Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej Nr 21

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

BADANIE STRUKTURY AKADEMICKIEGO SPOŁECZEŃSTWA INFORMACYJNEGO Z WYKORZYSTANIEM METODY MDS

Małgorzata KALICZYŃSKA

Politechnika Opolska, Instytut Automatyki i Informatyki
45-272 Opole, ul. Sosnkowskiego 31
tel.: (77)4006142 fax: (77)4006388 e-mail: mkal@po.opole.pl

W referacie zaprezentowano metodologię badań webometrycznych, których celem jest określenie podobieństw obiektów należących do akademickiego społeczeństwa informacyjnego wykorzystującego Internet i jego zasoby WWW do komunikowania się i wymiany informacji. W szczególności przeprowadzone badania dotyczą wybranych polskich uczelni, również na tle innych organizacji. Badania przeprowadzono w kilku etapach – na różnych grupach danych – z uwzględnieniem zmian w przestrzeni internetowej. Ich wyniki zostały przedstawione w postaci map topograficznych w przestrzeniach 2D oraz 3D. W analizie zastosowano komputerowe metody skalowania wielowymiarowego. Badania będą prowadzone dalej z uwzględnieniem dłuższego horyzontu czasowego dla wybranej grupy obiektów. Kolejny etap to badania bardziej zróżnicowanych obiektów, także w obszarze międzynarodowym.

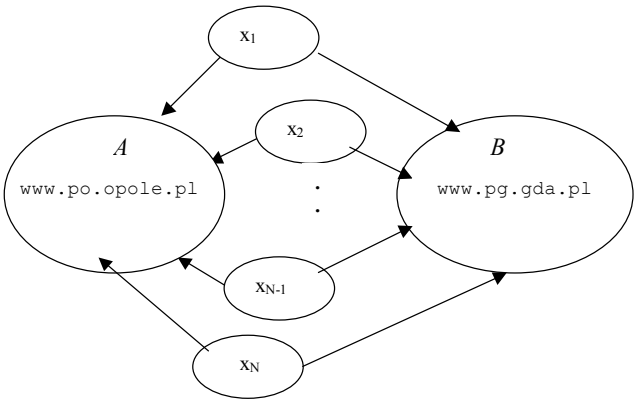
1. SPOŁECZEŃSTWO INFORMACYJNE – PODSTAWOWE POJĘCIA

Powszechny dostęp do Internetu zapoczątkował niekontrolowany proces samoczynnego tworzenia się społeczeństwa informacyjnego – kontaktującego się ze sobą, wymieniającego się informacjami i wiedzą, współpracującego na wielu płaszczyznach, czyli organizującego się (ang. *self-organization*) [1]. Pojawia się pytanie, czy istnieją zależności między różnymi ośrodkami uczestniczącymi w opisanym procesie – jak silne są powiązania między nimi, jakie występują między nimi zależności. Ogromną rolę w procesie tworzenia społeczeństwa informacyjnego odgrywają ośrodki akademickie, to one wymogły rozwój *cyberprzestrzeni*. W czasie swobodnego dostępu do Internetu, przy braku cenzury, serwisy internetowe uczelni i innych ośrodków naukowych są postrzegane jako skarbnice wiarygodnych danych. Również one są obiektem badań webometrycznych [2, 3]. Przed przystąpieniem do badań należy zdefiniować, co należy rozumieć przez *społeczeństwo informacyjne*. W tym celu wykorzystane zostaną zasoby internetowe, a w szczególności strony internetowe.

Załóżmy, że x_1, x_2, x_3, … x_N są serwisami (stronami) internetowymi organizacji, firm, uczelni, szkół, osób prywatnych i innych podmiotów. Każda witryna internetowa zawiera

Recenzent: Dr inż. Longin Stolc - Wydział Elektrotechniki i Automatyki Politechniki Gdańskiej

liczne powiązania hipertekstowe, tzw. *linki* do innych portali. Jeżeli znajdziemy grupę stron $x_1, x_2, x_3, \ldots x_N$, które mają odnośniki do portali *A* oraz *B* (rys. 1), to można przypuszczać, że portale *A* i *B* są podobne, gdyż wielu internatów uważa je za istotne i cytuje je (umieszcza linki) do tych portali jednocześnie. Takie wnioskowanie nie pozwala określić przyczyn, jakie wpływają na podobne traktowanie. Fakt powiązań jest jednak oczywisty.



Rys. 1. Uproszczony model fragmentu społeczeństwa informacyjnego

2. METODOLOGIA BADAŃ

By stwierdzić, jak postrzegane są strony internetowe polskich uczelni, przeprowadzone zostały badania, które pozwalają grupować badane obiekty z wykorzystaniem metod statystycznych, a w szczególności skalowania wielowymiarowego. Przedmiotem badań jest struktura sieci WWW, a w szczególności jej fragment odnoszący się do polskich ośrodków akademickich i ich udziału w sieci. W pierwszej fazie do badań wybrane zostały uczelnie techniczne biorące udział w projekcie *Wirtualna Politechnika*[1] oraz Politechnika Opolska. Badania powinny doprowadzić do uzyskania spójnych i umożliwiających jednoznaczną interpretację wyników.

2.1. Zbieranie danych

Ponieważ badania realizowane są w sieci WWW, uwagę skoncentrowano na serwerach sieciowych na poziomie poddomen (ang. *sub-domain*), oznaczane skrótem **sdws** (ang. *sub-domain web server*). Przykładowo `pg.gda.pl` to **sdws** Politechniki Gdańskiej.

Danymi do badań są informacje o liczbie powiązań między ośrodkami zgromadzone w macierzy odniesień. Aby otrzymać stosowne dane, przeszukano sieć stosując polecenia popularnej niegdyś wyszukiwarki *AltaVista*:

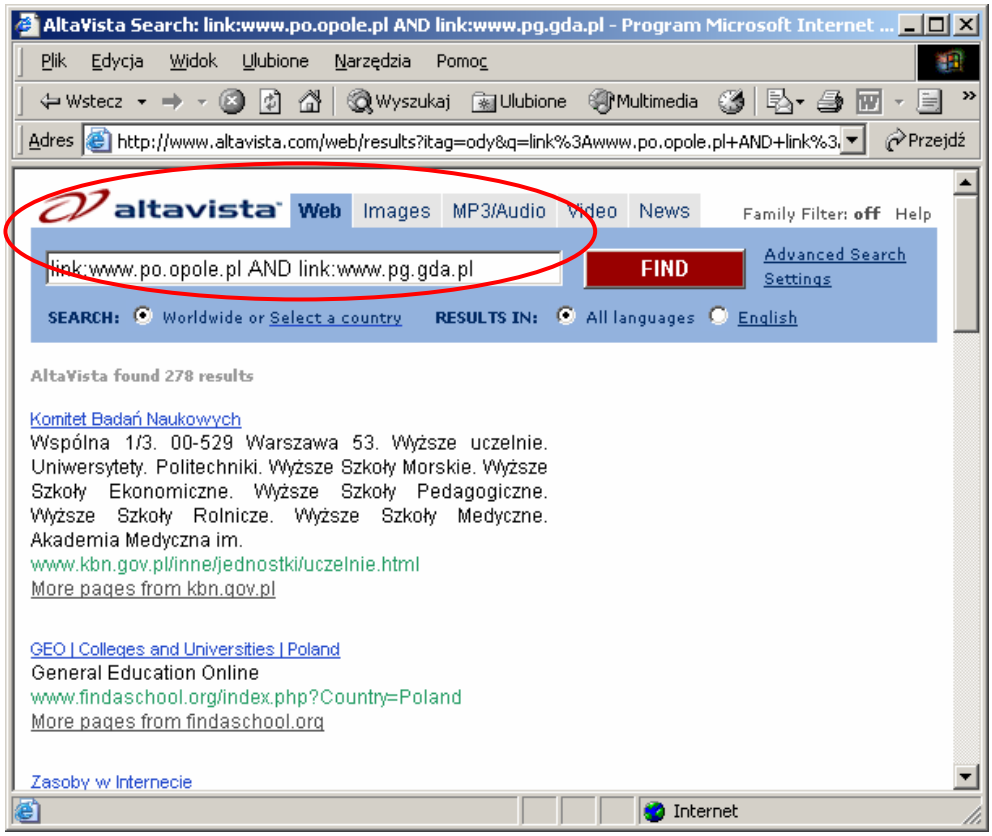
```
link:sdws(i) AND link:sdws(j),
                                    dla i,j = 1, ..., 8,
```

Podczas badań będą zadawane pytania typu:

1 Wirtualna Politechnika została powołana 10 XII 2002r.

```
link:www.po.opole.pl AND link:www.pg.gda.pl
```

w przypadku zadanego pytania, wyszukiwarka *AltaVista* znajduje 278 adresów internetowych (rys. 2), które mają linki do stron Politechniki Opolskiej oraz Politechniki Gdańskiej (w dniu 29 września 2005 roku).



Rys. 2. Wykorzystanie wyszukiwarki *AltaVista* do zbierania danych

2.2. Przetwarzanie danych – skalowanie wielowymiarowe MDS

Skalowanie wielowymiarowe [4, 5] **MDS** (ang. *multidimensional scaling*) należy do metod niezupełnej analizy skupień. Ideą analizy tego typu jest dzielenie zbioru obiektów na klasy bez określonego wcześniej kryterium zewnętrznego. Kryterium klasyfikacyjne, czyli reguła pozwalająca przypisać obiekty do poszczególnych grup jest tworzone w trakcie analizy na podstawie uwzględnionych cech. Analiza skupień wyodrębnia grupy obiektów według zasady podobieństwa. Klastry (skupienia) są tworzone przez obiekty, które są bardziej podobne do obiektów współtworzących dane skupienie niż do obiektów innych skupień. Kryterium klasyfikacyjnym jest matematycznie zdefiniowane podobieństwo. Innymi słowy – im bliżej siebie (na mapie) położone są obiekty lub cechy – tym bardziej są do siebie podobne lub tym częściej współwystępują. Pod pojęciem MDS rozumie się całą rodzinę technik ukierunkowanych na topograficznie dokładne rozróżnianie danych w ma-

łowymiarowej przestrzeni. Zachowanie topografii struktury danych to zarówno zachowanie ich jakościowych własności topologicznych, jak i własności metryki.

W skalowaniu wielowymiarowym [4, 5] rozpatrywana jest przestrzeń danych $\Re^N$, wektory danych **X** mapowane są na przestrzeń docelową, zwykle $\mathbf{Y} \in \Re^2$. Rozpatrywane są odległości $R_{ij} = D\left(X^i, X^j\right)$ pomiędzy X^i i X^j w $\Re^N$ oraz odległości $r_{ij} = d\left(Y^i, Y^j\right)$ w $\Re^2$. Celem metody MDS jest znalezienie mapy $X \to Y = M\left(X\right)$ minimalizującej globalne miary zgodności topograficznej, czyli różnicę między R_{ij} i r_{ij}.

Skalowanie wielowymiarowe rozpoczyna się od wyznaczenia macierzy przedstawiającej odległości między parami badanych obiektów. Geometrycznym obrazem pozycji obiektów w wybranych wymiarach odpowiadających skalom jest tzw. przestrzeń postrzegania – zwana także mapą percepcji. Wyznaczają je współrzędne prostopadłe.

	PO				**Macierz odległości**		
AGH	110	**AGH**					
PG	148	133	**PG**				
PKr	170	198	189	**PKr**			
PB	99	86	103	97	**PB**		
PP	153	142	184	246	98	**PP**	
PWr	155	147	168	194	104	165	**PWr**
PW	151	162	188	209	105	196	205

Interpretując znaczenie danych, liczba umieszczona na pozycji (`i`, `j`) w macierzy odległości oznacza liczbę stron internetowych zawierających łącza zarówno do `i`-tego, jak i do `j`-tego **sdws**. Mówiąc dokładniej, pary ośrodków wskazywane łączami z różnych domen są podobne. Dane przedstawione w postaci pierwotnej macierzy odległości o rozmiarze 8 × 8 zostaną poddane dalszemu przetwarzaniu. Ponieważ jest to macierz symetryczna, przytoczona została jedynie część pod główną przekątną.

Zaprezentowana macierz odległości ośrodków akademickich pozwala, przy wykorzystaniu technik skalowania wielowymiarowego, odtworzyć wzajemne rozłożenie ośrodków w przestrzeni wielowymiarowej. Macierz ta została przekształcona do macierzy podobieństw (przedstawiona poniżej). Zastosowano tu zależności korelacji liniowej Persona, gdzie współczynniki przyjmują wartości z przedziału [-1, 1]. Im większa jest wartość współczynnika na pozycji (`i`, `j`), tym bardziej podobne są ośrodki `i` oraz `j`. Wartości na głównej przekątnej oznaczają samopodobieństwo.

PO	1,000					**Macierz podobieństw**		
AGH	0,833	1,000						
PG	0,690	0,810	1,000					
PKr	0,119	0,000	0,143	1,000				
PB	0,286	0,548	0,262	0,405	1,000			
PP	0,452	0,690	0,405	-0,333	0,500	1,000		
PWr	0,286	0,571	0,476	0,095	0,333	0,595	1,000	
PW	0,738	0,429	0,310	-0,167	-0,048	0,190	-0,071	1,000

Algorytmy skalowania wielowymiarowego zostały zrealizowane za pomocą programu ***XL*STAT**[2] firmy Addinsoft – nakładki na program MS Excel. Program wykorzystuje procedurę iteracyjną do minimalizacji wartości naprężeń (ang. *stress*). Użytkownik może kontrolować iteracje i sprawdzać zmiany tych wartości. Konfigurację końcową można przeglądać w arkuszach lub na dwu- lub trójwymiarowych wykresach rozrzutu (mapach) dla przestrzeni wielowymiarowej z opisanymi punktami - obiektami. Wizualizacja w niskowymiarowych przestrzeniach wymaga oceny stopnia zniekształcenia - miary liczbowej. Zazwyczaj stosowane są proste miary zgodności topograficznej:

– współczynnik naprężeń i alienacji (Kruskal); może to być dowolna funkcja o nieujemnych przyczynkach, np. funkcje entropowe [5]:

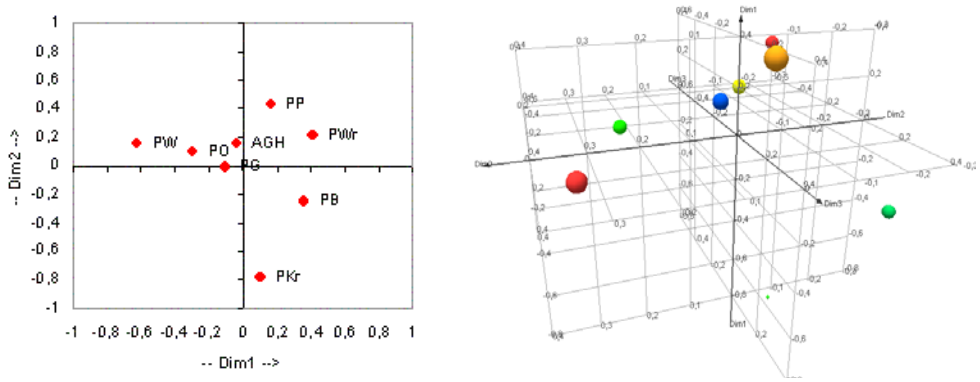
$$E(r) = \sum_{i>j}^{K} \left(R_{ij} - r_{ij}\right)^2 , \tag{1}$$

$$E_a(r) = \sum_{i>j}^{K} \left(1 - \frac{R_{ij}}{r_{ij}}\right)^2 , \tag{2}$$

– miara transmisji informacji określa, ile informacji uległo straceniu przy redukcji wielowymiarowości [4, 5]:

$$0 \le A(r) = \frac{\sum_{i>j}^{K} \left(R_{ij} - r_{ij}\right)^2}{\sum_{i>j}^{K} R_{ij}^2 + \sum_{i>j}^{K} r_{ij}^2} \le 1 . \tag{3}$$

Należy zauważyć, że nie istnieje ogólna funkcja opisująca badane obiekty i ich podobieństwo. Postać funkcji opisującej ściśle zależy od ustalonego zbioru punktów. Umieszczenie nowego punktu na mapie wymaga nowej minimalizacji.



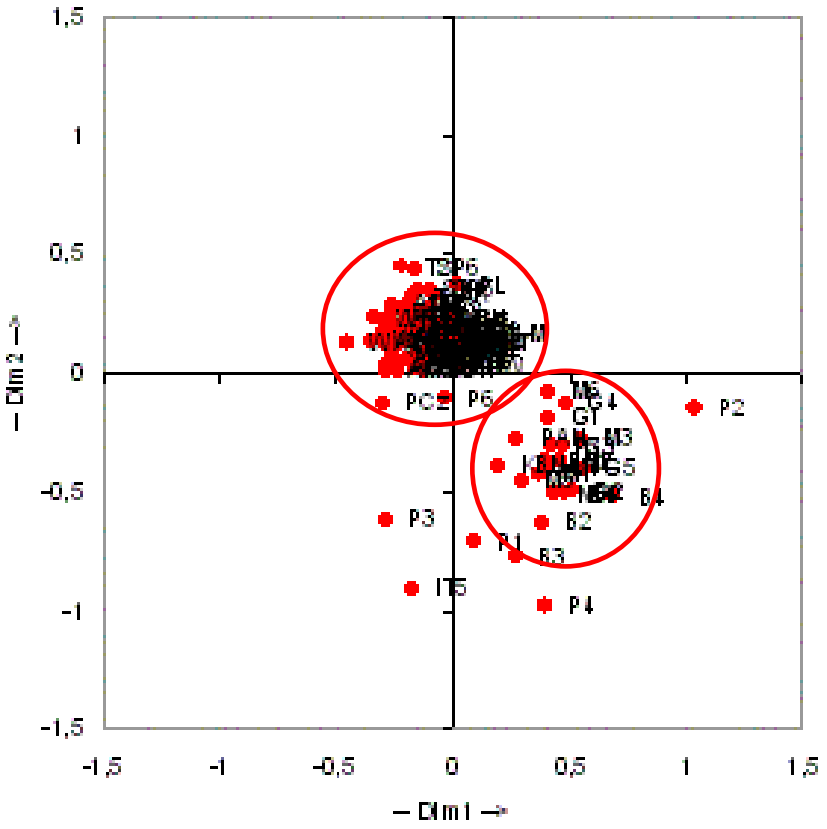
Rys. 3. Mapy w przestrzeni 2D oraz 3D

2 www.xlstat.com

Przeprowadzone badania pozwalają znaleźć podobieństwa oraz różnice między rozpatrywanymi obiektami. Bez względu na stosowane kryteria oceny – minimalizację zniekształceń, wyraźnie powtarzają się grupy obiektów leżących blisko siebie (rys. 3), a więc podobnych. Najbardziej zbliżone (podobne) są do siebie **AGH**, **PG** oraz **PO**, najbardziej oddalony od pozostałych obiektów (a więc inny) jest **PKr**. Należy stwierdzić, że liczba badanych obiektów jest zdecydowanie zbyt mała by uogólniać uzyskane wyniki.

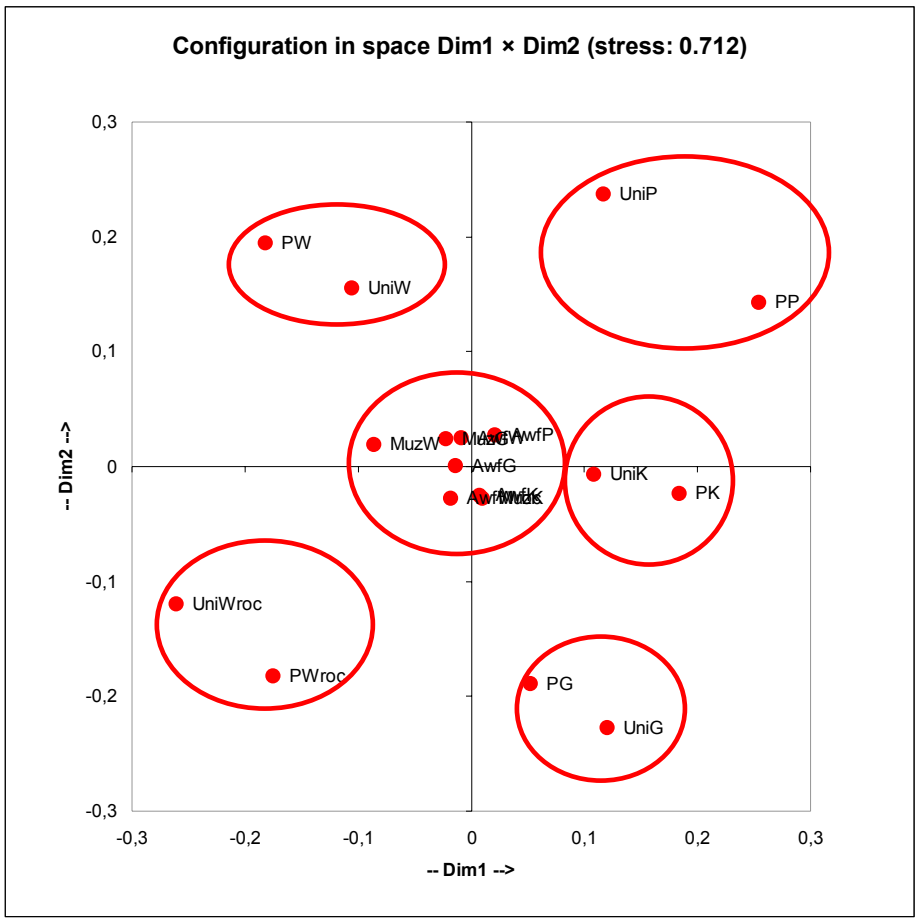
3. STRUKTURA POLSKIEGO AKADEMICKIEGO SPOŁECZEŃSTWA INFORMACYJNEGO NA TLE INNYCH ORGANIZACJI

Kolejne badania zostały przeprowadzone na grupie 99 obiektów. Poprzedni zbiór został powiększony o uniwersytety, wyższe szkoły pedagogiczne, akademie medyczne, akademie wychowania fizycznego, a także banki, ministerstwa, organizacje naukowe – KBN, PAN, wydawców czasopism takich, jak *Gazeta Wyborcza*, *Rzeczpospolita*, *Wprost*, *Polityka*, przemysł oraz firmy informatyczne. Wyniki analizy MDS zostały przedstawione w postaci mapy 2D na rysunku 4. Wyraźny jest tutaj podział na dwie grupy. Łatwo zauważyć, że w grupie pierwszej znajdują się uczelnie oraz nieliczne firmy związane z przemysłem. Odrębną grupę stanowią banki, prasa, ministerstwa, a wśród nich także PAN i KBN.



Rys. 4. Mapa w przestrzeni 2D dla 99 obiektów [opracowanie własne]

Wyniki kolejnych badań, przedstawione na rysunku 5, są ukierunkowane na położenie geograficzne. Możemy zauważyć odrębne zgrupowania uczelni Gdańska, Krakowa, Poznania, Wrocławia czy Warszawy. Środkowa grupa to uczelnie muzyczne, akademie wychowania fizycznego, które w tym przypadku są do siebie zbliżone bez względu na ich położenie geograficzne.



Rys. 5. Zależności geograficzne obiektów akademickich [opracowanie własne]

4. PODSUMOWANIE

Przedstawione wyniki wskazują, że metoda skalowania wielowymiarowego pozwala grupować badane obiekty w zależności od znanych parametrów. By wykazać jej przydatność przeprowadzono cały szereg kolejnych badań, których wyniki przedstawione są na kolejnych rysunkach.

Przeprowadzone badania pozwalają znaleźć podobieństwa oraz różnice między rozpatrywanymi obiektami. Trudno jednak określić, jakie czynniki mają tu decydujące znaczenie, chociaż już teraz widać, że takie podobieństwa są, obserwuje się grupy obiektów

o podobnym profilu specjalnościowym czy też położeniu geograficznym. Można również badać trendy rozwoju struktury społeczeństwa informacyjnego i przewidzieć ich kierunki na najbliższą przyszłość. Należy zaznaczyć, że stosowana do badań wyszukiwarka internetowa *AltaVista* nie pozwala sięgnąć po informacje do „głębokiego" Internetu – nie obsługuje baz danych, więc nie uwzględnia stron WWW tworzonych dynamicznie, gdzie stosowane są skrypty *php* lub *asp*.

5. BIBLIOGRAFIA

1. Boudourides M., Sigrist B., Alevizos P.: Webometrics and the Self-Organization of the European Information Society. *http://hyperion.math.upatras.gr/webometrics/*, Draft Report, 1999
2. Kaliczyńska M.: Webometrics: Can we measure the Internet? Photonics Applications in Astronomy, Communications, Industry and High-Energy Physics Experiments II. SPIE Proceedings Series Vol. 5484, 2004, s. 580 – 585, ISBN 0-8194-5415-X
3. Kaliczyńska M.: Webometria, czyli co można zmierzyć w Internecie. IX Konferencja Automatyzacja i Eksploatacja Systemów Sterowania i Łączności, Gdynia 2003, s. 137 – 144, ISBN 83-87280-60-7
4. Cox T., Cox M.: Multidimensional Scaling. Chapman & Hall/CRC Press, Inc. 2001, s. 328, ISBN 1584-8809-45
5. Kruskal J. B., Wish M.: Multidimensional Scaling. Series: Quantitative Applications in the Social Sciences. SAGE Publications, 1978, s. 96, ISBN 0-803-90940-3

RESEARCH OF STRUCTURE OF THE ACADEMIC INFORMATION SOCIETY USING MDS METHOD

The article presents the methodology of webometric research and analysis aiming at determining similar features of objects belonging to the Polish information society, which uses the Internet and its WWW resources for communication purposes. In particular, the analysis applies to the selected Polish technical universities and other organizations. The research was carried out in several phases – on different data groups – with regards to the Internet space and time changes. The results have been presented in a form of two and three-dimensional topography maps. For the purposes of this analysis, the computer methods of multidimensional scaling were used. The research will be further continued for a selected group of objects over a longer time frame. Its next stage will be the research on more diversified objects, also in a multinational aspect.