewczynek, będących w wieku nastoletnim jest wysokiego wzrostu	0.129	
11.	Większość dziewczynek w odniesieniu do chłopców, będących w wieku wczesnoszkolnym jest niskiego wzrostu	0.124	
12.	Około połowa dziewczynek w odniesieniu do chłopców, będących w wieku nastoletnim jest niskiego wzrostu	0	
13.	Większość dziewczynek będących w wieku wczesnoszkolnym, w odniesieniu do chłopców jest niskiego wzrostu	0.101	(15)
14.	Większość dziewczynek będących w wieku nastoletnim, w odniesieniu do chłopców jest niskiego wzrostu	0.004	
15.	Większość chłopców będących w wieku nastoletnim, w odniesieniu do dziewczynek jest wysokiego wzrostu	0.098	
16.	Około dwie trzecie chłopców będących w wieku wczesnoszkolnym, w odniesieniu do dziewczynek jest wysokiego wzrostu	0	
17.	Więcej chłopców niż dziewczynek jest wysokiego wzrostu	0.534	(18)

1	2	3	4
18.	Więcej dziewczynek niż chłopców jest niskiego wzrostu	0.5	
19.	Więcej chłopców niż dziewczynek jest w wieku nastoletnim	0.49	
20.	Więcej dziewczynek niż chłopców jest w wieku nastoletnim	0.51	
21.	Więcej chłopców niż dziewczynek jest w wieku wczesnoszkolnym	0.506	
22.	Około połowa dzieci to dziewczynki	1	(1)
23.	Większość dzieci jest w wieku wczesnoszkolnym	0,32	
24.	Około dwie trzecie chłopców jest wysokiego wzrostu	0	
25.	Większość chłopców będących wysokiego wzrostu jest w wieku nastoletnim	0,031	(2)

Podsumowania od 17 do 20 potwierdzają brak znacznych dysproporcji pomiędzy liczbą wysokich chłopców i wysokich dziewcząt oraz nastoletnich chłopców i nastoletnich dziewcząt. Zgodnie z podsumowaniem 17 i 18, jest nieco więcej wysokich chłopców niż wysokich dziewczynek oraz odnosząc się do podsumowań 19 i 20 – nastoletnich dziewcząt w bazie jest kilka więcej niż nastoletnich chłopców. Podsumowanie 21 potwierdza, że nastoletnich dziewczynek jest nieco więcej (liczba chłopców w wieku wczesnoszkolnym jest nieco większa niż dziewczynek).

Wykorzystanie klasycznych form podsumowań, tj. (1) i (2), wzbogaca wyniki o dodatkowe informacje. Rozszerzenie tab. 4 o podsumowania od 22 do 25 uzupełnia informacje na temat analizowanego zbioru danych. Przykładem jest potwierdzenie braku dysproporcji pomiędzy zbiorami chłopców i dziewczynek. Dedykowany algorytm może w łatwy sposób wyznaczyć stopnie prawdziwości, wybrać najlepsze z nich (niosące najwięcej informacji) oraz zaprezentować je w jasnej i czytelnej formie, np. *Około połowa dzieci to dziewczynki. Większość chłopców w odniesieniu do dziewczynek jest wysokiego wzrostu. Około dwóch trzecich dziewczynek, będących w wieku wczesnoszkolnym, w odniesieniu do chłopców jest wysokiego wzrostu.* Ostatni wniosek pokazuje, że nowe formy podsumowań nie wykluczają klasycznych, ale mogą być stosowane równolegle w celu rozszerzenia oraz ulepszenia procesu ekstrakcji wiedzy dotyczącej dużych zbiorów danych.

Literatura

Bosc P., Pivert O., *Fuzzy Querying in Conventional* Databases [w:] *Fuzzy Logic for the Management of Uncertainty*, eds. L.A. Zadeh, J. Kacprzyk, Wiley, New York 1992.

Flexible Query Answering System, eds. T. Andreasen, H. Christiansen, H.L. Larsen, Kluwer, Boston 1997.

Kacprzyk J., Yager R.R., *Linguistic Summaries of Data Using Fuzzy Logic*, „International Journal of General Systems" 2001, 30.

Kacprzyk J., Yager R.R., Zadrożny S., *A Fuzzy Logic Based Approach to Linguistic Summaries of Databases*, „International Journal of Applied Mathematics and Computer science" 2000, 10.

Kacprzyk J., Yager R.R., Zadrożny S., *Fuzzy Linguistic Summaries of Databases for an Efficient Business Data Analysis and Decision Support* [w:]: *Knowledge Discovery for Business Information Systems*, eds. W. Abramowicz, J. Zurada, Kluwer Academic Publisher, Boston 2001.

Kacprzyk J., Zadrożny S., *Flexible Querying Using Fuzzy Logic: An Implementation for Microsoft Access* [w:] Flexible Query Answering Systems, eds. T. Andreasen, H. Christiansen, H.L. Larsen, Kluwer, Boston 1997.

Niewiadomski A., *News Generating via Fuzzy Summarization of Databases*, „Lecture Notes in Computer Science" 2006, 3831.

Niewiadomski A., *Six New Informativeness Indices of Data Linguistic Summaries* [w:] *Advances in Intelligent Web Mastering*, eds. P.S. Szczepaniak, K. Węgrzyn-Wolska, Springer-Verlag, 2007.

Raschia G., Mouaddib N., *SAINTETIQ: A Fuzzy Set-Based Approach to Database Summarization*, „Fuzzy Sets and Systems" 2002, 129.

Rasmussen D., Yager R.R., *A fuzzy SQL Summary Language for Data Discovery* [w:] *Fuzzy Information Engineering: A Guided Tour of Application's*, eds. D. Dubois, H. Prade, R.R. Yager, Wiley, New York 1997.

Yager R.R., *A New Approach to the Summarization of Data*, „Information Science" 1982, 28.

Yager R.R., *On Ordered Weighted Averaging Operators in Multicriteria Decision Making*, „IEEE Transactions on Systems, Man, and Cybernetics" 1988, 18.

ACQUIRING KNOWLEDGE FROM RELATIONAL DATABASES: MULTI-SUBJECT LINGUISTIC SUMMARIES

Summary

The aim of this article is to show how fuzzy logic based algorithms can be applied to analyze large datasets and present its results in a human-friendly form: using natural language. A new concept of linguistic summaries is demonstrated: multi-subject linguistic summaries of relational databases, that extends the classic manner. This paper focuses on new, interesting forms of linguistic summaries, which are represented by equations (4), (11), (15) and (18). This article also contains discussion about calculating degrees of truth of the new forms. From the potential end user point of view simplified form of presenting results using natural language expressions is the most important thing. This paper includes demonstration and description of standalone application that generates analysis of large dataset and presents results using short and intuitive expressions in natural language. Possibilities given by the multi-subject linguistic summaries, e.g. description and comparison of more than one subject in one summarization, makes them great extension and complementation of existing forms of linguistic summaries.