Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej Nr 21

XV Seminarium
ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE' 2005
Oddział Gdański PTETiS

ALGORYTM PORZĄDKOWANIA TABLIC WYNIKOWYCH INFORMACJI STATYSTYCZNYCH

Ewa WĘDROWSKA[1], Marcin FORKIEWICZ[2]

1. Katedra Metod Ilościowych, Wydział Nauk Ekonomicznych,
   Uniwersytet Warmińsko-Mazurski, ul. Oczapowskiego 2, 10-900 Olsztyn
   tel.: (89) 523-37-21 fax: (89) 523-36-85 e-mail: ewaf@uwm.edu.pl
2. Katedra Zarządzania i Technik Informatycznych, Wydział Zarządzania i Ekonomii,
   Politechnika Gdańska, ul. Narutowicza 11/12, 80-952 Gdańsk
   tel.: (58) 347-10-21 fax: (58) 348-60-24 e-mail: mfork@zie.pg.gda.pl

W artykule przedstawiono algorytm filtrowania danych służący do porządkowania tablic wynikowych. Celem artykułu jest zdefiniowanie miary ilości informacji, tak aby możliwe stało się wyselekcjonowanie takich tablic, które niosą największy ładunek informacyjny – największą ilość informacji. Autorzy skoncentrowali się na badaniu ilości informacji strukturalnej zawartej w tablicach statystycznych. Zadaniem proponowanej miary ilości informacji strukturalnej dostarczanej przez tablice wynikowe jest szeregowanie tych tablic pod względem dostarczanego przez nie ładunku informacyjnego.

1. WPROWADZENIE

Złożoność otoczenia wymusza opracowywanie odpowiednich metod zarządzania zasobami informacyjnymi. Również w wyniku szeregu badań statystycznych, a także sprawozdawczości statystycznej pojawia się duża liczba tablic wynikowych dostarczających różnorodnych treści informacyjnych. Zasadne staje się zatem poszukiwanie kryteriów (filtrów) pozwalających na skuteczne selekcjonowanie informacji ważnych i niosących dla odbiorcy istotne treści oraz odrzucania szumów. Należy poszukiwać obiektywnych kryteriów selekcji informacji, tak aby odbiorca miał do swej dyspozycji wystarczającą ilość informacji do podejmowania decyzji i organizowania swoich działań. Naturalnie, na pierwszym miejscu należy w tych poszukiwaniach identyfikować osobiste, subiektywne poglądy i postawy poszczególnych odbiorców informacji.

Kwestią słabo rozpoznaną w standardach dotyczących budowy i prezentacji zestawień tabelarycznych są zasady porządkowania zbioru tablic wynikowych. Kolejność prezentowania wyników w dużej mierze zależy od toku prowadzonego wywodu, czyli jest podporządkowana koncepcji badania, opartej na pewnej koncepcji intelektualnej. Jednakże

Recenzent: Dr inż. Mirosław Kwiesielewicz – Wydział Elektrotechniki i Automatyki Politechniki Gdańskiej

w wielu procesach badawczych, szczególnie o charakterze poszukiwawczym (eksploracyjnym), użytkownik wyników nie jest do końca ściśle określony, zaś badana rzeczywistość jest złożona, co prowadzi do tworzenia w opisie rezultatów badania znacznego zbioru tablic wynikowych. W takich sytuacjach, kolejność prezentacji tablic wynikowych ma wpływ na kreowaną informację i w ten sposób na interpretację rozpoznawanych zjawisk. W referacie uwagę skupiono na kryterium służącym do uszeregowania tablic wynikowych według zawartego w nich ładunku informacyjnego. Celem referatu jest zdefiniowanie miary ilości informacji. Przy tym autorzy skoncentrowali się na badaniu ilości informacji strukturalnej zawartej w wynikowych tablicach statystycznych. Proponowany mechanizm [1, 2], oparty na obiektywnym (datalogicznym) kryterium, powinien przyczynić się do racjonalizacji procesów przetwarzania danych.

2. INFOLOGICZNA KONCEPCJA INFORMACJI

Pojęcie informacji jest obecnie jednym z najbardziej fundamentalnych i najważniejszych pojęć stosowanych we współczesnej filozofii, w naukach teoretycznych i stosowanych oraz w praktyce sterowania systemami technicznymi, społecznymi, ekonomicznymi i biologicznymi. Każda z tych dziedzin definiuje je na własny sposób, ze względu na swoją charakterystykę i potrzeby. Powoduje to, że pojęcie informacji staje się jednym z tych pojęć, których w pełni nie wyczerpuje żadna formalna definicja. Początkowo (w latach 40-tych) pojęcie informacji oraz cała teoretyczna problematyka z nim związana odnoszone były do zastosowań w telekomunikacji. Jako klasyczną teorię informacji traktuje się teorię matematyczną. Taką bowiem postać nadał jej Shannon, uważany za twórcę ilościowej teorii informacji. Sam Shannon nie zdefiniował pojęcia informacji, definiując jedynie pojęcie jej ilości [3]. Co więcej, wielu cybernetyków nie definiuje pojęcia informacji, pomimo że cybernetyka jest nauką zajmującą się głównie informacją.

W opracowaniu algorytmu porządkującego wynikowe tablice statystyczne według ładunku informacyjnego, jaki one dostarczają, autorzy przyjęli za punkt wyjścia infologiczną koncepcję informacji przedstawioną po raz pierwszy przez Sundgrena [4] i Langeforse'a [5]. Formalna istota tej koncepcji wymaga zdefiniowania pojęcia komunikatu.

Def. 1. Niech dany będzie układ:

$$K := (O, X, x, t, q), \tag{1}$$

gdzie: O – obiekt, X – cecha (atrybut) obiektu O, x – wartość cechy X, t – czas, w którym cecha X obiektu O ma wartość x, q – wektor dodatkowych charakterystyk związanych z obiektem O, cechą X i (lub) czasem t.

Układ K jest komunikatem infologicznym [6]. Elementy O, X, x, t oraz q komunikatu K zapisane za pomocą odpowiednich znaków zgodnych z normami języka obowiązującego w systemie z jakiego pochodzą, noszą nazwę danych [6].

Komunikat K można rozpatrywać dwojako. Z punktu widzenia strukturalnego, gdy rozumiany jest jako pewien układ danych O, X, x, t, q oraz z punktu widzenia semantycznego, gdy K rozpatrywany jest jako opis obiektu O ze względu na cechę X w czasie t, przy dodatkowych charakterystykach q. Z wzajemnej relacji zachodzącej pomiędzy danymi O, X, x, t oraz q wynika znaczenie semantyczne i sens komunikatu K. Relację tę można nazwać informacją. „Informacja w interpretacji infologicznej to treść komunikatu K, dostarczana przez dane O, X, x, t, q i wynikająca ze wzajemnych zależności zachodzących między tymi danymi" [6].

Komunikat *K* pełni rolę nośnika informacji. *K* stanowi minimalny wystarczający zestaw danych do przekazania jednoznacznej treści. Treść zawarta w elementarnym komunikacie opisanym formułą (1) jest informacją elementarną.

Rozważmy teraz zbiór wszystkich obiektów (procesów lub zdarzeń) $\boldsymbol{O}$, zbiór cech $\boldsymbol{X}$ charakteryzujących obiekty ze zbioru $\boldsymbol{O}$, zbiór wartości cech ze zbioru $\boldsymbol{X}$ oznaczony $\{x\}$ oraz przedział czasu $\boldsymbol{T}$ wraz ze zbiorem dodatkowych charakterystyk $\boldsymbol{Q}$ uzupełniających pełny opis obiektów. Zbiory $\boldsymbol{O}$ oraz $\boldsymbol{X}$ są dyskretne, natomiast przedział czasu *T* jest zbiorem ciągłym. Wartości $\{x\}$ cech ze zbioru $\boldsymbol{X}$ przyjmują zarówno wartości mierzalne (dyskretne lub ciągłe) oraz niemierzalne. Zbiór wszystkich komunikatów *K* określonych formułą (1) o elementach odpowiednio ze zbiorów $\boldsymbol{O}$, $\boldsymbol{X}$, $\{x\}$, $\boldsymbol{T}$, $\boldsymbol{Q}$ jest zbiorem $\boldsymbol{K}$ komunikatów.

Dla zbiorów $\boldsymbol{O}$, $\boldsymbol{X}$, $\{x\}$, $\boldsymbol{T}$ możliwe jest zdefiniowanie produktu kartezjańskiego $\boldsymbol{O}\times\boldsymbol{X}\times\{x\}\times\boldsymbol{T}$. W ten sposób określona przestrzeń wielowymiarowa stanowi przestrzeń informacyjną $\boldsymbol{PI}$. Zatem przestrzeń informacyjna określona jest następująco:

$$\boldsymbol{PI} := \boldsymbol{O}\times\boldsymbol{X}\times\{x\}\times\boldsymbol{T}\,. \tag{2}$$

3. ILOŚĆ INFORMACJI STRUKTURALNEJ DOSTARCZANEJ PRZEZ TABLICE STATYSTYCZNE

Wynikowe informacje statystyczne stanowią końcowy produkt określonego metodologicznie ciągu kolejnych operacji i określane są w statystyce publicznej jako „wyniki obliczeń, opracowań i analiz dokonanych na podstawie zebranych w badaniach statystycznych statystyki publicznej danych statystycznych" (Dz. U. 1996 r., nr 88, poz. 439). Znaczna część informacji statystycznych dostarczana jest przez odpowiednie służby sprawozdawcze urzędów statystycznych, instytucji, jednostek administracyjnych, instytutów badań opinii publicznej, badań koniunktury gospodarczej.

W wyniku przetwarzania zebrane dane opracowane zostają do poziomu przewidywanego w informacjach wynikowych i zawartego w projekcie tablic. Zagregowane i opracowane informacje przedstawione są zazwyczaj w formie tablic, które mogą być dopuszczone do publikacji. Działania podejmowane w ostatniej fazie badania, czyli publikowaniu, nie są już nakierowane na dodanie nowych danych lub poprawienie ich wiarygodności. Część z tych działań ma na celu wręcz ograniczenie i selekcję informacji wynikowych możliwych do rozpowszechniania. Pierwszym ograniczeniem jest zastosowanie środków zapobiegających ujawnieniu poufnych danych. Drugim – wyselekcjonowanie zbiorów informacji o maksymalnej użyteczności.

Załóżmy zatem, że dana jest tablica *T* o *m* wierszach i *n* kolumnach. Przedmiot opisu uwzględniony w tablicy nazwiemy obiektem i oznaczymy symbolem $O_j \quad (j=1,\ldots,m)$. Zbiór wszystkich obiektów, jakie zostały uwzględnione w tablicy, oznaczony zostanie $\boldsymbol{O}$. Obiekty w tablicy *T* opisywane są ze względu na określoną cechę *X*. Przy powyższych założeniach, każdy obiekt O_j może być scharakteryzowany przez wektor [1]:

$$\boldsymbol{x}_j = \left(x_{j1}, x_{j2}, \ldots, x_{jn}\right) \in R_+^n\,, \tag{3}$$

gdzie x_{jk} oznacza *k*-ty wariant cechy *X* występujący w tablicy *T*, a $R_+^n = \left\{x \in R^n \,\middle|\, x \geq 0\right\}$.

Analizując tablicę *T* nietrudno dostrzec jej wielowymiarowość. Przyjmując powyższe założenia otrzymujemy następujący układ:

(obiekt, cecha obiektu, wartość cechy *X*, czas t, dodatkowe charakterystyki),
czyli:

$$\left(O_j, X_j, x_{jk}, t, q\right) \quad \left(j = 1,\ldots,m;\ k = 1,\ldots,n\right). \tag{4}$$

Wektor ***q*** dodatkowych charakterystyk może być dowolnego wymiaru i zależy od umieszczonych w boczku lub główce tablicy objaśnień bądź uwag wyjaśniających.

Układ (4) stanowi określony wcześniej formułą (1) komunikat będący jednostką elementarną układu o wyższym stopniu złożoności, czyli tablicy *T*. Komunikat (4) nazwiemy komunikatem prostym i oznaczymy *K*, natomiast zbiór wszystkich komunikatów prostych dla każdego $j = 1,\ldots,m$ oraz $k = 1,\ldots,n$ – komunikatem złożonym ***K***. Komunikat prosty – to jedno pole z tablicy z boczkiem, z główką oraz objaśnieniami. Tablica *T* analizowana w ujęciu datalogicznym stanowi zwarty i systematyczny zapis komunikatu złożonego ***K***, a dowolny zbiór tablic ***T*** wynikowych informacji statystycznych rozpatrywany jako zbiór komunikatów złożonych ***K*** stanowi podprzestrzeń przestrzeni informacyjnej ***PI***. Treść, jaką niesie komunikat ***K***, jest informacją w sensie datalogicznym.

Niech obiekty O_j będą scharakteryzowane przez wektory $\left[x_{jk}\right]$ lub $\left[n_{jk}\right]$ $\left(j = 1,\ldots,m;\ k = 1,\ldots,n\right)$, gdzie n_{jk} oznacza liczbę występujących *k*-tych wariantów cechy *X* w *j*-tym obiekcie badania. Dla każdego obiektu można wyznaczyć odpowiednio współczynniki struktury lub współczynniki udziału, oznaczane α_{jk}. Dysponując wskaźnikami struktury lub udziału, można zbudować macierz wskaźników:

$$\left[\alpha_{jk}\right] \quad \left(j = 1,\ldots,m;\ k = 1,\ldots,n\right). \tag{5}$$

Każdy obiekt O_j $\left(j = 1,\ldots,m\right)$ jest scharakteryzowany przez wiersz macierzy, czyli przez wektor wskaźników struktury $S_j = \left[\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jn}\right]$.

W niniejszym opracowaniu struktura będzie interpretowana jako obiekt opisany wektorem wskaźników struktury (lub udziału). Wyznaczanie wektora S_j jest zasadne tylko wtedy, gdy cecha *X* podlegająca badaniu spełnia własność addytywności, to znaczy wtedy, gdy suma wartości poszczególnych wariantów cechy przejawia sens interpretacyjny. W macierzy (5), w *j*-tym wierszu $\left(j = 1,\ldots,m\right)$ zapisane są wskaźniki struktury właściwe dla *j*-tego obiektu O_j będącego przedmiotem opisu tablicy *T*. Wiersz taki jednoznacznie odpowiada badanej strukturze rozumianej jako obiekt złożony z ciągu wskaźników struktury α_{jk}, spełniających warunki:

$$0 \le \alpha_{jk} \le 1 \quad \left(j = 1,\ldots,m;\ k = 1,\ldots,n\right), \tag{6}$$

$$\sum_{k=1}^{n} \alpha_{jk} = 1 \quad \left(j = 1,\ldots,m;\ k = 1,\ldots,n\right). \tag{7}$$

Dysponując pełnymi danymi o składnikach struktur S_j $\left(j = 1,\ldots,m\right)$, można wyznaczyć entropię rzeczywistą (empiryczną) struktury S_j obiektu O_j, przyjmując za podstawę

logarytmu liczbę 2 [3]:

$$H(S_j) = -\sum_{k=1}^{n} \alpha_{jk} \log_2 \alpha_{jk} . \tag{8}$$

Entropia $H(S_j)$ zależy wyłącznie od częstości występowania *k*-tego wariantu cechy *X* w *j*-tej strukturze, a więc od wskaźników struktury (lub udziału) charakteryzujących dany obiekt O_j.

Entropia $H(S_j)$ osiągnie maksimum równe $\log_2 n = H_{max}$ dla struktury S_j o wskaźnikach α_{jk}, takich, że:

$$\alpha_{j1} = \alpha_{j2} = \ldots = \alpha_{jn} . \tag{9}$$

Dla struktury S_j $(j = 1,\ldots,m)$ zdefiniowana zostanie miara dekoncentracji [2].

Def. 2. Wskaźnik dekoncentracji struktury S_j jest stosunkiem entropii rzeczywistej $H(S_j)$ do maksymalnej wartości entropii H_{max}:

$$DC_{S_j} = \frac{H(S_j)}{H_{max}} . \tag{10}$$

Wskaźnik dekoncentracji struktury S_j jest miarą dekoncentracji rozkładu wartości cechy *X* dla badanego obiektu O_j a także rozkładu cechy *X* w czasie [1].

Na podstawie wskaźnika dekoncentracji DC_{S_j}, który jest unormowany w przedziale $[0,1]$, można wnioskować o rozkładzie wartości badanej cechy w zbiorowości statystycznej. Gdy $DC_{S_j} = 1$ oznacza to, że w zbiorowości statystycznej występuje maksymalne zróżnicowanie $(H(S_j) = H_{max})$ wariantów badanej cechy, czyli równomierny rozkład jednostek statystycznych na wszystkie warianty cechy *X*. Osiąganie przez współczynnik dekoncentracji największych wartości bliskich 1, towarzyszy sytuacji, w której współczynnik zmienności liczony na podstawie wariantów cechy *X*, jest jak najmniejszy.

W *m*-elementowym zbiorze ***O*** wszystkich obiektów opisanych w tablicy *T* łączymy obiekty w pary (O_i, O_j), gdzie *i*, *j* są numerami obiektów oraz $i \neq j$ $(i, j = 1,\ldots,m)$. Dla *m* obiektów, można wyróżnić $m!/(m-2)!$ par obiektów. Dla każdej pary (O_i, O_j) $(i \neq j;\ i, j = 1,\ldots,m)$ określony zostanie wskaźnik struktury pary obiektów oznaczony α_{ijk}.

Def. 3. Wskaźnik struktury pary obiektów (O_i, O_j) $(i \neq j;\ i, j = 1,\ldots,m)$ jest stosunkiem *k*-tej wartości zmiennej *X* dla *i*-tego obiektu badania, do sumy realizacji wartości zmiennej *X* dla obiektów O_i oraz O_j [1]:

$$\alpha_{ijk} = \frac{n_{ik}}{\sum_{k=1}^{n} n_{ik} + \sum_{k=1}^{n} n_{jk}} \quad (i, j = 1,\ldots,m;\ i \neq j;\ k = 1,\ldots,n), \tag{11}$$

gdzie n_{ik} jest liczbą jednostek o *k*-tym wariancie cechy w *i*-tym obiekcie badania, n_{jk} – liczbą jednostek o *k*-tym wariancie cechy w *j*- tym obiekcie badania.

Wskaźniki struktury α_{ijk} wyrażają częstość występowania *k*-tej realizacji cechy *X* obiektu O_i w łącznej sumie realizacji cechy *X* dla dwóch obiektów badania: O_i, O_j.

Wskaźniki α_{ijk} oraz α_{jik} spełniają relację:

$$\sum_{k=1}^{n} \alpha_{ijk} + \sum_{k=1}^{n} \alpha_{jik} = 1 \quad (i, j = 1,\ldots,m;\ i \neq j;\ k = 1,\ldots,n). \tag{12}$$

Dysponując wskaźnikami struktury par obiektów (O_i, O_j) dla wszystkich $i, j = 1,\ldots,m$, takich, że $i \neq j$, można zbudować $m!/(m-2)!$ tablic współczynników struktury par obiektów. Wskaźniki α_{ijk} są podstawą do wyznaczenia entropii warunkowej pary obiektów.

Def. 4. Entropia warunkowa $H(O_i, O_j)$ **pary obiektów** (O_i, O_j) określona jest wzorem:

$$H(O_i / O_j) = -\sum_{k=1}^{n} \alpha_{ijk} \log_2 \alpha_{ijk} \quad (i, j = 1,\ldots,m;\ i \neq j;\ k = 1,\ldots,n). \tag{13}$$

Ponieważ $\sum_{k=1}^{n} \alpha_{ijk} \neq \sum_{k=1}^{n} \alpha_{jik}$ (równość zachodzi tylko wtedy, gdy wektory charakteryzujące struktury S_i oraz S_j są sobie równe, to znaczy $\alpha_{ijk} = \alpha_{jik}$ $(i, j = 1,\ldots,m;\ i \neq j;\ k = 1,\ldots,n)$, oczywistym jest fakt, że entropia $H(O_i / O_j)$ nie spełnia warunku symetrii.

Jeśli ***O*** jest zbiorem obiektów badania opisanych w tablicy *T*, można zdefiniować wskaźnik struktury obiektu O_j w całym zbiorze ***O*** następująco:

Def. 5. Wskaźnik struktury obiektu O_j w zbiorze ***O*** jest stosunkiem sumy wartości zmiennej *X* dla obiektu O_j do sumy wszystkich realizacji zmiennej *X* w całym zbiorze ***O*** [1]:

$$\alpha_j = \frac{\sum_{k=1}^{n} n_{jk}}{\sum_{j=1}^{m} \sum_{k=1}^{n} n_{jk}} \quad (j = 1,\ldots,m;\ k = 1,\ldots,n). \tag{14}$$

Uwzględnienie realizacji cechy *X* dla wszystkich obiektów jednocześnie zmniejsza

entropię warunkową będącą wartością oczekiwaną informacji. Średnia entropia warunkowa jest postaci:

$$H(O_j/\boldsymbol{O}) = \sum_{i=1}^{m} H(O_j/\boldsymbol{O}) \cdot \alpha_i \quad (i \neq j;\ i, j = 1, \ldots, m). \tag{15}$$

Znajomość jednocześnie entropii rzeczywistej $H(O_j)$ oraz średniej entropii warunkowej $H(O_j/\boldsymbol{O})$ pozwala na zastosowanie wzoru Shannona określającego ilość informacji [7, 8].

Def. 6. Ilość informacji strukturalnej $I(O_j/\boldsymbol{O})$ wynikającej ze struktury obiektu O_j, jest różnicą pomiędzy entropią rzeczywistą obiektu $H(O_j)$ oraz średnią entropią warunkową $H(O_j/\boldsymbol{O})$ [7]:

$$I(O_j/\boldsymbol{O}) = H(O_j) - H(O_j/\boldsymbol{O}). \tag{16}$$

Ilość informacji strukturalnej wyraża ilość informacji o strukturze obiektu O_j w zbiorze obiektów $\boldsymbol{O}$, opisanych za pomocą komunikatów elementarnych *K*, należących do komunikatu złożonego $\boldsymbol{K}$. Wielkość (16) zależy nie tylko od struktury obiektu O_j, ale również od wzajemnych relacji i powiązań pomiędzy strukturą tego obiektu, a strukturami pozostałych obiektów ze zbioru $\boldsymbol{O}$, uwzględnionych w opisie. Tablica *T*, rozumiana w sensie datalogicznym jako komunikat, niesie treść o zjawiskach uwzględnionych w badaniach statystycznych. Treść tę rozumiemy jako informację. W szczególności treść wynikająca ze struktury obiektów ze zbioru $\boldsymbol{O}$ jest informacją strukturalną. $\boldsymbol{E}$-miara ilości informacji strukturalnej określona zostanie w następujący sposób [1, 2]:

Def. 7. Ilość informacji strukturalnej dostarczanej przez tablicę *T* jest średnią geometryczną ilości informacji strukturalnej przekazywanej przez obiekty $O_j \in \boldsymbol{O}$ $(j = 1, \ldots, m)$:

$$E(T) = \sqrt[m]{\prod_{j=1}^{m} I(O_j/\boldsymbol{O})}. \tag{17}$$

Ilość informacji strukturalnej $\boldsymbol{E(T)}$ – wyraża wielkość ładunku informacyjnego dostarczanego przez tablicę *T,* wynikającego ze struktury obiektów w niej uwzględnionych. Przedstawiona miara jest propozycją teoretyczną, opartą na datalogicznej interpretacji informacji wynikającej wyłącznie ze struktury obiektów. Użyteczność praktyczna miary $\boldsymbol{E(T)}$ ujawnia się w sytuacji wymagającej uszeregowania tablic pod względem dostarczanego przez nie ładunku informacyjnego. W szczególności $\boldsymbol{E}$-miara może dopomóc osobom publikującym informacje wynikowe w wyborze takich tablic, które niosą jak największy potencjał treściowy o strukturze obiektów.

4. ILUSTRACJA EMPIRYCZNA

Wynikowe informacje statystyczne zgromadzone są w bazach danych, a możliwości przyjęcia i wykorzystania tych informacji maleją wraz ze wzrostem rozmiarów nagroma-

dzonych baz danych. Rodzi się zatem problem wykorzystania wszystkich informacji wynikowych, które zostały zgromadzone, bowiem część tablic wynikowych zawiera ważny materiał informacyjny, wykorzystywany między innymi przez statystyków i ekonomistów w prowadzonych przez nich analizach. Natomiast pozostała część tablic nie jest zazwyczaj wykorzystywana, ze względu na mniej interesujący materiał statystyczny. Specjaliści – statystycy zajmujący się analizą tablic oceniają przydatność tablic wynikowych w badaniach statystycznych i podejmują decyzje, które z tablic powinny ukazać się w publikacjach. Tablice wynikowe, które wybrane zostały ze względu na ich przydatność mogą zostać uporządkowane wg ilości informacji strukturalnej dostarczanej przez te tablice, wyznaczonej zgodnie z formułą (17).

Zgodnie z powyżej przedstawioną metodą, opracowano algorytm obliczania $\boldsymbol{E}(T)$ (tab. 1), który został zaimplementowany w języku *Visual Basic for Application* – wykorzystującym strukturę danych arkusza kalkulacyjnego Excel. Wybór środowiska został podyktowany łatwością wprowadzenia danych tablicowych do Excela oraz wieloma narzędziami umożliwiających konwersję różnych struktur danych (np. z baz danych) do skoroszytów arkusza kalkulacyjnego. Również wyniki pośrednie mogą być zapisywane w odpowiednich skoroszytach i być wykorzystywane do analizy danych statystycznych innymi metodami.

Przeanalizowanych zostało sześć tablic informacji wynikowych $T_I,\ldots,T_{VI}$ pochodzących z [9]. Znajomość ilości $\boldsymbol{E}(T)$ informacji strukturalnej zawartej w każdej z tablic $T_I,\ldots,T_{VI}$ posłużyła do ustalenia odpowiedniego ich porządku (tab. 2).

Tablica 1. Algorytm wyznaczania $\boldsymbol{E}(T)$ w pseudokodzie

```
Begin
  For each obiekt O_j do
    Oblicz wektor wskaźników struktury S_j
  EndFor
  For each struktura S_j do
    Oblicz entropię rzeczywistą H(S_j)
    Oblicz wskaźnik dekoncentracji struktury DC_Sj
  EndFor
  For k=1 to n do
    For i=1 to m do
      For j=1 to m do
        If i≠j then oblicz α_ijk
      EndFor
    EndFor
  EndFor
  For each obiekt O_j do
    Oblicz wskaźnik struktury obiektu O_j w zbiorze O
    Oblicz średnią entropię warunkową H(O_j/O)
    Oblicz ilość informacji strukturalnej I(O_j/O)
  EndFor
  Oblicz ilość informacji strukturalnej E(T)
End
```

Źródło: Opracowanie własne

Tablica 2. Ustalony porządek tablic zgodnie z kryterium $\boldsymbol{E}(T)$

Ranga	Tytuł tablicy	*E(T)*
1	Prognoza gospodarstw domowych wg liczby osób [9, Tabl. 58, s. 212]	0,8266
2	Ludność w wieku 15 lat i więcej wg poziomu wykształcenia i województw w roku 2002 [9, Tabl. 27, s. 167]	0,7520
3	Ludność wsi w wieku 15 lat i więcej wg poziomu wykształcenia i województw w roku 2002 [9, Tabl. 27, s. 169]	0,6624
4	Rozwody w 2003 r. wg liczby małoletnich dzieci oraz województw [9, Tabl. 48 (106), s. 291]	0,6046
5	Urodzenia żywe w 2003 r. wg kolejności urodzenia dziecka oraz województw [9, Tabl. 68 (126), s. 311]	0,5776
6	Dzieci pozostające z małżeństw rozwiedzionych w 2003 r. wg wieku oraz województw [9, Tabl. 49 (107), s. 291]	0,4896

Źródło: Opracowanie własne

Ustalona hierarchia tablic może być traktowana jako wskazówka w sprawie kolejności analizy tablic $T_I,\ldots,T_{VI}$, może to być także wskazówka, w jakiej kolejności należałoby je drukować (publikować), gdyby ich odbiorca nie określił żadnego innego kryterium lub gdyby owi odbiorcy nie byli znani w momencie drukowania $T_I,\ldots,T_{VI}$.

Analiza uzyskanych wyników pozwala na sformułowanie kilku istotnych wniosków dotyczących ***E***-miary:

1. ***E***-miara, a w konsekwencji wynikający z jej stosowania porządek tablic, pozwala na pełniejsze poznanie właściwości zbioru wartości liczbowych zamieszczonych w tych tablicach, rzeczywistych różnic pomiędzy liczbami zamieszczonymi w tablicy oraz ich wzajemnych relacji.
2. ***E***-miara jest jedną z nielicznych prób podejmowanych przez specjalistów przedstawienia niektórych aspektów datalogicznego ujęcia informacji i zastosowania go w procesach analizy wynikowych informacji statystycznych.
3. ***E***-miara i wyznaczany przez nią porządek rozpatrywanych tablic jest obiektywną konsekwencją występowania w tablicach niejawnych związków (relacji) pomiędzy danymi (wartościami) zamieszczonymi w tych tablicach. Oznacza to występowanie w każdej takiej tablicy nie tylko informacji faktograficznej zawartej w danych, lecz także informacji strukturalnej wynikającej ze wzajemnych relacji między tymi danymi.
4. Przedstawiony porządek tablic, ustalony na podstawie ***E(T)***, stanowi odzwierciedlenie jednego z przejawów różnorodności charakteryzujących obiekty, w rzeczywistości dotyczy bowiem jednego z elementów komunikatu *K* – liczbowych wartości cechy *X*.
5. Proponowana ***E***-miara stanowi datalogiczne kryterium sortowania tablic wynikowych. W ustalonym na jej podstawie porządku tablic nie uwzględnione są inne przejawy różnorodności (różnorodność wynikająca z odmienności rozpatrywanych obiektów i cech oraz subiektywnej interpretacji odbiorcy) charakteryzujące obiekty. W konsekwencji nie została uwzględniona treść dostarczana przez tablicę rozumianą jako komunikat złożony ***K***. Porządek ten nie jest zatem porządkiem infologicznym. Może jednak stanowić wstępne rangowanie tablic wynikowych, umożliwiając dostarczanie odbiorcy wynikowych informacji statystycznych w pewnej, ustalonej obiektywnie, hierarchii. Umożliwi to zatem analizowanie i wykorzystanie informacji wynikowych w sposób usystematyzowany.
6. ***E-miara*** jako kryterium porządkowania zbioru tablic statystycznych wykorzystać można również do odpowiedniego ich „porcjowania" w procesie przekazywania użytkowni-

kowi. Zaprezentowane przez autorkę kryterium porządkujące tablice wynikowe przyczyni się zatem do usprawnienia i racjonalizacji procesów przetwarzania i interpretacji danych.

5. ZAKOŃCZENIE

W artykule zaprezentowana została miara ilości informacji strukturalnej, stanowiąca kryterium porządkowania tablic statystycznych, traktowanych jako komunikat złożony ***K***. Ustalony porządek tablic dokonany jest według ilości informacji strukturalnej dostarczanej przez te tablice. Porządek ten może stanowić kryterium na poziomie datalogicznym kolejności publikowania tablic w warunkach, kiedy użytkownik informacji nie jest ściśle określony. Artykuł stanowi przede wszystkim pewien wkład do szerszych badań na temat formułowania miar ilości informacji. Dalsze badania, zarówno teoretyczne jak i praktyczne, konieczne są w zakresie wypracowania teoretycznie poprawnych i praktycznie użytecznych miar ilości informacji, odpowiednich dla charakteru i postaci informacji.

6. BIBLIOGRAFIA

1. Wędrowska E.: Datalogiczna miara ilości informacji strukturalnej jako instrument zarządzania zasobami informacji statystycznej, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, nr 975, Wrocław 2003, s. 447–459.
2. Wędrowska E.: Datalogiczne aspekty informacji w procesach analizy zasobów informacyjnych, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 999, Wrocław 2003, s. 102–122.
3. Shannon C.E., Weaver W.: The mathematical theory of communication, Urbana 1948.
4. Sundgren B.: An infological approach to data bases, Skriftserie Statistiska Centralbyran, Lund, Sztokholm 1973.
5. Langeforse B.: Infological modeles and information users view, Information Systems, Vol. 5, 1980.
6. Stefanowicz B.: Różnorodność informacji, Wiadomości Statystyczne, nr 4, GUS 1996.
7. Nowakowski J., Sobczak W.: Teoria informacji, WNT, Warszawa 1970.
8. Sobczak W., Malina W.: Metody selekcji informacji, WNT, Warszawa 1978.
9. Rocznik Demograficzny 2004, GUS, Warszawa 2004.

STATISTICAL TABLES RANKING ALGORITHM

The proposed article is an attempt to establish an appropriate filtering algorithm. Its purpose is to define the measure of the information amount, so that selection of the tables carrying the largest information load becomes possible. The authors have concentrated on the study of the amount of structural information included in the statistical tables. The purpose of proposed measure of structural information amount is to provide the ranking of the charts as per the supplied information load.