Nauka - Dydaktyka - Praktyka

Veslava Osińska

Wizualizacja i wyszukiwanie dokumentów

WYDAWNICTWO SBP

Veslava Osińska - wykształcenie zdobyła na Uniwersytecie Wileńskim (magister fizyki) oraz na Uniwersytecie Mikołaja Kopernika w Toruniu (doktorat z zakresu bibliologii i informacji naukowej). Od 1998 wykłada Technologię Informatyczną i przedmioty pokrewne w Instytucie Informacji Naukowej i Bibliologii na Uniwersytecie Mikołaja Kopernika.

Główne obszary działalności naukowej autorki koncentrują się wokół zastosowania współczesnych metod wizualizacji struktur wiedzy i zasobów informacyjnych ze szczególnym uwzględnieniem dynamiki zmian w w/w dziedzinach. W projektach badawczych autorka wykorzystuje umiejętności programistyczne oraz zainteresowania plastyczne.

Veslava Osińska jest autorem szeregu publikacji w polskich i zagranicznych czasopismach naukowych, jak również dwóch podręczników dla studentów: *Technologia Informatyczna i Multimedia.*

Autorka jest członkinią Polskiego Towarzystwa Informatycznego oraz polskiej sekcji ISKO (International Society of Knowledge Organization).

WIZUALIZACJA I WYSZUKIWANIE DOKUMENTÓW

Polish Librarians Association
SCIENCE-DIDACTICS-PRACTICE

Veslava Osińska

VISUALISATION VS. DOCUMENTS' SEARCHING

WYDAWNICTWO SBP

Warsaw 2010

Stowarzyszenie Bibliotekarzy Polskich
NAUKA-DYDAKTYKA-PRAKTYKA

Veslava Osińska

WIZUALIZACJA I WYSZUKIWANIE DOKUMENTÓW

WYDAWNICTWO SBP

Warszawa 2010

«NAUKA – DYDAKTYKA – PRAKTYKA»

**Publikacja dofinansowana przez Instytut Informacji Naukowej
i Bibliologii Uniwersytetu Mikołaja Kopernika**

Recenzent
Dr hab. Wiesław BABIK

Projekt okładki
Tomasz KASPERCZYK

Redakcja techniczna i korekta
Marta LACH

ISBN 978-83-61464-36-5

CIP - Biblioteka Narodowa

Osińska, Veslava
Wizualizacja i wyszukiwanie dokumentów / Veslava
Osińska ; Stowarzyszenie Bibliotekarzy Polskich
. - Warszawa ; Wydawnictwo Stowarzyszenia
Bibliotekarzy Polskich, 2010. - (Nauka, Dydaktyka,
Praktyka ; 116)

Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich
00-335 Warszawa, ul. Konopczyńskiego 5/7 tel. (22) 827-52-96
Warszawa 2010. Wyd 1. Ark. wyd 9,5 Ark. druk. 11,5
Łamanie: Funky Worky Studio Składu Komputerowego
Druk i oprawa: MKJdruk, 15-703 Białystok
ul. Zwycięstwa 3A, tel./fax : (85) 652-52-30
e-mail: biuro@mkjdruk.com.pl

Spis treści

Wstęp 9

Rozdział 1
Wizualizacja informacji – obszar badań informacji naukowej 15
1.1. Geneza wizualizacji, pojęcie i historia 15
1.2. Nowoczesne metody oraz przykłady wizualizacji informacji 27
1.3. Wizualizacja informacji w poszukiwaniu strategii mapowania nauk (Mapping Science) 37
Rozdział 2
Przedmiot badań – klasyfikacja nauk komputerowych CCS 49
2.1. Dyscyplina: Nauki Komputerowe i jej pokrewne 49
2.2. Schemat klasyfikacji CCS ACM 57
a) Przesłanki historyczne 57
b) Rozwój schematu klasyfikacji CCS 60
c) Analiza aktualnej klasyfikacji CCS 62
d) Ontologia w klasyfikacji CCS 66
e) Aktualizacja klasyfikacji CCS 69
2.3 Systematyka przedmiotu w serwisach sieciowych 71
Rozdział 3
Opis prac badawczych 75
3.1. Objaśnienie podstawowej metodyki badań 75
3.2. Przebieg procesu podstawowej analizy danych 82
a) Narzędzia i implementacja 82
b) Etap kolekcjonowania danych 84
c) Etap przetwarzania danych 86
d) Etap wizualizacji danych 90
3.3. Interpretacja wyników 93
a) Powierzchnia sfery 93
b) Mapy 96
c) Nowa klasyfikacja 102
3.4. Rozszerzone metody obróbki danych 105
a) Modyfikacja zestawu danych 105
b) Kwantyfikacja struktury map wizualizacji 106

Rozdział 4
Implementacje praktyczne systemu wizualizacji 113
4.1. Charakterystyki czasowe domeny naukowej 113
4.2. Zastosowanie: wyszukiwanie dokumentów 123
Podsumowanie i wnioski 135
Bibliografia 141
Spis tabel z wynikami badań 151
Spis ilustracji 183
Indeks rzeczowy 184

Contents

Introduction 9

Chapter 1
Wizualizacja informacji – obszar badań informacji naukowej..... 15
Visualisation of information – a research field in information science 15
1.1 Background of visualization – the concept and history 27
1.2 Modern methods and examples of information visualization .....
1.3Information visualization – looking for a strategy of science mapping..... 37
Chapter 2
The research subject –the Computer Classification System (CCS) 49
2.1 The discipline – computer and related sciences..... 49
2.2 The CCS ACM classification scheme..... 57
a) Historical background..... 57
b) The CCS classification scheme development..... 60
c) Analysis of the current CCS classification scheme 62
d) Ontology in the CCS 66
e) The CCS updating 69
2.3 The subject's taxonomy in the Internet services 71
Chapter 3
Description of the study..... 75
3.1 Methodology 75
3.2 Process of data analysis..... 82
a) Tools and their implementation 82
b) Data collection stage 84
c) Data processing stage..... 86
d) Data visualization stage..... 90
3.3 Interpretation of the results 93
a) The sphere's surface 93
b) The maps..... 96
c) The new classification..... 102
3.4 Expanded methods of data analysis..... 105
a) Modification of data set 105
b) Quantification of structure of visualization maps 106

Chapter 4
Implementations of the visualization system 113
4.1 Time characteristics of the science domain 113
4.2 Application: documents' searching 123
Conclusions 135
References 141
List of tables with study results 151
List of illustrations 183
Subject index 184

Wstęp

Cel i zakres pracy

Przeszukiwanie współczesnych zasobów sieciowych pozostaje dla przeciętnego użytkownika wciąż ograniczone na skutek wykorzystywania danych niejednakowych systemów indeksujących, katalogujących, dokumentów wielojęzykowych i multimedialnych o różnych formatach. Dotychczas ugruntowały się dwie podstawowe metody przeglądania danych w serwisach sieciowych: przy pomocy słów kluczowych, które się wpisuje w polu wyszukiwawczym lub przy pomocy struktury tematycznej. W pierwszej metodzie problemy stwarza między innymi niekonsekwencja słownikowa pomiędzy zapytaniem użytkownika a prezentowanymi danymi; przykładem mogą być słowa-synonimy. W drugiej wykorzystuje się katalogi tematyczne, które są prezentowane w postaci drzew o zadanym poziomie hierarchii. W nawigacji tematycznej charakterystyczne są przejścia liniowe w strukturze drzewa katalogu bądź klasyfikacji. Tematy główne (klasy główne) lub nadrzędne dostępne są przy nawigowaniu w górę drzewa hierarchii; aby rozwinąć tematy podrzędne, trzeba zagłębić się w dół. Taki liniowy porządek ogranicza zakres przeszukiwania i przeglądania badanych dokumentów do stopniowo zawężanego tematu. W systemach bibliotecznych tworzy się wieloaspektową charakterystykę wyszukiwawczą dokumentów. Takim sposobem konstruuje się klasyfikacje polihierarchiczne, tzw. fasetowe. Jednostki grupujące – **fasety** – określają wspólną cechę pojęć, tematów, należących do różnych kategorii klasyfikacji. Stwarza to możliwość wertowania zasobów o tematyce bardziej lub mniej zbliżonej do wybranego początkowo. Drugim jaskrawym przykładem zapotrzebowania na nawigację szerokotematyczną jest rozwiązanie, stosowane na przykład w księgarniach internetowych, które jest skierowane na logiczne przewidywanie zainteresowań czytelnika. Do każdej książki wytypowanej w bazie generowane są pozycje w kategorii „inne książki tegoż autora" lub „osoby które kupiły tę książkę, kupowały też...".

W obliczu wymienionych wymagań użytkowników priorytetowe znaczenie przyjmuje wizualna prezentacja przestrzeni informacyjnej i jej elementów w celu ułatwienia ich szybkiego przyswojenia i zrozumienia. W wizualizacji informacji nowoczesną i sprawdzoną metodą jest przedstawienie dendrogramów czyli drzew (*tree map*). Tworzą one

pogrupowane lub poklasteryzowane tematycznie zasoby w obszary o sprecyzowanej (prostokąty, okręgi) lub nie geometrii.

Posługując się specjalistycznym schematem klasyfikacji w zakresie informatyki (klasyfikacją *Computing Classification System* autorstwa *Association for Computing Machinery – ACM*), wyznaczono za cel znalezienie odpowiedniej przestrzeni reprezentującej tematy wszystkich klas i podklas aktualnego drzewa *CCS*. Jakie były powody wytypowania danej klasyfikacji? Przede wszystkim jest to najważniejsza specjalistyczna klasyfikacja piśmiennictwa w zakresie przedmiotu nauk komputerowych. Po drugie, dynamikę zmian tego schematu powinno się analizować w odniesieniu do istotnych procesów w historii rozwoju technologii informacyjno-komunikacyjnych. W ich ocenie autorka kierowała się własnym doświadczeniem osoby uprawiającej zawód informatyka i trenera technologii informacyjno-komunikacyjnych od 1990 r. Po trzecie, uniwersum klasyfikacyjne wraz ze strukturą klasyfikacji są publiczne dostępne (z ograniczeniem do pełnych tekstów publikacji), co ułatwiło kolekcjonowanie danych. Jako podstawowy obiekt badań posłużył zbiór metadanych dokumentów biblioteki cyfrowej *ACM*, opublikowanych w 2007 r. Na podstawie przypisanej artykułom symboliki klas wyznaczono odległości tematyczne pomiędzy klasami i podklasami. Jako przestrzeń wybrano powierzchnię sfery ze względu na jej symetrię, bezbrzegowość, jak również dużą pojemność przestrzeni topologicznej, uzależnioną od promienia sfery. Otrzymane mapy sprawdzono pod względem przydatności w wyszukiwaniu dokumentów oraz przy ich pomocy przeanalizowano możliwości modernizacji klasyfikacji pierwotnej.

Badania te powtórzono dla innych okresów czasowych z odstępem co 10 lat. Pozwoliło to na obserwację zmian jakie zachodziły w naukach komputerowych na przestrzeni ich historii rozwoju. Przestrzeń specjalistycznych klasyfikacji przedmiotowych uformowana na powierzchni sfery pozwoliła na głębszą interpretację dynamiki ewolucji danej dziedziny. Co więcej, jeśli dla każdej klasyfikacji bibliotecznej skonstruujemy własną sferę, to zmapowanie powierzchni jednej na drugą pozwoli na dogłębne zbadanie różnic i zmian w strukturach klasyfikacji. Obecnie są one rejestrowane w tablicach klasyfikacyjnych, mających takie wady, jak objętość, asymetryczność, powtarzanie się terminów itp. W aspekcie współczesności zmiana topologii takich obiektów jest potrzebna zarówno bibliotekarzom, jak i użytkownikom.

Struktura rozdziałów

Niniejsza praca liczy cztery rozdziały. Wizualizacja informacji jest typowo interdyscyplinarną dziedziną, wykorzystującą osiągnięcia naukowe ostatnich lat w analizie danych, interakcji człowiek-komputer, jak również grafice komputerowej. Dlatego rozdziały pierwszy i drugi charakteryzują tę dyscyplinę z perspektywy współczesnej problematyki badań, odpowiednio w Informacji Naukowej i Naukach Komputerowych.

Rozdział pierwszy zatytułowano *Wizualizacja informacji – obszar badań informacji naukowej*. Nasz mózg wraz ze swoim systemem percepcji wizualnej jest zawsze „stacją końcową" metod wizualizacji. To właśnie w nim odbywa się najważniejsza część analizy danych tak, aby była najlepiej przydatna w kolejnych procesach kognitywnych. Pierwszy podrozdział charakteryzuje działanie ludzkiej percepcji, opisuje w jaki sposób zachodzi

przetwarzanie obrazu – ta wiedza pomaga w formułowaniu reguł wyświetlania informacji. Zasady te znajdują zastosowanie w projektowaniu współczesnych interfejsów wizualizacyjnych, tak aby prezentowały informację w sposób efektywny. W podrozdziale drugim dokonano przeglądu nowoczesnych technik wizualizacyjnych, które są implementowane w interfejsach aplikacji służących zarówno do przeglądania, nawigacji, jak i wyszukiwania danych. W zależności od typu informacji, które zostały pogrupowane w 5. kategoriach stosuje się techniki wizualizacyjne, które też są uwarunkowane wymiarem topologicznym danych i ich poziomem abstrakcji. W ostatnim podrozdziale analizowana jest historia oraz ostatnie odkrycia w poszukiwaniu metod wizualizacji wiedzy z perspektywy nauk interdyscyplinarnych. Opisane zostały też paradygmaty naukowe, towarzyszące technikom wizualizacji w procesie działania struktur intelektualnych. Ten podrozdział nabiera szczególnej wagi w momencie dyskusji na ile proponowana metoda może się przydać w wizualizacji nauk komputerowych i im pokrewnych zawartej w rozdziale podsumowującym.

Rozdział drugi poświęcono przedmiotowi badań czyli klasyfikacji nauk komputerowych *CCS*. Aktualne schematy klasyfikacyjne nauk komputerowych i technologii informacyjnej z reguły nie odwzorowują poprawnie szybko zmieniających się taksonomiczno-statystycznych własności nauk komputerowych. W rozdziale tym zagadnienie klasyfikacji bibliotecznych w obszarze zastosowań komputerowych omawiane jest w oparciu o historię rozwoju nauk komputerowych, zapotrzebowanie współczesnych użytkowników serwisów o profilu informatycznym oraz relacje ontologiczne. System klasyfikacji *Computing Classification System* (*CCS*), stworzony przez *ACM*, uważany jest na świecie za standard w identyfikacji i kategoryzacji literatury komputerowej oraz działalności badawczej w zakresie nauk komputerowych. Zostały tu szczegółowo opisane i przeanalizowane podziały tematyczne 11. klas głównych klasyfikacji *CCS* oraz przedstawiony został czteropoziomowy schemat drzewa. Omówiono tematyczną organizację informacyjnych zasobów serwisów sieciowych specjalizujących się w ogólnie pojętych zagadnieniach informatycznych. Spróbowano wskazać charakterystyczne cechy ontologii opisanych serwisów. Podkreślono znaczenie aktualizacji takiej struktury w obliczu szybko zachodzących zmian w dziedzinie nauk komputerowych.

Rozdział trzeci dotyczy przeprowadzonych prac badawczych. Zagadnieniu temu poświecono szczególną uwagę ze względu praktyczny charakter niniejszej rozprawy doktorskiej. Wyjaśniono tu koncepcję metodologii badań, opisano etapy ekspertmentu włącznie ze schematem architektury stosowanych algorytmów. W odniesieniu do dotychczasowych badań omówiono innowacyjność proponowanej metody oraz wypunktowano pozytywne strony jej zastosowania w obszarze struktur klasyfikacyjnych. Przebieg procesu analizy danych omówiony został z perspektywy dwóch faz eksperymentu w dwóch podrozdziałach: podstawowa analiza (I) oraz dodatkowe metody obróbki danych (II). W fazie pierwszej wyodrębniono trzy etapy podstawowej analizy: kolekcjonowanie danych, przetwarzanie oraz wizualizacja. Przedstawione również algorytmy obliczeń i obróbki danych. Druga faza zawiera metody rozszerzone, które usiłowano zaimplementować do badania struktury oraz dynamiki kolekcji danych, a które jednak nie wniosły wiele istotnych informacji. Wyniki wizualizacji danych prezentowane są w postaci zrzutów ekranowych końcowej sfery w wielu projekcjach oraz map kartograficznych jej powierzchni. Interpretując wyniki udowodniono, że wybrany sposób reprezentacji klas i podklas klasyfikacji *CCS* na powierzchni sfery jest właściwy.

Rozdział czwarty zawiera praktyczne implementacje stworzonego systemu wizualizacji dokumentów. Jako, że jednym z podstawowych celów danej pracy jest sprawdzenie w stworzonym systemie wizualizacji potencjału wyszukiwania informacji, w jednym z podrozdziałów rozważa się czy analizowane mapy mogą służyć jako interfejs wyjściowy wyszukiwania dokumentów. Pozytywne wyniki testów, zaprojektowanych w obu kierunkach (wizualizacja – wyszukiwanie; wyszukiwanie – wizualizacja) zachęcają do praktycznej implementacji proponowanej metody w systemach wyszukiwawczych. Skompletowano dane i otrzymano mapy wizualizacji schematu klasyfikacji dla okresu 50 lat z cyklem 10-letnim.

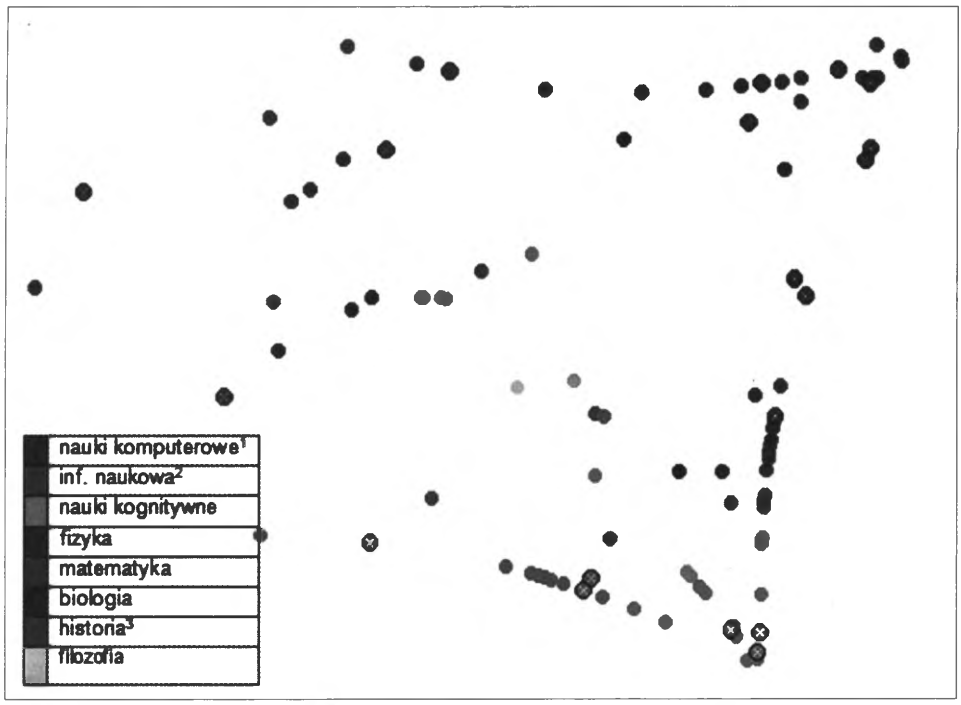
Interpretacja wyników przy wsparciu wiedzy z zakresu historii dziedziny nauk komputerowych pomogła we wnioskowaniu o ewolucji klasyfikacji *CCS* na przestrzeni tych lat. Uszczegółowienie przeprowadzonych prac badawczych oraz wnioski z nich wynikające zamieszczono w zakończeniu zatytułowanym *Podsumowanie i wnioski*.

Niniejszą rozprawę uzupełniają materiały pomocnicze, takie jak np. spisy tabel czy bibliografia. Tabele z wynikami analizy danych klasyfikacji odseparowano od podstawowego tekstu ze względy na ich układ poziomy oraz objętość, zmuszającą czytelnika do przełączenia uwagi w dłuższym czasie. Ich numeracja zaczyna się od oznaczenia literami „A", „B" itp. a odnalezienie ułatwia załączony *Spis tabel z wynikami badań*. Ilustracje (częściowo – w wersji kolorowej), zawierające wynikowe mapy wizualizacji ponumerowano osobno i umieszczono w miejscach, gdzie są do nich odwołania. Na końcu pracy przytoczony został również *Spis ilustracji*. Interdyscyplinarny charakter tekstu zmusił autorkę do wprowadzenia *Indeksu rzeczowego* odsyłającego do omówionych terminów, dołączonego na końcu pracy.

Kolorowe wersje map wizualizacyjnych i wybranych rysunków z tekstu czytelnik może obejrzeć pod adresem www.umk.pl/~wieo/Infoviz.

Mapa bibliografii

Wykorzystaną literaturę przedmiotu wykazano w *Bibliografii*. Jej zawartość dodatkowo przedstawiono poniżej w postaci mapy. Skompletowane dane prezentują wielowymiarowy charakter, co jest istotne w wyborze metody analizy i wykorzystaniu odpowiednich środków na etapie wizualizacji. W pracy posługiwano się literaturą z zakresu multidyscyplinarnego – przynależność do danej dyscypliny naukowej nie dało się określić za pomocą logiki binarnej. A więc do wieloaspektowego charakteryzowania źródeł bibliograficznych zastosowano bardzo podobną do badanej metodologię. Każda pozycja bibliograficzna przedstawiona jest za pomocą wektora o dziewięciu cechach. Pierwsze osiem reprezentowały dominację jednej z następujących dyscyplin: nauki komputerowe, informacja naukowa i bibliotekoznawstwo, fizyka, matematyka, biologia, historia i filozofia. Tylko nieliczne punkty mają „czysty" kolor, odpowiadający danej nauce (p. legendę na Mapie 1). Większość danych – o mieszanych kolorach, wskazuje na pozycje dwu- i trój-dyscyplinarne. Dziewiąty parametr wskazuje czy źródło było polskojęzyczne, a takowe stanowiły 16% całej listy (krzyżyki na punktach). Zdecydowana większość literatury na temat wizualizacji informacji jest anglojęzyczna co świadczy potrzebie takich badań w Polsce.



Mapa 1. Dwuwymiarowa mapa pozycji bibliograficznych. Krzyżyki oznaczają źródła polskojęzyczne:
[1] Włączając sztuczną Inteligencję, grafikę komputerową i edukację *Information Technology (IT)*;
[2] Obejmuje takie obszary badań jak: bibliotekoznawstwo, bibliometrię, naukometrię, zarządzanie wiedzą, wizualizacja, wyszukiwanie informacji;
[3] Jako sposób badań danego tematu.

Źródło: Opracowanie własne.

Wartości cech ustalono na podstawie tytułów, abstraktów lub danych opisów czasopism. Otrzymana mapa bibliografii wydaje się być jest czytelna i wiarygodna. Mapy bibliografii nie należy mylić z mapami domen naukowych. Powstała ona na podstawie analizy cytowań jednej pracy i na niej są odwzorowywane relacje dziedzinowe pomiędzy wykorzystywanymi publikacjami i czasopismami. W przełożeniu tych związków na interdyscyplinarny skład treści niniejszej rozprawy doktorskiej trzeba też brać pod uwagę objętość, wagę i rozbudowę wątków, powstałych na skutek studiowania konkretnego źródła. Zaprezentowana mapa sygnalizuje potrzebę opracowania metody automatyzacji procedury określania wektorów cech (czyli reprezentacji wielostronnych charakterystyk za pomocą liczb) przestudiowanych artykułów. To byłby zauważalny krok w kierunku aktualnego zapotrzebowania w świecie naukowym na „mierzenia" stopnia interdyscyplinarności.

Rozdział 1

WIZUALIZACJA INFORMACJI – OBSZAR BADAŃ INFORMACJI NAUKOWEJ

1.1. Geneza wizualizacji, pojęcie i historia

Wizualizacja jako swoista kompilacja *nauki i sztuki* od dawna jest obiektem zainteresowania naukowców, informatyków, grafików i wszystkich, którzy używają technologii komputerowych. Początkowo analizowana jako proces reprezentacji wyrazów i pojęć w formie wizualnej, rozszerzyła swój zakres zastosowań do narzędzi i metod interpretacji danych graficznych oraz generowania obrazów na podstawie zbiorów danych wielowymiarowych.

Na przestrzeni wieków, wizualne reprezentacje takie jak rysunki na starożytnych budowlach, mapy, reprezentacje geometrii euklidesowej a później diagramy statystyczne rozwijały się, aby wspomagać myślenie i wyobraźnię. Na tym podłożu wyrosły dyscypliny zajmujące się graficznym obrazowaniem: kartografia, kreślarstwo, projektowanie urządzeń technicznych za pomocą technik typu *CAD*[1], grafika, drukarstwo, projektowanie etykiet czy wizualizacja danych statystycznych. Dzisiaj graficzna prezentacja danych jest już pełnoprawną częścią produkcji telewizyjnej i filmowej. Wystarczy spojrzeć na współczesne filmy, których nieodłącznym elementem są migające komunikaty, kolorowe okienka, wykresy, nie wspominając o efektach specjalnych. Nowoczesne środki grafiki komputerowej pozwalają na estetyczną, przystępną, przejrzystą, a często zaskakującą prezentację danych, informacji,

[1] CAD (*Computer Aided Design*) – projektowanie wspomagane komputerowo. Programy typu CAD służą do obliczeń inżynierskich, rysunków konstrukcyjnych, przedstawiania rysowanych elementów w perspektywie itp. Por. *Słownik pojęć komputerowych*. Pod red. V. Illingworth i J.Daintitha. Warszawa: Świat Książki 2004, s. 44.

czy wiedzy. Powstał nowy termin określający *najbardziej udany związek słowa z obrazem*[2] – **infografika** czyli grafika informacyjna.

Problem czy „wizualizacja jest bardziej grafiką czy nauką" ciągle powraca w momentach pojawienia się nowych trendów w metodach wizualizacji bądź zaawansowanego programu wizualizacyjnego. Z kombinacji tych dwóch podstawowych perspektyw wyłania się kilka podejść do wizualizacji. Można ją badać w ramach tradycji artystycznej projektowania graficznego. Z kolei w obszarze zastosowań nauk komputerowych może ona określać algorytmy wyświetlania danych. W konstruktywnym przybliżeniu systemów symbolicznych wizualizacja może być również przydatna jako część semiotyki. Nowoczesne podejście, bazując na teorii poznania wykorzystuje w regułach projektowania aktualną wiedzę o ludzkim systemie percepcji wzrokowej.

Wizualizację w praktycznych zastosowaniach naukowych i informatycznych nazwano **wizualizacją naukową** (ang. *scientific visualization*) lub **wizualizacją danych** (ang. *data visualization*). Uformowała się ona w odpowiedzi na zapotrzebowanie świata nauki na wydajną prezentację danych eksperymentalnych w formie graficznej i późniejszą sprawną ich analizę. Od kiedy naukowcy uznali wizualizację za dziedzinę wyłaniającą się bezpośrednio z nauk komputerowych powstało dużo prac metodologicznych definiujących tę dyscyplinę badawczą poprzez jej cele. Wizualizacja naukowa jest procesem reprezentacji danych wyjściowych w postaci obrazów, który ma pomóc zrozumieć znaczenie wyników eksperymentu. Podstawowym celem takiego procesu, co podkreśla się w wielu pracach[3] jest „pogłębianie wiedzy" (z angielskiego: *insight*; inne tłumaczenie to: „wnikliwe poznawanie") poprzez mapowanie danych do podstawowowych figur graficznych: prostokątów, kwadratów, okręgów itp. Ma on również polepszać możliwości poznawcze człowieka w trakcie badania, podejmowania decyzji i eksploracji. Wizualizacja wspiera naukowców w udowadnianiu i obalaniu hipotez naukowych, w odkrywaniu nowych zjawisk fizycznych i symulacji pomiarów eksperymentalnych.

Z najnowszych definicji, które można znaleźć w literaturze naukowej, warto przytoczyć następującą, ze względu na jej aktualność, przejrzystość i ścisłość. *Wizualizacja jest to więcej niż metoda komputerowa. Wizualizacja jest to proces, w którym dane, informacja i wiedza przekazywane są formie wizualnej. Zaangażowane są w tym momencie następujące elementy: komputer do przetwarzania informacji, ekran – do prezentacji oraz ludzki mózg – do jej percepcji i analizy.*[4] Jest to proces pozwalający odbiorcy na obserwację, przeglądanie oraz zrozumienie informacji. Wizualizacja zatem to zarówno sam akt tworzenia graficznej reprezentacji danych, jak i proces logicznej analizy jej treści[5].

---

[2] M. Burns, T. Bitner: *Sztuka informowania.* Digit online [on-line] 2003, nr 6 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.digit.pl/artykuly/34291_6/sztuka.informowania.html.

[3] I. Niskanen. *An interactive ontology visualization approach for the domain of networked home environments* [on-line]. Oulu: Julkaisija-Utgivare, 2007 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.vtt.fi/inf/pdf/publications/2007/P649.pdf; E. R. Tufte. *Envisioning Information.* Connecticut, USA: Graphics Press 1990, s. 12-51; W. J. Yurcik. *Scientific Visualization* [on-line]. BookRags [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.bookrags.com/research/scientific-visualization-csci-03/scientific-visualization-csci-03.html.

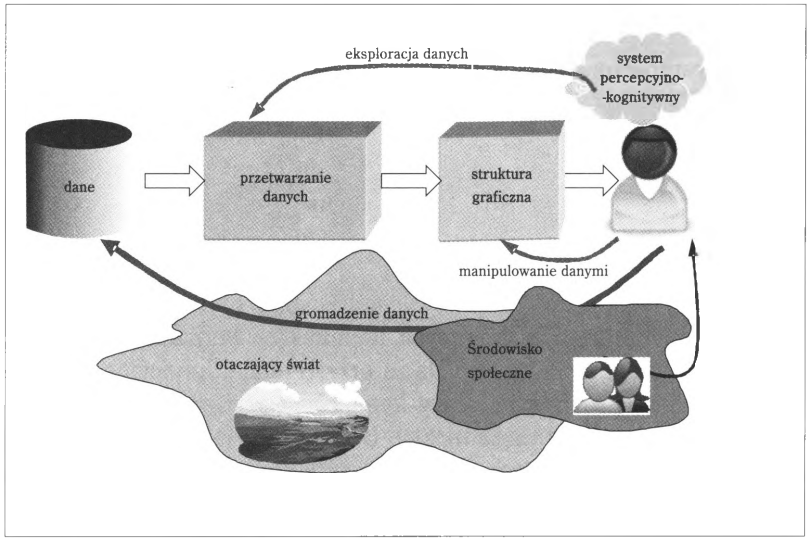
[4] *What is Visualization?* [on-line]. Infovis. Information Visualization Resources [dostęp 19 maja.2009]. Dostępny w World Wide Web: http://infovis.org/.

[5] I. Niskanen, dz.cyt.

Jak zatem wygląda proces wizualizacji? Według C. Ware, dyrektora *Data Visualization Research Lab* i autora monografii dotyczącej wizualizacji informacji składa się on z czterech podstawowych etapów z trzema pętlami oddziaływania zwrotnego[6]:

- skompletowanie zbioru danych;
- przetwarzanie danych, które prowadzi do ich konwersji w formę zrozumiałą dla człowieka;
- użycie algorytmów graficznych do wyświetlenia obrazu wizualizacji na ekranie;
- włączenie ludzkiego systemu percepcyjno-kognitywnego.

Analityk postępuje w następującej kolejności: kolekcjonuje lub mierzy serie danych i wprowadza je w stadium obróbki, która włącza konwersję formatów i filtrowanie, po czym następuje mapowanie danych zgodnie z koncepcją wizualizacji oraz renderowanie[7] w przypadku przestrzeni trzywymiarowych (Rysunek 1). Końcowy obraz ma zaktywować ludzkie systemy wizualny i kognitywny – czyli ostatni etap procesu wizualizacji informacji. Trzy pętle zwrotne to: zbieranie, manipulowanie i eksploracja danych. Najdłuższą pętlą jest kolekcjonowanie danych, kiedy naukowiec lub analityk gromadzi interesujące go informacje ze środowiska. To ostatnie jest źródłem danych, *gdy środowisko społeczne kompleksowo determinuje obiekt kolekcjonowania* – zaznacza C. Ware[8]. Manipulowanie danymi oznacza zakres kontroli prezentowania wyników odbiorcy, np. wybór kolorystyki, rozdzielczości obrazu lub transformacji geometryczno-przestrzenne obiektów badanych. Dogłębne zbadanie danych można przeprowadzić, jeśli cofniemy się do etapu przetwarzania, aby wybrać stosowną metodę transformacji – pętla eksploracji danych.



Rysunek 1. Diagram procesu wizualizacji

Źródło: Na podst. C. Ware. *Information Visualization: Perception for Design.* Wyd. 2. San Francisco: Morgan Kaufmann, 2004, s. 4.

[6] C. Ware. *Information Visualization: Perception for Design.* Wyd. 2. San Francisco: Morgan Kaufmann, 2004, s. 4-5.

[7] *Renderowanie* (ang. *rendering*) lub obrazowanie w grafice 3D: komputerowa analiza modelu danej sceny i utworzenie na jej podstawie dwuwymiarowego obrazu wyjściowego w formie statycznej lub w formie animacji. Por. *Słownik pojęć komputerowych...*, s. 271.

[8] C. Ware, dz. cyt., s. 5.

Inni badacze tej problematyki także podejmowali się wykonania schematyzacji procesu wizualizacji. E. Tufte w swym dziele *Readings in Information Visualization. Using vision to think* (1992) z perspektywy badań neuropsychologicznych i kognitywistycznych prowadzonych w ciągu ostatnich 15 lat stworzył diagram, który ilustruje etapy przekształecenia wierszy danych (ang. *raw data*) do postaci graficznej. Pierwszy etap włącza konwersję danych do tabel lub macierzy, drugi – reprezentację graficzną, w trzecim dokonuje się transformacji obrazu końcowego.

Rozważmy zatem, jak funkcjonowała ta dyscyplina badawcza w okresach przedkomputerowym i ogólnej dostępności komputerów. Badacze już od XII wieku potrafili wypracować metody graficznego odwzorowywania elementów rzeczywistości: system układu współrzędnych, wszelkiego rodzaju wykresy (liniowy, strumieniowy, rozrzutu), mapy konturowe (1594 r.), kolorowe (1741 r.) warstwowe i stereogramy (1896 r.). Mimo że mapy wykonywano odręcznie i powstanie kartografii datuje się na rok 1664, to komputery zrewolucjonizowały tę dziedzinę. Dzisiejsze mapy wysokiej jakości o zastosowaniu komercyjnym wykonuje się przy pomocy programów *CAD, GIS* (*Geographic Information System*) lub w innych systemach specjalistycznych.

Wizualizacja naukowa wykonywana za pomocą komputera rozwijała się równolegle z grafiką komputerową. W połowie lat osiemdziesiątych XX w. zaawansowane procesy informatyczne, takie jak symulacje na superkomputerach, dostarczały tak dużych ilości danych, iż wzmogło to konieczność poszukiwania nowych narzędzi i algorytmów do kompleksowej wizualizacji. Rok 1986 rozpoczął etap integracji grafiki komputerowej i nauk komputerowych. Przyczyniła się do tego Narodowa Fundacja Nauki w USA (*National Science Foundation*), która poleciła instytucjom naukowym nabywać oprogramowanie graficzne oraz odpowiedni sprzęt komputerowy. Raport z pierwszych warsztatów wizualizacji naukowej rok później głosił:

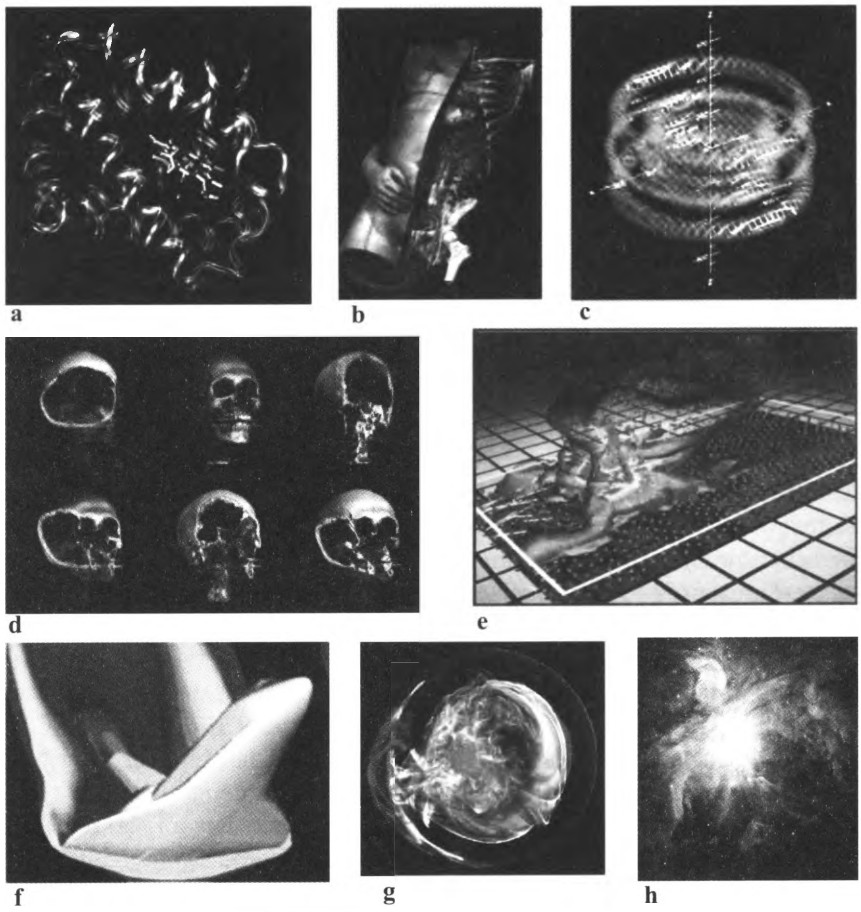
Naukowcy potrzebują alternatywy dla liczb. Wykorzystanie obrazu zamiast liczb jest technicznie realne dzisiaj i jutro stanie się niezbędnym rekwizytem wiedzy. Celem obliczeń naukowych jest poszukiwanie, nie wyliczenie. Dowiedziono, że 50% neuronów w mózgu odpowiada za widzenie. Wizualizacja w naukowych obliczeniach pomaga całą neuronową maszynerię puścić w ruch[9].

Dzisiaj już doskonale wiemy, że dwie półkule mózgu funkcjonują w różny sposób. Lewa pomaga w obliczeniach analitycznych, komunikacji werbalnej i operowaniu symbolami abstrakcyjnymi. Prawa odpowiada za przestrzenne, intuicyjne lub holistyczne myślenie. Aktywuje się w trakcie oceny kompleksowości sytuacji. Graficzne reprezentacje stymulują właśnie tę część mózgu. Używając takiego przybliżenia naukowcy otrzymują całościowy obraz danych. W późniejszych etapach do wykrywania pewnych anomalii we wzorach danych stosuje się metody bardziej analityczne.

Do wzmożonego zainteresowania wizualizacją już od końca lat dziewięćdziesiątych ubiegłego wieku przyczyniło się kilka czynników. Dostępność cenowa komputerów, kolorowych monitorów i wydajnych kart graficznych dało początek rozpowszechnieniu grafiki prezentacyjnej i oprogramowaniu do tego służącemu. Popularyzacja Linuxa oraz programów Open Source pozwoliła na rozwój technik

[9] X. Berenguer: *The synthetic image as language*. Temes de Disseny [on-line] 1991, nr 5 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://tdd.elisava.net/coleccion/5/berenguer-en.

wizualizacyjnych i programów graficznych w środowisku naukowym, gdzie wzrosło też zapotrzebowanie na zaawansowaną wizualną analizę obszernych kolekcji danych. I wreszcie, błyskawicznie taniejące monitory LCD i karty graficzne z akceleratorem 3D wspomogły dalszy rozwój metod wizualizacji przestrzennej.



Rysunek 2. Przykłady nowoczesnej wizualizacji naukowej: **a**) dynamiki molekuł; **b**) wnętrza ludzkiego ciała; **c**) chmury elektronowej; **d**) tomografia komputerowa czaszki; **e**) klimatycznej; **f**) dynamiki cieczy; **g**) symulacja wybuchu gwiazdy supernowej; **h**) międzygwiezdnego pyłu

Źródła: a, b, d, e, g) Kwan-Liu Ma. *Introduction to Visualization* [on-line]. DOE Office of Science Homepage [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.er.doe.gov/ascr/Research/scidac/intro_datavis.pdf; c) Help programu *MATLAB*; f) *Orbiter model* [on-line]. NASA [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.nasa.gov/multimedia/imagegallery/image_feature_431.html; h) *Wnętrze Wielkiej Mgławicy w Orionie (M42)* [on-line]. *teleskopy.pl* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.teleskopy.pl/obiektymglawicowe.html.

Nowoczesne technologie wizualizacyjne interesowały nie tylko naukowców, lecz także środowiska komercyjne. Informacja biznesowa potrzebowała rzetelnego przedstawienia danych, np. w organizowaniu, identyfikacji i komunikacji trendów

rynku i danych konsumenta. Wizualizacja naukowa wykorzystywana jest w astronomii do wizualizacji obiektów i zjawisk z kosmosu, których nie sposób obserwować bezpośrednio: czarnych dziur, fal grawitacyjnych i kolizji gwiazd neutronowych (Rysunek 2g, h). W fizyce technologie wizualizacyjne są obecne, kiedy bada się dynamikę płynów (*Computational Fluid Dynamics – CFD*) (Rysunek 2f) lub konstruuje się mapy rozkładu gęstości elektronów (Rysunek 2c). Fascynujące są ilustracje przedstawiające naukowe procesy w sposób zrozumiały dla zwykłego czytelnika które możemy obejrzeć w czasopismach naukowo-popularnych np. „Wiedza i Życie", „Świat Nauki". W medycynie i biologii technik wizualizacji używa się, najczęściej aby zobrazować procesy biologiczne np. rozwój komórki, dynamikę molekuł, funkcjonowanie ludzkich narządów (Rysunek 2a, b), wspomagając diagnostykę medyczną lub w przygotowaniu operacji chirurgicznych. Godną uwagi jest tomografia komputerowa, metoda diagnostyczna obrazowania ludzkiego ciała, wykorzystująca osiągnięcia wizualizacji naukowej (Rysunek 2d). Ogólnie znanym przykładem jest wizualizacja danych meteorologicznych, którą możemy obserwować w prognozach pogody – Rysunek 2e.

Zaadoptowanie metod wizualizacyjnych w różnych sferach nauki i aktywności człowieka spowodowało, że oprócz pierwotnego przeznaczenia do analizy zbiorów danych naukowych, wykształciły się inne ścisłe kierunki zastosowań wizualizacji: edukacja, komunikacja, informacja naukowa i organizacja wiedzy. **Wizualizacja edukacyjna** (ang. *Educational Visualization*) używa przede wszystkim symulacji komputerowej w celu zaprezentowania zjawisk lub tematów, których nie jest prosto zaobserwować, np. strukturę atomu, wnętrze ludzkiego ciała albo życie prehistoryczne dinozaurów[10]. W nienaukowym odniesieniu wizualizacja jest jednym ze sposobów komunikacji międzyludzkiej. Od sztuki jaskiniowej, znaków, ideogramów, alfabetu, książki, obrazów, fotografii aż po strony WWW – taką ścieżką podążała historia **komunikacji wizualnej**[11], którą w Wikipedii określa się jako *komunikację idei poprzez wizualne wyświetlanie informacji*[12].

Jeśli wizualizacja danych jest skoncentrowana wokół danych mierzalnych, takich jak wyniki medycznych badań ludzkiego ciała lub dane geograficznych systemów informacyjnych, to **wizualizacja informacji** zajmuje się danymi nierzeczywistymi czyli np. tekstem lub strukturami hierarchicznymi. W myśl ogólnej definicji, wizualizacja informacji jest wizualną prezentacją przestrzeni informacyjnych i struktur w celu ułatwienia ich szybkiego przyswojenia i zrozumienia[13]. W rzeczywistej (nie abstrakcyjnej) reprezentacji informacji wykorzystywana jest wiedza o naturalnej zdolności człowieka do szybkiego rozpoznawania obrazów. Jednak nie każdą informację da się sprowadzić do jej bezpośredniej interpretacji w świecie fizycznym.

Info*r*is (popularnie używany skrót w literaturze naukowej i biznesowej od ang. *Information Visualization*; tak również nazywają się corocznie odbywające się kon-

[10] I. Niskanen, dz. cyt., s. 40-45.

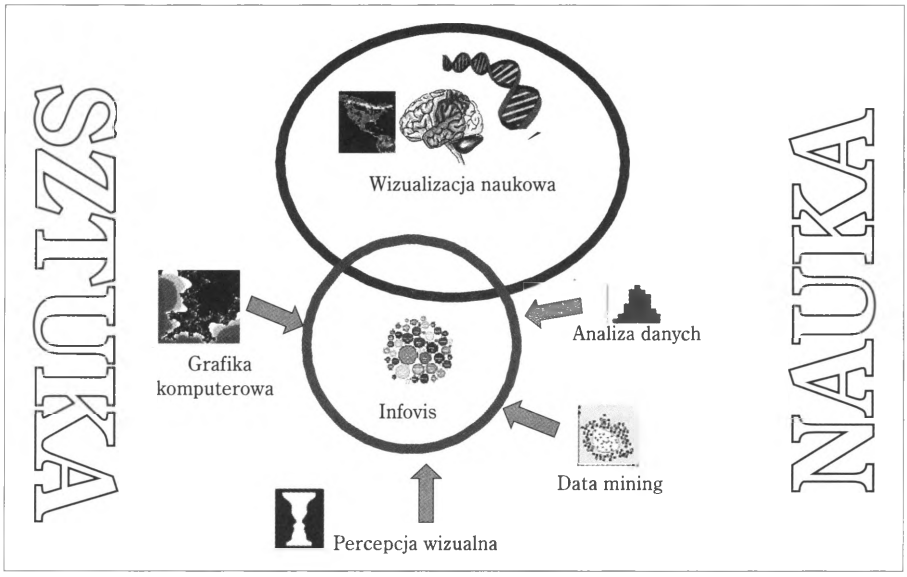
[11] *The History of Visual Communication* [on-line]. Sabanci University, Istanbul [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.citrinitas.com/history_of_viscom/.

[12] *Visual Communication*. W: *Wikipedia. The Free Encyclopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/Visual_communication.

[13] N. D. Gershon, S. G. Eick. *Guest Editors' Introduction: Information Visualization. The Next Frontier*. Journal of Intelligent Information Systems. 1998, Vol. 11 (3), s. 199-201.

ferencje pod patronatem *IEEE*[14]) jest to dyscyplina, która poszukuje nowych metafor graficznych w celu przedstawienia informacji nie mającej naturalnej i oczywistej reprezentacji. Należy nadmienić, że wizualizacja informacji jest stosunkowo młodą dyscypliną badawczą o zaledwie 10-letniej historii, intensywnie się rozwija i jest przedmiotem rozważania w niniejszej pracy. Infovis wykorzystuje osiągnięcia takich pokrewnych mu dziedzin jak: wizualizacja naukowa czy eksploracja danych, interakcja człowiek-komputer (*Human Computer Interaction*[15]), percepcja wizualna (*Vision, Visual Perception*) i grafika komputerowa. Wzajemne relacje pomiędzy wymienionymi dyscyplinami przedstawiono na Rysunku 3.

Znamiennym cytatem o Infovis jest: *Oko wypatruje podobne obiekty, aby je porównać, dokonuje ich analizy pod różnym kątem z różnej perspektywy, aby dopasować ich elementy składowe* (S. Feldman)[16]. Wizualizacja może być jednym z etapów procesu analitycznego, jeśli pozwala na szybkie wykrycie związków pomiędzy poszczególnymi cechami lub nieprawidłowych wartości tych cech. Taka analiza wizualna koncentruje się na procesach rozumowania i odkrywania sensu danych. Techniki wizualizacji są stosowane jako jedna ze skuteczniejszych form eksploracji danych (ang. *data mining*); mogą one w niektórych przypadkach wykryć więcej korelacji niż klasyczne metody statystyczne.



Rysunek 3. Infovis i dyscypliny pokrewne

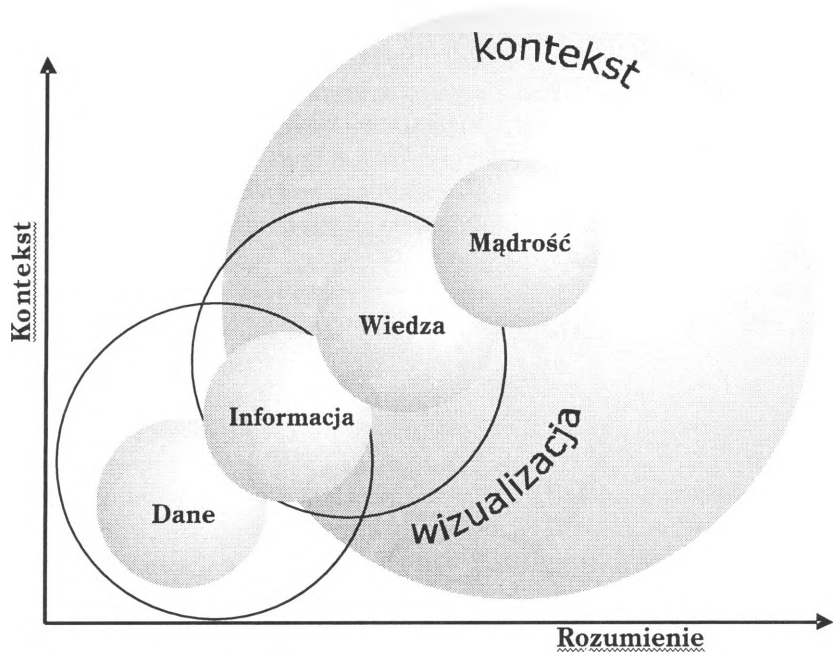
Źródło: Opracowanie własne.

[14] IEEE (Institute of Electrical and Electronics Engineers) – Instytut Inżynierów Elektryków i Elektroników – jedna z głównych organizacji skupiająca informatyków – praktyków. Ustala standardy konstrukcji urządzeń elektronicznych. Por. Concise Encyclopedia of Computer Science. Ed. by E. D. Reilly. Chichester, UK: Wiley, 2004, s. 395.

[15] Interdyscyplinarną naukę zajmująca się projektowaniem interfejsów użytkownika oraz badaniem i opisywaniem zjawisk związanych z używaniem systemów komputerowych przez ludzi.

[16] J. Luther, M. Kelly, D. Beagle: *Visualize This*. Library Journal [on-line] 2005, nr 3 (1) [dostęp 19.05.2009]. Dostępny w World Wide Web: http://www.libraryjournal.com/article/ CA504640.html.

Niektórzy badacze rozpatrują wizualizację informacji w kontekście zarządzania wiedzą jako stymulator jej zrozumienia. Wówczas definicja w nawiązaniu do tej koncepcji brzmi[17]: Infovis - jest to *proces uwewnętrznienia* (ang. *internalization*) *wiedzy poprzez percepcję informacji*. W efekcie rozumowanie można zinterpretować jako kontinuum, które rozciąga się od danych pierwotnych do mądrości poprzez informację i wiedzę, „zespolone" w procesie wizualizacji – por. Rysunek 4. Diagram ten zawiera cztery koncepcyjne koła: dane, informację, wiedzę i wreszcie mądrość jako umiejętność praktycznego wykorzystywania posiadanej wiedzy. Dane figurują jako czyste fakty, w oderwaniu od kontekstu.



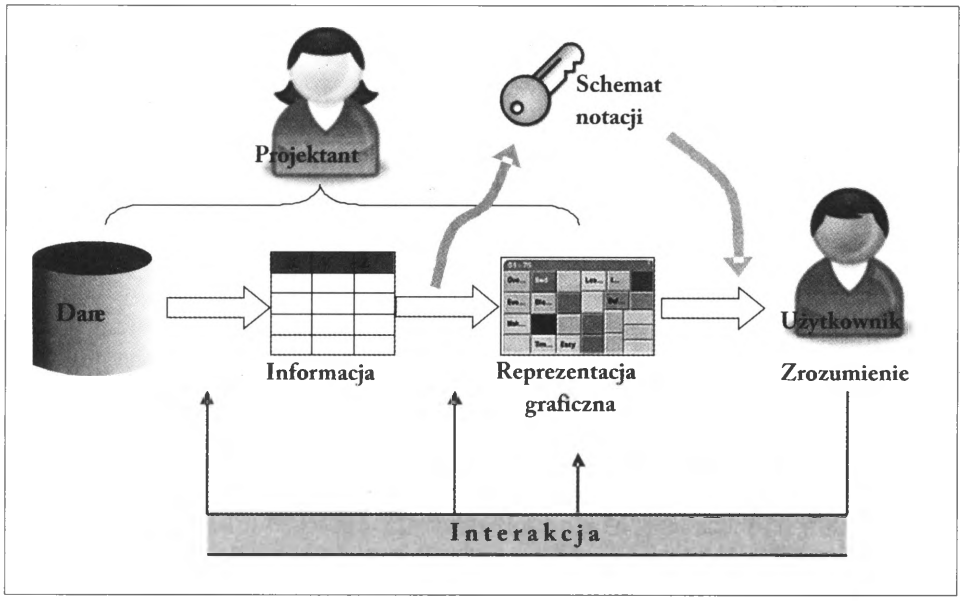
Rysunek 4. Koncept zamiany danych wejściowych w wiedzę i mądrość

Źródło: Na podst. J. C. Dürsteler. *Diagrams for Visualisation*. The digital magazine of InfoVis.net [on-line] 2007, nr 186 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.infovis.net/printMag.php?num=186&lang=2.

Dane przekształcają się w informację pod warunkiem że przenoszą informację, którą zdołamy zrozumieć i która jest dla nas wartościowa. Informacja – to są dane wzbogacone o kontekst; zamienia się ona w kolejny element konceptualny – wiedzę. W słowniku termin „wiedza" jest określony jako „zasób zorganizowanych informacji zaopatrzony w szeroki kontekst" oraz „wiedzę nabywa się poprzez badanie lub doświadczenie". Wizualizacja jako konstrukcja w umyśle rozpościera się poza percepcję sensoryczną, przez co zbliża się do wiedzy. Staje się intelektualnym ujęciem obiektów. Zrozumieć oznacza otoczyć, zawierać, uwewnętrznić.

[17] J. C. Dürsteler. *Diagrams for Visualisation*. The digital magazine of InfoVis.net [on-line] 2007, nr 186 [dostęp 19.05.2009]. Dostępny w World Wide Web: http://www.infovis.net/printMag.php?num= 186&lang=2.

W oparciu o takie abstrakcyjne budulce jak fenomen percepcji i wiedzę postarajmy się odtworzyć pełniejszy diagram procesu wizualizacji informacji. Propozycja autora pracy (Dürsteler 2008) włącza nowe aspekty. Po pierwsze, wprowadzono wizualny język kodowania – schemat notacji, aby ułatwić użytkownikowi wnioskowanie. Po drugie, wizualizację potraktowano dwojako: zarówno jako proces jej tworzenia, jak i analizy prezentowanych treści graficznych. Na Rysunku 5 widzimy trzy fazy przejściowe: dane → informacja, informacja → reprezentacja wizualna, reprezentacja wizualna → zrozumienie. Na pierwszym etapie – konwersji danych w informację, wykonywane są trzy czynności: gromadzenie i przechowywanie danych, obróbka danych oraz organizacja danych stosownie do ich znaczenia, np. konstruowanie tabel lub macierzy. W fazie transformacji informacji do jej graficznej reprezentacji ważna jest możliwość późniejszej identyfikacji struktury graficznej i percepcyjnej od strony użytkownika. Pomocnym elementem może tu się okazać schemat notacji – taki język wizualny, przedstawiający strumienie informacji za pomocą symboli graficznych.



Rysunek 5. Diagram procesu Infovis

Źródło: Na podst. J. C. Dürsteler. *Diagrams for Visualisation.* The digital magazine of InfoVis.net [on-line] 2007, nr 186 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.infovis.net/printMag.php?num=186&lang=2.

W końcowej fazie wizualizacja ma zmusić użytkownika do percepcji, a następnie do wnioskowania. A zatem projektant interfejsu musi wziąć pod uwagę wpływ takich czynników jak: percepcja wizualną, psychologia kognitywna, lingwistyka i inne. Odbiorca powinien mieć możliwość interakcji z programem, np. modyfikować formę organizacji danych, generować nowe dane lub struktury i manipulować parametrami reprezentacji graficznej.

Infovis w swoim założeniu ma służyć przede wszystkim użytkownikowi, zwiększać jego zdolności percepcyjne – to jest główna różnica w porównaniu z zadaniami wizualizacji naukowej. Właśnie teraz, kiedy dociera do nas nadmiar informacji, w różnym stopniu przydatnej oraz czytelnej dla zainteresowanego odbiorcy, priorytetowe znaczenie ma skuteczna metoda wizualizacji. Ma ona zapewnić użytkownikowi pełną kontrolę nad procesem wyszukiwania informacji (ang. *Information Retrieval – IR*). W obliczu problemów przed jakimi staje społeczeństwo informacyjne, aplikacje Infovis maja być projektowane z uwzględnieniem trzech zasad[18]. Po pierwsze efektywny interfejs programu z zaimplementowanymi algorytmami wizualizacyjnymi ma wykonywać realizację takich zadań jak: obserwacja, wyszukiwanie, nawigacja, rozpoznanie, filtrowanie danych, oraz rozumienie i odkrywanie ukrytych wzorów korelacji. Projektanci Infovis za cel stawiają ergonomię systemów informacyjnych. Wiedza o działaniu ludzkiej percepcji i tego w jaki sposób zachodzi przetwarzanie obrazu pomaga w formułowaniu reguł wyświetlania informacji, które są niezbędne w budowaniu interfejsów graficznych. Kolejny podrozdział przybliży problematykę badań neuropsycho-logicznych i kognitywistycznych.

Drugą zasadą jest trzymanie się walorów estetycznych w prezentacji informacji. Wizualizacja wykształciła się po części z grafiki komputerowej, w związku z czym potrafiła zaadoptować nowoczesne środki technologii graficznych. Ładna i przejrzysta wizualizacja skutecznie przykuwa uwagę odbiorcy.

I wreszcie ważną rolę w programie odgrywa komunikacja z użytkownikiem. Funkcja interaktywna, a o niej tu mowa, stwarza możliwość manipulowania danymi na bieżąco. A parametry, na które użytkownik może wpłynąć, to miedzy innymi: zakres danych, rozdzielczość, kąt widzenia i perspektywa (3D), sposób sortowania, filtrowania, kolorystyka obrazu itp.

Projektowanie graficzne ewoluowało skokowo, równolegle z rozwojem technologii komputerowych[19]. Ta dynamika znajduje odwzorowanie w przebiegu zmian w stosowanych metodach wizualizacji. Dzisiaj ogólną tendencją jest projektowanie interfejsów 3D. Można spróbować wymienić przyczyny powodujące, że wizualizacja (włącznie z Infovis) w XXI wieku przybiera nowe kompleksowe oblicze:

- zaawansowany sprzęt komputerowy i rozwój technologii informatycznych;
- wzrastające rozdzielczość i złożoność struktur danych;
- rozwój grafiki 3D;
- rozwój gier komputerowych i technologii rzeczywistości wirtualnej;
- postępy w robotyce i systemach sztucznej inteligencji;
- zaawansowany stan badań nad ludzką percepcją;
- społeczny Web 2.0;
- używanie semantyki i zapowiedź semantycznego WWW.

Współczesne aplikacje do wizualizacji można podzielić na trzy grupy według przeznaczenia i typów danych: matematyczne dla danych numerycznych (*Matlab, Mathematica*), naukowe dla danych przestrzennych (*IDL, Insight, Arcgis*) i informacji naukowej dla danych symbolicznych, np. ze stron internetowych. W sieci moż-

[18] J. C. Dürsteler: *InfoVis Diagram...*

[19] W. Osińska: *Dynamika historycznego rozwoju stron WWW*. Biuletyn EBIB [on-line] 2007, nr 7 (88) [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.ebib.info/2007/88/.

na bez większego wysiłku znaleźć imponującą ilość darmowych aplikacji Infovis. Do ich historycznego przeglądu przeznaczony został podrozdział 1.2. Wizualizacja informacji cieszy się stałym zainteresowaniem nie tylko informatyków, grafików, specjalistów informacji naukowej, lecz także użytkowników Internetu. Świadczą o tym liczne czasopisma elektroniczne (np. *Infovis*, *Information Visualization*, *International Journal of Human-Computer Studies*), a także portale sieciowe z zasobami oferujące źródła programów, moduły, dyskusje i programy on-line do testowania sposobów wizualizacji. Dzisiejszym trendem w Infovis jest skierowanie na masowego odbiorcę. Pod hasłami „infografika dla mass" i „wiele oczu" (*Infovis for Masses*[20], *Many Eyes*[21]) odbyła się konferencja Infovis w roku 2007 w Sacramento, USA.

Eksperci Infovis, np. Ch. Chen[22] – główny redaktor czasopisma *Information Visualization* i wiodący specjalista w wizualizacji zasobów bibliotek cyfrowych, sygnalizują konieczność powstania rozwiniętej teorii wizualizacji. Infovis jest młodą subdyscypliną, która rozwija się niejednorodnie. Literatura bibliograficzna przedstawia szerokie spektrum zaimplementowanych metod lub technik, gdzie trudno o solidną systematyzację. Brakuje jednolitej teorii, aby móc dokonać ewaluacji danego schematu reprezentacji. Wobec tego na razie podstawowym wymogiem jest zapewnienie aplikacjom, opisanych powyżej funkcji w interakcji z użytkownikiem, do którego należy ocena końcowa.

Kolor i kształt w projektowaniu informacji

Kolor jest niezbędny w wizualizacji informacji, ponieważ monochromatyczne kodowanie danych jest nieefektywnym wykorzystaniem zasobów percepcyjnych człowieka. Kolorowa wizja udowodniła swoją przydatność w procesach ewolucji. Na przykład pomagała ujawniać kamuflaż, wynajdywać użyteczne obiekty, odczytywać świeżość pożywienia itp. Taka naturalna rola koloru sugeruje, że poprawniej byłoby traktować go jako atrybut obiektu niż jego pierwotną charakterystykę, przez co idealnie się nadaje zadań kategoryzacji i etykietowania. U podstaw percepcji kolorów leży trójchromatyczna teoria, według której informacja o kolorze przenoszona jest w trzech kanałach. Przyczyna tkwi w budowie ludzkiego oka: otóż światłoczułe receptory siatkówki, nazywane czopkami są trzech rodzajów. Każdy ma inną charakterystykę widmową, czyli reaguje na światło z innego zakresu barw. Pierwszy rodzaj reaguje głównie na światło czerwone (ok. 700 nm), drugi na światło zielone (ok. 530 nm) i ostatni na światło niebieskie (ok. 420 nm). Wyróżnienie tych trzech rodzajów czopków wpłynęło na opracowanie modelu kolorów RGB. Oprócz ludzi takim trójchromatycznym widzeniem charakteryzują się niektóre naczelne. Badania neuropsychologiczne postrzegania barw przez ludzi i małp wykazały ciekawe właściwości i zależności. Do zapamiętania łatwiejsze są kolory zbliżone do „idealnych",

[20] Strona konferencji InfoVis 2007: *InfoVis 2007 Welcome* [on-line] [dostęp 19 maja 2009]. Dostepny w World Wide Web: http://conferences.computer.org/infovis/infovis2007/.

[21] Portal „publicznej wizualizacji": *Many Eyes* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://manyeyes.alphaworks.ibm.com/manyeyes/.

[22] *Chaomei Chen's Homepage* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web http://www.pages.drexel.edu/~cc345/.

za które się uważa następujące: czerwony, zielony, niebieski, żółty, różowy, morski, pomarańczowy, fioletowy oraz biały i czarny. W kodowaniu kolorowej informacji kluczowym może się okazać nie tylko kontrast, lecz i luminancja. Z perspektywy wizualizacji danych obrazy lub mapy powinny być konstruowane w palecie kanałów „kolorów przeciwnych"[23]: czerwony-zielony albo żółty-niebieski. Da się też zauważyć trend projektowania współczesnych interfejsów aplikacji infovis w dwóch gamach kolorów: ciepłej lub zimnej.

Możemy uzyskać wizualizacje wielowymiarową posługując się różnymi mapami kolorów. Kolory są dobrze rozpoznawalne i klasyfikowane w korze mózgowej człowieka i dzięki właściwej manipulacji ich parametrami – nasyceniem, kontrastem i głębią – możemy uzyskać bardzo dobre własności percepcyjne prezentowanych wizualizacji. Na korzyść koloru świadczą wyniki ewaluacji percepcji klastrów, gdzie za pomocą mapowania kolorem udało się zwiększyć liczbę wymiarów obserwowanych danych.

W symbolizowaniu danych zamiast dobrze znanych kropek i kółek z powodzeniem używa się różnego rodzaju **glify**[24]. W zależności od ilości takich symboli w zbiorze i czasu wyświetlanej sceny, jesteśmy w stanie zapamiętać od trzech do siedmiu jednostek.

Atrybuty i zmienne graficzne **Tabela 1.**

Zmienna graficzna	Wymiarowość
Pozycja przestrzenna	3 wymiary: X,Y,Z
Kolor	3 wymiary zgodnie z teorią trójchromatyczności
Kształt	2 ub 3 w zależności od otoczenia
Orientacja	3 wymiary, w zależności od układu współrzędnych
Tekstura powierzchni	3 wymiary: orientacja, rozmiar i kontrast
Ruch	2 lub 3, może być przydatna faza
Miganie	1 wymiar (używa się wspólnie ze zmienną ruchu)

Źródło: Na podst. C. Ware. *Information Visualization: Perception for Design*. Wyd. 2. San Francisco: Morgan Kaufmann, 2004, s. 183.

Badania nad detekcją glifów świadczą, że powinno się je rozmieszczać w sąsiedztwie własnym, lecz z dala od innych obiektów. Uważa się, że ludzie postrzegają obrazy świata, sprowadzając je do określonych symboli, gdzie znaczenie ma np. kolor i kształt. Eksperymentalnie wykryto cechy istotne w projektowaniu informacji, i które są przydatne w wizualizacji kategoryzacji i klasyfikacji obiektów to:
- forma: orientacja, szerokość i rozmiar linii; *współliniowość*[25], krzywizna, grupowanie przestrzenne, rozmycie, znakowanie, liczebność;
- kolor: jaskrawość, jasność;

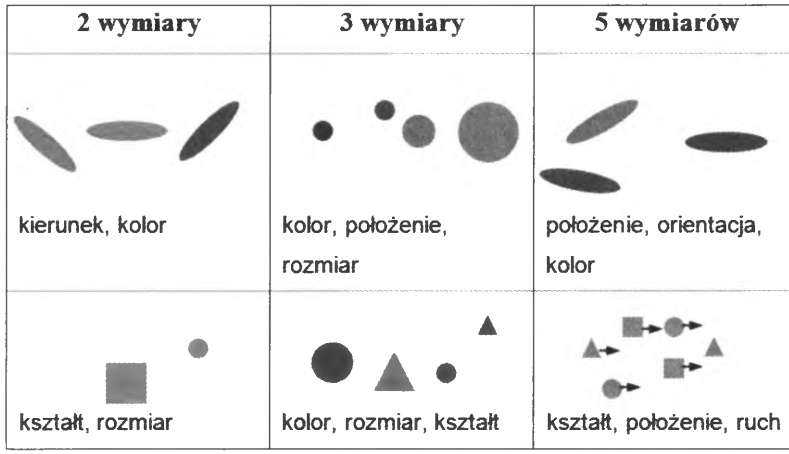
[23] M. L. Hurvich, D. Jameson. *An opponent-process theory of color vision*. Psychological Review 1957, Vol. 64 (6), s. 384-404.

[24] *Glyf* – obiekt graficzny, w typografii określany jako kształt przedstawiający w określonym kroju pisma konkretny graf lub symbol.

[25] *Współliniowość* – cecha, która determinuje czy dane punkty leżą do jednej krzywej.

- ruch: kierunek, migotanie;
- pozycja przestrzenna: 2D, głębokość stereoskopowa, wypukłość/wklęsłość.

Widzimy, iż do mapowania danych wielowymiarowych oprócz koloru mogą także służyć orientacja, rozmiar, pozycja elementów, tekstura, prosty ruch. Są to niskopoziomowe kanały informacji, działające osobno. Problem jest zwykle związany z wymiarem mapowanych danych. Tabela 1 prezentuje podstawowe atrybuty graficzne, które się stosuje w projektowaniu glyfów. Jak efektywnie zakodować dane wielowymiarowe najlepiej zobaczyć na praktycznych przykładach. Rysunek 6 ilustruje glify kodowane według zadanej ilości cech; opisane są również atrybuty mapujące zmienne.



Rysunek 6. Przykłady glifów kodowanych według zadanej ilości cech

Źródło: Na podst. C. Ware. *Information Visualization: Perception for Design.* Wyd. 2. San Francisco: Morgan Kaufmann, 2004, s. 181.

Wizualizacja wielowymiarowych danych staje się jednym z ważniejszych czynników decydujących o właściwym zrozumieniu danych. Percepcja wizualna informacji złożonej posługuje się najbardziej naturalnymi drogami przetwarzania w ludzkim mózgu, wykorzystując również najstarszy i z tego względu najbardziej stabilny limbiczny szlak przetwarzania informacji.

1.2. Nowoczesne metody oraz przykłady wizualizacji informacji

Stosowane w aplikacjach techniki wizualizacyjne uwarunkowane są przede wszystkim rodzajem informacji, jej wymiarem i poziomem abstrakcji. Do klasycz-

nych technik zalicza się podstawowe formaty wizualne pośredniczące w przekazywaniu wiedzy, np. diagramy, listy, wykresy, tabele i macierze.

Dokładniejsze przyjrzenie się metodom wizualizacji warto jest rozpocząć od określenia rodzajów informacji, jaką ludzie potrafią zinterpretować. Poniżej zostaną scharakteryzowane wyróżniane typy informacji wraz z przykładami ich prototypów przestrzeni informacyjnej[26]:

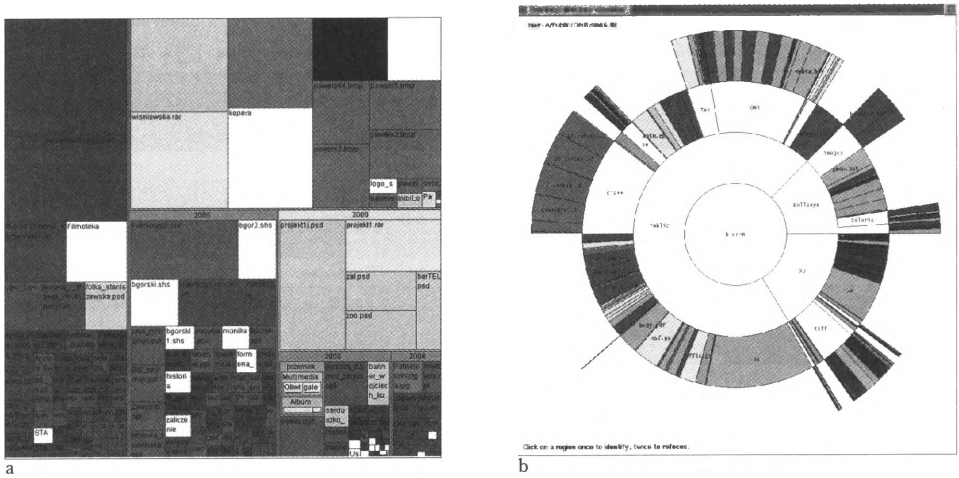
- Liniowa: listy alfabetyczne, chronologiczne, tabele, kody programów;
- Hierarchiczna: drzewa klasyfikacji, hierarchie struktury plików/katalogów;
- Sieci: topologie sieciowe, struktury **grafów**[27], sieci semantyczne;
- Wielowymiarowa: dane kompleksowe, metadane, takie jak: typ, rozmiar, autor dokumentu itp.;
- Przestrzenie wektorowe: reprezentacja dokumentów za pomocą macierzy liczb w zagadnieniach wyszukiwania informacji (*Information Retrieval*),
- Przestrzenne: mapy topologiczne, obrazy 2D lub 3D, modele w systemach *CAD*.

Najprostszym typem informacji jest informacja liniowa, składająca się z sekwencji liczb i cyfr. Dane w postaci różnego rodzaju list i tabel, powszechne w historii piśmienniczej i obliczeniowej ludzkiej działalności, znane są jeszcze z czasów starożytnych. Znaki alfanumeryczne trudno jest przedstawić w innej formie niż tekst, np. graficznej. Nie przeszkadzało to, aby w latach dziewięćdziesiątych XX w. inżynierowie wiodących koncernów programistycznych poszukiwali nowych, na miarę ówczesnego rozwoju technologicznego, rozwiązań wizualizacji danych liniowych. Wartości liczbowe w tabelach zastąpili oni odpowiednią ilością kolorowych pikseli, w ten sposób powstawały kolorowe spektra, przedstawiające zależności co najwyżej dwóch wartości. Za przykłady mogą posłużyć wizualizacje wyników sondaży ankiet i statystyk odwiedzin portali sieciowych.

Informacja hierarchiczna jest najliczniejszą, wytypowaną grupą danych, ponieważ większość współczesnej informacji interpretowana jest poprzez struktury hierarchiczne. Hierarchia jest obecna w organizacji systemów katalogów i plików, bibliotecznych systemach klasyfikacji, danych genealogicznych, a również w definicjach klas języków programowania zorientowanego obiektowo. Hierarchiczne struktury drzewiaste najczęściej są prezentowane za pomocą **dendrogramów** (z greckiego: *dendron* – drzewo, *gramma* – rysować). Dendrogram w istocie przypomina rozgałęzione drzewo z tą różnicą, że korzeń (element główny) umieszczony jest zazwyczaj u samej góry, a liście (elementy najniższego poziomu) – na samym dole drzewa. Obiekty w dendrogramie łączone ze sobą za pomocą relacji pod- i nadrzędnych (ang. *parent-child*).

[26] V. Osińska: *Przybliżenie semantyczne w wizualizacji informacji w Internecie i bibliotekach cyfrowych.* Biuletyn EBIB [on-line] 2006, nr 7 (77) [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.ebib.info/2006/77/osinska.php.

[27] *Grafy* – w matematyce: struktury składające się z wierzchołków i krawędzi; są wykorzystywane powszechnie w algorytmice.



a b

Rysunek 7. Strategia prostokątna (**a**) i pierścieniowa (**b**) wizualizacji zasobów katalogowych

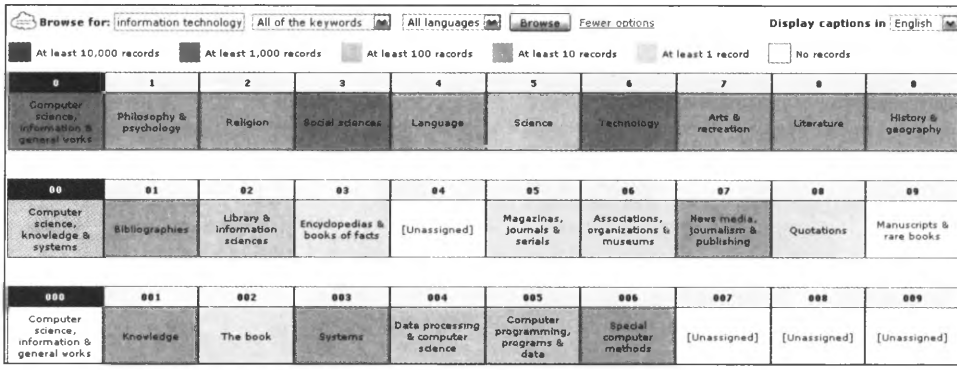
Źródła: **a**) Opracowanie własne: katalog zawiera pliki zaliczeniowe studentów, kolor indykuje format pliku, np. zielony – XLS, niebieski – DOC, brązowy – MDB, pole prostokąta – wskazuje na rozmiar pliku; **b**) J. Stasko. *SunBurst* [on-line]. College of Computing, Georgia [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.cc.gatech.edu/gvu/ ii/sunburst/ – ta sama zasada wizualizacji, lecz inna topologia.

Na początku lat dziewięćdziesiątych ubiegłego wieku. szybkość procesorów nie nadążała za dynamiką zwiększania zasobów na twardych dyskach. Dlatego inżynierowie i naukowcy intensywnie poszukiwali nowych, wydajnych metod wizualizacji struktur katalogowych. Drzewa hierarchiczne przedstawiano nie w postaci gałęzi, lecz map – topologię jednowy-miarową poszerzono do dwóch wymiarów. Generację oprogramowania, zapoczątkowanego przez B. Shneidermana[28], służącego do takich zadań nazwano **TreeMap**[29]. Rozwiązanie to opiera się na zagnieżdżaniu prostokątów mniejszymi prostokątami o polach proporcjonalnych do pojemności zasobów folderów, co ilustruje Rysunek 7a. Kolejnym pomysłem na przeniesienie struktury drzewa katalogowego na dwuwymiarową przestrzeń jest schemat hierarchii kreślony za pomocą koncentrycznych pierścieni, np. program pod nazwą *SunBurst* autorstwa J. Stasko[30]. Katalog główny znajduje się w środkowym kole mapy, segmenty kolejnych kół reprezentują podkatalogi z ich zawartością. Takie cechy jak ogólna pojemność katalogu i typ pliku identyfikowane są odpowiednio za pomocą kąta segmentu i koloru (Rysunek 7b).

[28] B. Shneiderman: *Treemaps for space-constrained visualization of hierarchies* [on-line]. University of Maryland. Department of Computer Science [dostęp 15 marca 2009]. Dostępny w World Wide Web: http://www.cs.umd.edu/hcil/treemap-history/.

[29] Program w wersji demo jest dostępny pod adresem http://www.cs.umd.edu/hcil/treemap/#download.

[30] J. Stasko: *HCC Education Digital Library: Information Vizualization.* [on-line] [dostęp 19.05.2009]. Dostępny w World Wide Web: http://hcc.cc.gatech.edu/ taxonomy/cat.php?cat=86.



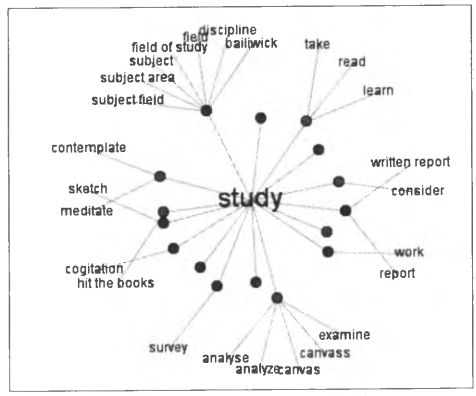
Rysunek 8. Interfejs aplikacji on-line autorstwa *OCLC DeweyBrowser* do przeszukiwania zbiorów bibliotecznych. Kolorowe komórki tabel są odpowiednikami klas i podklas trzech poziomów klasyfikacji KDD.

Źródło: V. Osińska. *Przybliżenie semantyczne w wizualizacji informacji w Internecie i bibliotekach cyfrowych*. Biuletyn EBIB [on-line] 2006, nr 7 (77) [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.ebib.info/2006/77/osinska.php.

Poziomy struktury hierarchicznej w innych projektach przedstawiano również w postaci oddzielnych tabel. Korporacja *OCLC* (*On-line Computer Library Center*) nadzorująca rozwój klasyfikacji Dewey'a udostępniła na swoim portalu eksperymentalne oprogramowanie pod nazwą *DeweyBrowser*[31], które umożliwia użytkownikom wyszukiwanie i przeglądanie zasobów bibliotecznych zorganizowanych zgodnie z klasyfikacją KDD. W tej aplikacji używa się hierarchii tabel do reprezentacji trzech górnych poziomów klasyfikacji. Rysunek 8 ilustruje rzut ekranowy programu w odpowiedzi na zapytanie „Information Technology".

Informacja o liczebności zbiorów w każdej z klas i podklas przekazywana jest za pomocą kolorów. W odróżnienie od dendrogramu, w sieciowych strukturach powiązania istnieją nie tylko w kierunku góra-dół, lecz także pomiędzy węzłami równorzędnymi. Powszechnie obserwujemy, że wiele domen rzeczywistości takich jak systemy danych geograficznych przedstawia się za pomocą węzłów i wektorów, czyli grafów. Liczne przykłady można znaleźć w aplikacjach sieciowych: hiperłącza w dokumentach WWW, mapy powiązań wyrazów bliskoznacznych w tezaurusach, relacje pomiędzy tabelami w bazach danych, algorytmy, procesy technologiczne i logistyczne, struktury organizacyjne firm, scenariusze lekcyjne itp.

[31] *OCLC DeweyBrowser Beta v.1.0*. OCLC Online Computer Library Center [on-line] [dostęp 19.05.2009]. Dostępny w World Wide Web: http://deweybrowser.oclc.org/ddcbrowser/wcat.



Rysunek 9. Mapa tezaurusa dla wyrazu „study" w programie *VisualThesaurus* autorstwa firmy *ThinkMap*. Kolory punktów czerwony, niebieski i żółty wskazują rzeczowniki, czasowniki i przymiotniki odpowiednio.
Źródło: *Thinkmap Visual Thesaurus* [on-line] [dostęp 19.05.2009]. Dostępny w World Wide Web: http://www.visualthesaurus.com/.

Rozwiązania map grafopodobnych wykorzystują interfejsy programów edukacyjnych. Oprogramowanie *Visual Thesaurus*[32] jest zintegrowanym słownikiem i tezaurusem w zakresie języka angielskiego. Interaktywny interfejs pozwala użytkownikowi na naukę poprzez eksplorację wyników zapytania. Na Rysunku 9 przedstawiony jest zrzut ekranowy wersji on-line programu. Diagram przedstawia mapę powiązań wyrazów bliskoznacznych dla słowa „study". Kolory kółek czerwony, niebieski i żółty są zarezerwowane do oznakowania rzeczowników, czasowników i przymiotników odpowiednio. Linie ciągłe łączą wyrazy - synonimy. Według zamysłu autorów, studenci i uczniowie za pomocą tego narzędzia mogą nie tylko nauczyć się nowych słów i pojęć, lecz ulepszyć swoje umiejętności czytania, pisania i komunikacji.

Informacja wielowymiarowa stanowi najodpowiedniejszy zasób danych dla badań nad strukturami semantycznymi. Metadane niosą informację o danych dokumentu i jednocześnie zawarte są w samym dokumencie. Według standardu *Dublin Core*[33] do metadanych należą informacje o tytule, autorze, wydawcy dokumentu, słowach kluczowych, opisie, języku itp. Dokumenty WWW przechowują te parametry w polach meta, opisywanych za pomocą znaczników.

Wraz z sukcesem wyszukiwarki Google firmy komercyjne intensywnie rozwijające oprogramowanie wizualizacyjne, takie jak: *KartOO*[34], *Grokker*[35], *AquaBrowser*[36], *Vivisimo*[37] w celu pozyskania nowych klientów, zaczęły profilować swoje produk-

[32] Wersja demo dostępna pod adresem: : http://www.visualthesaurus.com [dostęp 19 maja 2009].

[33] *Dublin Core (Dublin Core Metadata Element Set, DC)* – ogólny standard metadanych do opisu zasobów (np. bibliotecznych). DC definiuje 15 prostych elementów, np. tytuł, autor, słowa kluczowe, data itp. Wykorzystywany jest w Documnet Type Definition, formatach RDF, XML. Por. *Dublin Core*. W: *Wikipedia. The Free Encyclopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Dublin_Core.

[34] Wyszukiwarka dostępna pod adresem: www.kartoo.com [dostęp 19 maja 2009].

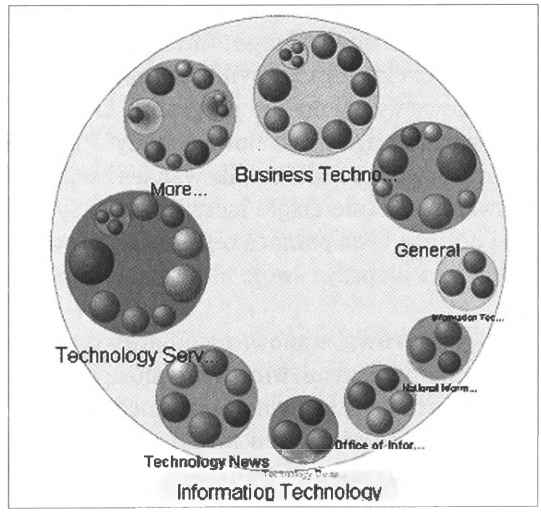
[35] Wyszukiwarka dostępna pod adresem: http://www.grokker.com/ [dostęp 19 maja 2009].

[36] Wyszukiwarka *AquaBrowser* dostępna pod adresem: www.medialab.nl/ [dostęp 19 maja 2009].

[37] *Vivisimo search engine* dostepny poda adresem: http://vivisimo.com/ [dostęp 19 maja 2009].

ty w kierunku integracji zadań wyszukiwania i nawigacji[38]. W zależności od koncepcji autorów i zastosowanych metod wizualizacji użytkownik może zapoznać się nie z listą rankingową, lecz z wielowymiarową przestrzenią nawigacyjną. Zgodnie z założeniem większej swobody w nawigacji, powinien on również mieć możliwość kolekcjonowania wyselekcjonowanych elementów. Tu można przytoczyć analogię do koszyka zakupów w sklepie internetowym. W takich wielowymiarowych mapach odrębne znaczenie przyjmują kolor, kształt, rozmiar, pozycja oraz połączenia obiektów.

Firma *Groxis*, działająca od 2001 r. zaprojektowała wyszukiwarkę z graficznym interfejsem *Grokker*, której w mediach przepowiadano konkurencyjną przyszłość wobec Google. Aplikacja ta korzysta z baz danych serwisów Yahoo, *ACM Digital Library*[39] i/lub *Amazon Books*[40]. Kolorowe koła wewnątrz innych kół (mogą to być też kwadraty) są odpowiednikami klas i podklas (Rysunek 10). Przy najechaniu myszką na obiekt w polu objaśnienia wyświetlane są metadane dla wybranego zasobu, takie jak tytuł, autor, czas utworzenia itp. Użytkownicy mogą posortować wyniki według dziedziny oraz zachować je do późniejszego wykorzystania. Zastosowanie filtrów powoduje zawężenie wyników. Ciekawostką jest to, że *Grokker* jest w stanie pokategoryzować pliki z naszego dysku według zawartości, pomijając informację o przynależności do folderów.



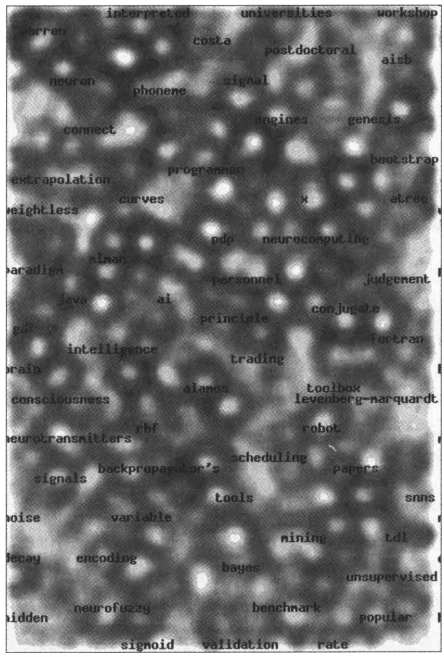
Rysunek 10. Wygenerowana mapa skojarzonych tematycznie obszarów z wyrażeniem „Information Technology" w wyszukiwarce *Grokker*.

Źródło: *Information Technology* [on-line]. Grokker, Inc. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://live.grokker.com/grokker.html?query=information%20technology &OpenSearch_Yahoo=true&Wikipedia=true&numResults=250.

Niejednorodna struktura bardzo licznych zasobów w sieci wymaga od systemów wizualizacji umiejętności wykrywania i reprezentowania złożoności tych danych. Pierwszym krokiem do wizualnej analizy dużych zbiorów danych jest automatyczna klasteryzacja zgodnie z miarą ich podobieństwa. W tym podejściu używa się staty-

[38] J. Luther, M. Kelly, D. Beagle, dz. cyt.
[39] Strona dostępna pod adresem: http://www.acm.org/dl [dostęp 19 maja 2009].
[40] Strona dostępna pod adresem: http://www.amazon.com [dostęp 19 maja 2009].

styczno-lingwistycznych algorytmów, uczenia się maszynowego i sztucznych sieci neuronowych, aby na bieżąco określić tematyczne kategorie zasobów. W wizualizacji grupowanych obiektów bardzo przydatne są tak zwane mapy samoorganizujące się (ang. *Self Organizing Maps - SOM*), rodzaje sztucznych sieci neuronowych o szerokim zastosowaniu. *SOM* zostały rozwinięte od roku 1982 przez T. Kohonena; przez co znane też są pod nazwą sieci Kohonena. Rysunek 11 ilustruje semantyczną mapę grup listy dyskusyjnej comp.ai.neural.nets, wygenerowanej na portalu grupy fińskich naukowców *WebSOM*[41]. Przy modulowaniu reprezentacji semantycznych w zadaniach filtrowania i wyszukiwania informacji wykorzystywany jest wektorowy model przestrzeni wielowymiarowej



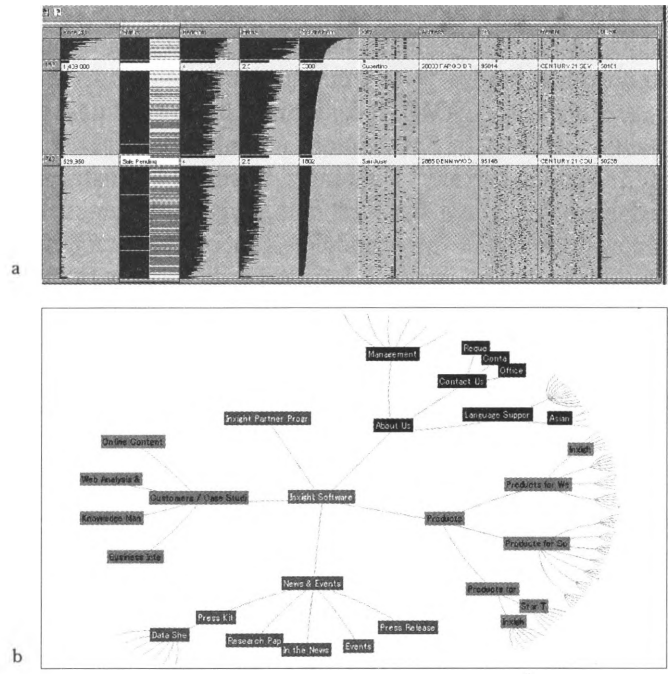
Rysunek 11. Grupowanie mapy *WebSOM* grupy dyskusyjnej comp.ai.neural.nets

Źródło: *WEBSOM map* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://websom.hut.fi/websom/comp.ai.neural-nets-new/html/root.html.

(*Vector Space Modelling - VRM*)[42]. Zagadnienia informacji wielowymiarowej w opar-ciu o przestrzenie wektorowe opisuje **semantyka wektorowa** (ang. *vectorial semantics*). Dokumenty są przedstawiane w sposób formalny przy użyciu **wektorów cech**, za które mogą nam posłużyć np. słowa kluczowe, sekwencje słów, odległość pomiędzy wyrazami, występowanie spójników, topologia obiektów w dokumentach, formaty i rozmiary plików itp. Procedurę tworzenia modelu przestrzeni wektorowej można podzielić na trzy etapy. Pierwszym jest indeksowanie dokumentów i wyłonienie słów oddających treść dokumentu.

[41] *WEBSOM – A novel SOM-based approach to free-text mining* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://websom.hut.fi/websom/.

[42] I. Larsen: *Vector Space Modeling* [on-line] THOR Center for Neuroinformatics [dostęp 19 maja 2009]. Dostepny w World Wide Web: http://eivind.imm.dtu.dk/thor/projects/multimedia/textmining/node5.html.



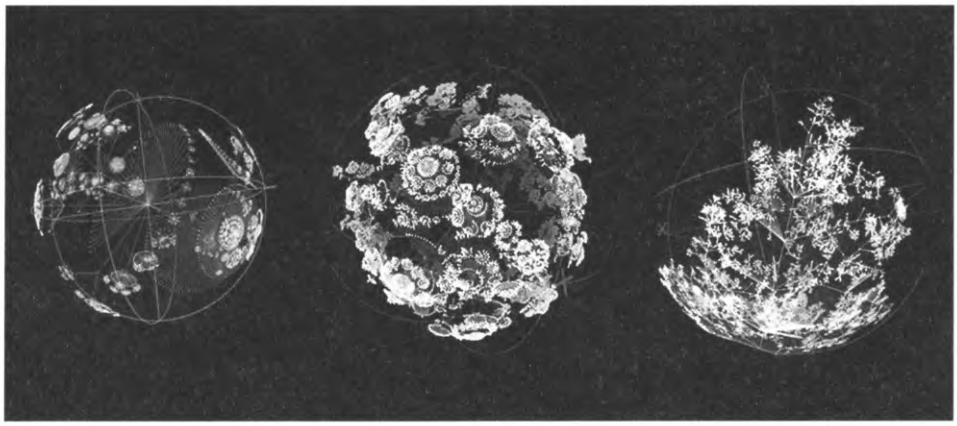
Rysunek 12. Strategie powiększenia przestrzeni informacyjnej firmy *Inxight*:
a) *TableLens* – wybrany fragment danych jest powiększany podczas eksploracji całej tabeli;
b) *StarTree* – projekcja hiperboliczna *fisheye* na wybrany fragment danych.

Źródło: *Inxight* [on-line] [dostęp 31 sierpnia 2008]. Dostępny w World Wide Web: http://www.inxight.com (adres aktualny do grudnia 2008 r.).

Na drugim etapie zachodzi ważenie słów indeksowanych, czyli określenie, w jakim stopniu termin jest ważny dla dokumentu w odniesieniu do zapytania. Na koniec ustalana jest pozycja rankingowa dokumentu na liście odpowiedzi.

W procesach przeglądania i wyszukiwania danych dużą rolę w aplikacji odgrywa przestrzeń eksploracyjna, która jest ograniczona oknem monitora. Zapotrzebowanie na wyjawienie oraz śledzenie szczegółów (niski poziom informacyjny) na obrazie koliduje z informacją kontekstową wysokiego poziomu. Użytkownik zazwyczaj wymaga zapewnienia obu poziomów. Problem ten występuje pod nazwą **focus+context**.

Znane rozwiązania to: użycie wielu widoków w osobnych oknach (*multiply view*), powiększenie/pomniejszenie fragmentów obiektów (ang. *zooming*, np. technika *table lens*) oraz zniekształcenia geometryczne (rybie oko, ang. *fisheye*) zaprezentowane odpowiednio na Rysunkach 12a, b. Geometrycznym sposobem na rozciągnięcie obszaru eksploracji jest reprezentacja hierarchicznych struktur w przestrzeni hiperbolicznej. Pierwszymi aplikacjami, które wykorzystały technikę **fisheye** były przeglądarki hiperboliczne. Przestrzeń Euklidesową zastępuje się hiperboliczną, którą rzutuje się na kolisty obszar widzenia. Ten mechanizm zapewnia więcej miejsca na wizualizację hierarchii (obwód koła rośnie wykładniczo z promieniem, co oznacza, że ze wzrostem odległości mamy eksponencjalne powiększenie przestrzeni).

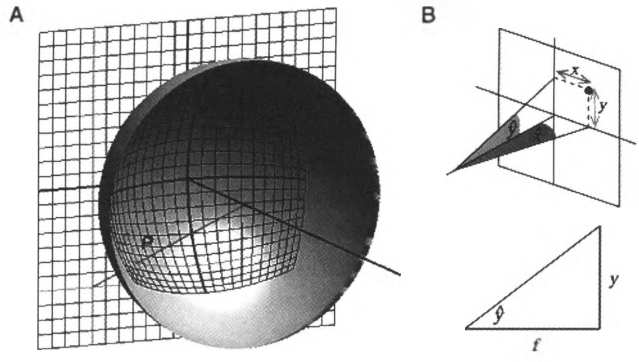


Rysunek 13. Wizualizacje sieci z ilością węzłów ponad 50 tys. w programie *Walrus*

Źródło: *Walrus – Gallery: Visualization & Navigation* [on-line] CAIDA, the Cooperative Association for Internet Data Analysis [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.caida.org/tools/visualization/walrus/gallery1/.

W automatycznej klasteryzacji pozycjonowanie dokumentów zachodzi w kierunku dół-góra: od najniższego poziomu do najwyższego. Próba wizualizacji takiej struktury w przestrzeni trójwymiarowej prowadzi do umieszczenia głównego węzła w centrum, a węzłów podrzędnych we wszystkich kierunkach wokół środka (program *Walrus*[43], dla którego na Rysunku 13 podane są przykłady wizualizacji). Narzędzie Walrus jest otwartym oprogramowaniem do interaktywnej wizualizacji zorientowanych grafów o dużej ilości węzłów w przestrzeni trzywymiarowej. Poprzez zastosowanie zniekształcenia *fisheye* zapewnienia jednoczesne wyświetlanie zarówno szczegółów, tak i całego kontekstu. Charakterystyczną obecnie tendencją w projektowaniu graficznym jest korzystanie z przestrzeni 3D. Znajdujemy sporo argumentów przemawiających za ostateczną dominacją systemów wizualizacji przestrzennej. Naturalnym jest stwierdzenie, że żyjemy w świecie trójwymiarowym i nasz mózg jest przystosowany do interakcji właśnie w trzech wymiarach. Obraz pierwotny odwzorowywany na sferycznej siatkówce (Rysunek 14) również posiada cechy kulistej struktury 3D. Wydaje się więc uzasadnione wykorzystywanie sferycznych metod wizualizacji danych dostosowanych do naturalnych predyspozycji ludzkiego aparatu wizualnego. Jednak jak do tej pory wizualizacja dwuwymiarowa pozostaje ciągle najprostszym sposobem przedstawiania wyników, gdyż jest zazwyczaj umieszczana na dwuwymiarowej kartce papieru lub podobnej płaszczyźnie monitora komputerowego.

[43] *Walrus – Graph Visualization Tool* [on-line] CAIDA, the Cooperative Association for Internet Data Analysis [dostęp 19.05.2009]. Dostępny w World Wide Web: http://www.caida.org/tools/ visualization/walrus/.



Rysunek 14. Reprezentacja obrazu na siatkówce. (**A**) Mapowanie powierzchni na sferze geometrycznej układu siatkówki. (**B**) Sposób konwersji obrazu sferycznego do płaszczyzny projekcyjnej.

Źródło: J. C. A. Read, B. G. Cumming. *Does depth perception require vertical-disparity detectors?* Journal of Vision 2006, Vol. 6 (12), A. 1, s. 1327.

Tabele z dużą ilością kolumn zawierających liczby są zazwyczaj zupełnie nieczytelne dla użytkownika – dopiero ich prezentacja na wykresach w postaci zbioru punktów lub linii staje się istotnym elementem procesów kognitywnych zachodzących w mózgu. I chociaż wykresy takie dobrze wykorzystują własności kory wzrokowej człowieka, która posiada wyspecjalizowane obszary do analizy krawędzi – poziomych, pionowych i pochyłych – to już jednak wizualizacje przedstawiające rozproszone zbiory punktów bez ich wcześniejszej konglomeracji stają się często nic nie znaczącym szumem. Tak więc zwiększenie przestrzeni o jeden wymiar znacząco rozciąga zakres informacji wizualnej. Kolejnym powodem zainteresowania wizualizacją przestrzenną, to zaawansowanie technologii graficznych 3D, które stają się coraz mniej kosztowne. Systemy rzeczywistości wirtualnej cieszą się dużą popularnością wśród miłośników gier komputerowych. Nie bez znaczenia jest również efekt estetyczny, który wywołują umiejętnie zilustrowane sceny 3D.

Aktualne prace nad rozwojem metod wizualizacji oprócz doskonalenia aplikacji 3D przewidują także wprowadzenie wymiaru czasowego i tym samym stworzenie dynamicznej eksploracji, wspomagającej badanie zmian w zasobach informacyjnych, np. w piśmiennictwie, bibliotekach cyfrowych, serwisach sieciowych.

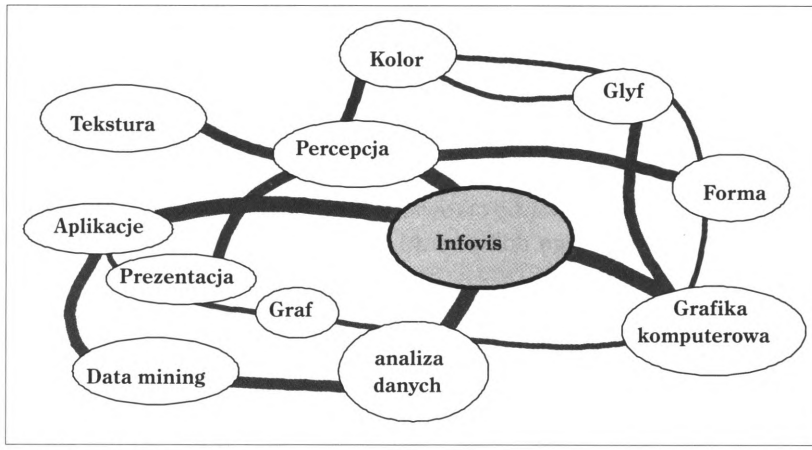
W mnogości opisanych sposobów graficznych reprezentacji, technik i coraz to nowych idei nie jest łatwo o porządną systematykę tych metod. Przedstawione w tym rozdziale funkcjonalne podejście odwołuje się do rodzaju informacji, wykrycia: czy to liniowa, hierarchiczna, sieciowa albo wielowymiarowa informacja? Na obecnym etapie rozwoju metodologii Infovis da się wyodrębnić główne modele wizualizacji:

- grafy, czyli węzły i linki;
- algorytmy wypełnionej przestrzeni takie jak treemap, mapy *SOM*;
- krajobrazy informacji, często używane przez naukowców w latach dziewięćdziesiątych ubiegłego wieku.

A zatem, metody wizualizacji danych ewaluowały od interfejsów programów z minimalną ilością elementów graficznych, wykorzystujących zagnieżdżone drze-

wa klasyfikacji, tabele oraz wykresy dwuwymiarowe, przez diagramy relacji między dokumentami przy użyciu takich abstrakcyjnych kształtów jak koła, kwadraty, linie oraz łącza, do przeglądarek hiperbolicznych i geoprzestrzennych map z włączoną osią czasu.

Próba wyselekcjonowania głównych typów informacji nie oznacza, że nie możemy spotkać w życiu przykładów kombinowanych albo zmieniających typów danych. Ponieważ w świecie zachodzą dynamiczne procesy, a zatem informacja też bezustannie się zmienia, pomiędzy pierwotnie równorzędnymi elementami niosącymi informację też mogą powstawać relacje hierarchiczne. W analizie informacji płynącej z otaczającego nas świata dążymy do sklasyfikowania występujących w rzeczywistości obiektów w grupy – klasy. Czynimy to na podstawie wspólnych ich cech (wygląd, przeznaczenie, pochodzenie itp.) lub zachowań (co obiekt może wykonać?). Obiekty – w trakcie poznawania coraz większej ich – ilości grupujemy w klasy, a następnie klasy nadrzędne. Dążenie do hierarchizacji elementów informacji jest więc naturalnym objawem, który sygnalizuje potrzebę mapowania przeszukiwanych wyników. Mapy, które obecnie są najpopularniejszym rozwiązaniem w sieciowych projektach wizualizacji informacji, znalazły szerokie spektrum zastosowań, przez co dzielimy je na kartograficzne, koncepcyjne (semantyczne) i domenowe[44]. Na koniec w ramach autodoświadczenia w percepcji map dla czytelnika zamieszczony został przykład mapy koncepcyjnej Infovis (Rysunek 15).



Rysunek 15. Mapa koncepcyjna Infovis

Źródło: Na podst. C. Ware. *Information Visualization: Perception for Design*. Wyd. 2. San Francisco: Morgan Kaufmann, 2004, s. 368.

1.3. Wizualizacja informacji w poszukiwaniu strategii mapowania nauk (Mapping Science)

W obliczu integracji wielu dyscyplin naukowych wykrywanie obszarów zainteresowania różnych dyscyplin naukowych staje się praktyką, która może przyczynić się do stymulacji badań interdyscyplinarnych. Aby lepiej zrozumieć strukturę i dy-

[44] *Exhibit Purpose and Goals* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://scimaps.org/.

namikę rozwoju poszczególnych działów nauki oraz znaleźć sposób wykrycia w nich trendów tematycznych, naukowcy posługują się wskaźnikiem „przesunięcia granic naukowych" (ang. *scientific frontiers*) To pole badawcze, luźno zdefiniowane jako **wizualizacja dyscyplin wiedzy** lub **nauki** (ang. *Knowledge Domain Visualisation - KDViz*), jest przedmiotem badań od zaledwie dziewięciu lat (Börner i in. 2003, Chen 2003). Ponieważ stosowane tutaj metody prowadzą do generowania map graficznych, równolegle istnieje inne nazewnictwo: **mapowanie nauk** (ang. *Mapping Science*)[45], lub rzadziej używana: **naukografia** (ang. *Scientography*)[46]. Ostatnie pojęcie zostało wprowadzone w 1960 r. przez Eugene Garfielda, założyciela Instytutu Filadelfijskiego (*Institute of Scientific Information - ISI*, obecnie znanego jako *Thomson Scientific*[47]). Na swoim portalu[48] umieścił on szereg artykułów, poświęconych problemom i przykładom naukografii (Garfield 1998). Mapowanie nauki jako metodologia wyrosła z tradycji bibliometrycznych. Wykorzystuje ona najbardziej znane bibliograficzne bazy danych instytutu *ISI*, czyli indeksy cytowań poszczególnych artykułów publikowanych w czasopismach naukowych (*ISI Citation Index*), jak również sumaryczny indeks liczby i dynamiki cytowań wszystkich artykułów w danym czasopiśmie *Journal Citation Reports - JCR*. W ten sposób mapowane są nauki ścisłe, nauki społeczne oraz nauki humanistyczne[49]. Przegląd historyczny oraz opinie, w jaki sposób nauka była mapowana od dziesięcioleci, przedstawiono szczegółowo w serii artykułów[50].

Do klasycznych technik mapowania zalicza się przestrzenną reprezentację danych bibliometrycznych (autorów, dokumentów, czasopism, kategorii) w oparciu analizę wspólnych cytowań autorów (ang. *author* **co-citation analysis**) lub czasopism (ang. *document co-citation analysis*). Metoda ta zainicjowana została w latach siedemdziesiątych XX w. w pracach Marshakowej[51], Small i Griffitha[52] oraz Garfielda[53]. Dwa dokumenty są wspólnie cytowane, jeśli równocześnie występują w wykazie bibliograficznym trzeciego dokumentu. Podobieństwo pomiędzy dokumentami

[45] Więcej: *Science Frontiers 1976–* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.science-frontiers.com/.

[46] E. Garfield: *Scientography: Mapping the tracks of science.* Current Contents: Social & Behavioural Sciences 1994, nr 7(45), s. 5-10.

[47] *Science - Thompson Reuters* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://scientific.thomson.com/.

[48] *Eugene Garfield, Ph. D. Home Page* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://garfield.library.upenn.edu/.

[49] K. Börner, Ch. Chen, K.W. Boyack: *Visualizing Knowledge Domains.* W: Blaise Cronin (red.). *Annual Review of Information Science & Technology.* Medford, NJ: Information Today, Inc./American Society for Information Science and Technology 2003, Vol. 37, s. 180-185; F. Moya-Anegón i in. *A new technique for building maps of large scientific domains based on the cocitation of classes and categories.* Scientometrics 2004, Vol. 61, nr 1, s. 129-145; E.C.M. Noyons, H.F. Moed. *Combining Mapping and Citation Analysis for Evaluative Bibliometric Purposes: A Bibliometric Study.* Journal of the American Society for Information Science 1999, nr 50(2), s. 115-131; E. Garfield. *Essays/Papers on „Mapping the World of Science"* [on-line]. *Eugene Garfield, Ph. D. Home Page* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://garfield.library. upenn.edu/mapping/mapping.html .

[50] F. Moya-Anegón i in., dz. cyt., s. 130-133; Boyack K. W. i in. *Mapping the backbone of Science.* Scientometrics 2005, Vol. 64, nr 3, s. 353-354; E.C.M. Noyons, H.F. Moed. dz. cyt., s. 115-116.

[51] I. V. Marshakova: *A system of document connection based on references.* Scientific and Technical Information Serial of VINITI 1973, Vol. 6 (2), s. 3-8.

[52] H. Small. *Co-citation in the scientific literature: A new measurement of the relationship between two documents.* Journal of the American Society of Information Science 1973, Vol. 24 (4), s. 265-269; H. Small, B. C. Griffith. *The structure of scientific literatures I: Identifying and graphing specialities.* Science Studies 1974, nr 4, s. 17-40.

[53] E. Garfield: *Essays/Papers on „Mapping the World of Science"...*

jest obliczane na podstawie częstości ich wspólnych cytowań. W innej metodzie bibliografii sprzężonych (ang. *bibliographic coupling*) – dwie prace odwołują się w bibliografii do tych samych dokumentów, których ilość determinuje podobieństwo. Analiza może być oparta również o powiązane cytowania lub przypisy kilku autorów: *inter-citation*. Koncepcja prefiksów *‚co-'* i *‚inter-'* jest dobrze udowodniona w pracy White & McCain[54]. Pierwszy implikuje wspólne występowanie badanych jednostek, drugi – ich wzajemne powiązania. Analiza cytowań pozwala na rozpoznawanie dominujących obszarów badań w danej dziedzinie, jak również struktur społecznych wewnątrz i na zewnątrz środowisk naukowych. O aspekcie społecznym takiej analizy wielokrotnie pisał B. Hjørland – wybitny znawca problematyki organizacji wiedzy oraz traktujący analizę domenową jako jeden z podstawowych paradygmatów informacji naukowej[55]. Ścieżki koo-cytowań mogą także ujawniać kierunki integracji interdyscyplinarnych.

Od roku 2008 w celu zademonstrowania 10-letnich osiągnięć w dziedzinie mapowania informacji uruchomiono wirtualną wystawę *Places@Spaces: Mapping Science*[56]. Portal redagowany jest przez K. Börner, *School of Library and Information Science, Indiana University*, której zespół prowadzi zaawansowane badania nad metodologiami wizualizacji dziedzin nauki, obszarów współpracy naukowców oraz bibliotek cyfrowych. Jak podkreśla Ch. Chen[57] – główny inicjator badań nad *KDViz* – do rzetelnej analizy wyników potrzebne są połączone wysiłki specjalistów różnych dziedzin: od filozofii nauki i naukometrii do zarządzania wiedzą i eksploracją danych. Perspektywę interdyscyplinarną zapewni tylko współpraca wielu naukowców. Nauka jest społecznym fenomenem, na który składają się poglądy i obserwacje badaczy. Przedmiotem badań wizualizacji dyscyplin wiedzy są tak zwane sieci naukowe, składające się ze współpracy naukowej, powiązań cytowań w piśmiennictwie naukowym, frontów i zakresów badań oraz kierunków rozprzestrzeniania się (dyfuzji) wiedzy.

Nauka rozwija się za pomocą specyficznych metod naukowych nazywanych też **paradygmatami nauki**. Uchwycenie dynamiki naukowych paradygmatów można zbadać za pomocą analizy cytowań w literaturze naukowej, co jest szczegółowo opisywane w wiodących pracach na temat *KDViz*[58]. W książce *The Structure of Scientific Revolutions* opublikowanej w 1962 r. T. S. Kuhn dowodzi, że nauka nie jest jednostajnym, kumulatywnym pozyskiwaniem wiedzy[59]. Zamiast tego nauka *jest serią spokojnych okresów przerywanych przez gwałtowne intelektualne rewolucje, po których*

[54] H. D. White, K. W. McCain: *Visualization of literatures*. Annual Review of Information Science and Technology 1997, Vol. 32, s. 99-168.

[55] B. Hjørland, H. Albrechtsen: *Toward a new horizon in information science: domain analysis*. Journal of the American Society for Information Science (JASIS) 1995, nr 46, s. 400-425; B. Hjørland. *Domain analysis in information science: eleven approaches – traditional as innovative*. Journal of documentation 2002, nr 58, s. 422-462.

[56] *Exhibit Purpose and Goals...*

[57] Ch. Chen: *Mapping Scientific Frontiers. The Quest for Knowledge Visualization*. London: Springer-Verlag, 2003, s. 1-5.

[58] Tamże; Ch. Chen. *Information Visualization. Beyond the Horizon*. 2nd ed . London: Springer, 2006, s. 5.

[59] F. Pajares: *The Structure of Scientific Revolutions by Thomas S. Kuhn. Outline and Study Guide* [on-line]. Emory University. Division of Education Studies [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.des.emory.edu/mfp/Kuhn.html.

jeden koncepcyjny światopogląd jest zamieniany przez inny. W przedstawionym ujęciu rozwój nauki przebiega w kilku fazach:

- W okresie przedparadygmatycznym nie ma między uczonymi zgody co do fundamentalnych podstaw danej dziedziny, dlatego ich badania nie są jednoznacznie ukierunkowane.
- W „normalnej" (dojrzałej) fazie - jakiejś szkole udaje się dokonać przełomu, który staje się obowiązującym paradygmatem, do którego przyłączają się inne szkoły.
- Następnie pojawiają się anomalie, których nie można rozwiązać w obrębie paradygmatu i które prowadzą do kryzysów. Następuje wówczas rewolucja naukowa i stary paradygmat zastępowany jest nowym.

Zgodnie z poglądami Kuhna, paradygmat kieruje wysiłkiem badawczym społeczności naukowych i jest tym kryterium, które najbardziej ściśle identyfikuje obszary nauk. Fundamentalnym argumentem Kuhna jest to, że dla dojrzałej nauki typową drogą rozwojową jest kolejne przechodzenie w procesie zmian dynamicznych od jednego do innego paradygmatu. Gdy ma miejsce zmiana paradygmatu, *świat naukowy zmienia się jakościowo i jest jakościowo wzbogacany przez fundamentalnie nowe zarówno fakty jak i teorie*[60].

Badacze *KDViz* od lat z powodzeniem wykorzystują techniki wizualizacji informacji, aby wykryć zmiany krytyczne w rozwoju wiedzy naukowej. Teorie i metody mapowania zakresów badawczych nauk formowały się od dziesięcioleci, wykorzystując osiągnięcia wielu dyscyplin: filozofii i socjologii nauki, naukometrii, wyszukiwania informacji, wizualizacji informacji. Większość tradycyjnych metod polega na wykrywaniu wzorów sieci cytowań w dokumentach naukowych. Intelektualną strukturę społeczności naukowych można analizować na podstawie dynamiki tej sieci. Przy takim założeniu, obserwacja ewolucji dyscypliny naukowej jest możliwa poprzez modulowanie oraz wizualizację struktur cytowań. Wraz z nagromadzeniem piśmiennictwa naukowego badania nad siecią cytowań rozwinęły się od studiów pionierskich z wykorzystaniem technik niekomputerowych przed 1980 r. do zaawansowanych metod, bazujących na algorytmach statystycznych (od roku 2000).

Mapowanie nauk może pomóc w identyfikacji subdyscyplin lub obszarów badawczych i ich wzajemnych powiązaniach w ramach konkretnej dziedziny. Niektórzy utrzymują, iż analiza ewolucji sieci cytowań pozwala nawet na predykcję trendów badawczych oraz oszacowanie rozpiętości oddziaływań społecznych nauki w przyszłości[61]. Ponadto tworzone mapy nauk stymulują poznawanie współczesnego stanu wiedzy i mogą pomóc w dokonaniu odkryć naukowych.

Ch. Chen opisuje w jaki sposób zaobserwował dwa paradygmaty, przez które przeszła teoria superstrun[62]. Teoria strun jest jednym z najbardziej aktywnych tematów badawczych w fizyce teoretycznej. Tradycyjne pojęcie cząstki elementarnej zastąpiono struną, czyli drgającą wielowymiarową strukturą topologiczną czasami utożsamianą z odcinkiem linii lub kołem. Cząstki są strunami mającymi rozmiary

[60] *Struktura rewolucji naukowych*. W: *Wapedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://wapedia.mobi/pl/Paradygmat.

[61] Ch. Chen: *Information Visualization...*, s. 5; I. Samoylenko, T.-C. Chao, W.-C. Liu, C. M. Chen. *Visualizing the scientific world and its evolution*. Journal of the American Society for Information Science and Technology 2006, Vol. 57 (11), s. 1461-1469.

[62] Ch. Chen: *Information Visualization...* , s. 157-171.

zbliżone do długości Plancka (około 10^{-35}m), które wibrują z pewnymi ściśle określonymi częstotliwościami. Struny te cechuje supersymetria. Do analizy wybrano dane bibliograficzne, zawierające w temacie pracy wyrazy „string (superstring) theory". Dalej zbadano częstotliwości występowania terminów, charakteryzujących daną teorię w różnych okresach jej rozwoju, np. „black holes", „branes" oraz „supegravity". Wyniki wykazały dwie krytyczne zmiany na skali czasu. Nawiązywały one do dwóch rewolucji paradygmatycznych w teorii superstrun: pierwsza w latach 1984-1985, kiedy rozwiązywano anomalie w supersymetrii wielowymiarowej oraz druga – zaczynając od 1994, kiedy podstawowe problemy zaczęły dotyczyć dualizmu i teorii strun. Warto zaznaczyć, iż doświadczalna weryfikacja tych podejść będzie prawdopodobnie możliwa już w 2009 r. dzięki nowym pomiarom satelity Planck wprowadzonej na orbitę w 2008 r.

Jak wspomniano wyżej podstawowe metody analizy bibliometrycznej skupiają się na wzajemnych i/lub wspólnych cytowaniach prac naukowych. W tym schemacie podobieństwo pomiędzy artykułami *i* oraz *j* wyznaczane jest za pomocą liczby dokumentów, powołujących się na obydwa źródła. Obiekty badane są sprowadzane do postaci macierzy podobieństwa (ang. *proximity/similarity matrix*)[63], które następnie są konwertowane do map przestrzennych. Aby zmniejszyć wymiary danych stosuje się dwie popularne techniki statystyczne: analizę czynnikową oraz skalowanie wielowymiarowe, wykorzystane w danej pracy. Oprócz statystycznych coraz częściej wykorzystuje się algorytmy oparte o sztuczne sieci neuronowe. Metody te prowadzą do poprawienia widoczności ukrytych struktur, zawartych w macierzy danych.

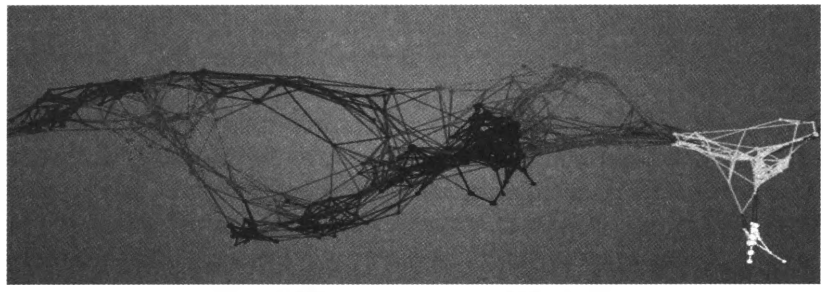
Dane wzajemnie powiązanych cytowań z czasopism rozmaitych dziedzin zgromadzono przy wykorzystaniu bazy danych *Science Citation Index* w okresie 1994-2001[64]. Autorzy na mapie zinterpretowali współczesną ewolucję świata naukowego, charakteryzującego się liniową strukturą i trzema głównymi kierunkami rozwoju: nauki matematyczno-fizyczne, nauki o życiu i nauki medyczne. Analiza cytowań pomiędzy czasopismami potwierdza semantyczne odseparowanie nauk matematyczno-fizycznych od innych domen, a badania w dwóch pozostałych domenach są powiązane, co wskazuje na wzajemne lansowanie odkryć w obu naukach. Większość map nauki generowana była raczej dla ograniczonych pól badawczych z niewielkich zbiorów danych, liczących setki lub tysiące węzłów w grafie. Takie prekursorskie wizualizacje ujawniały nie tylko relacje pomiędzy dokumentami i artykułami, lecz wspomagały detekcję najważniejszych badaczy w ramach konkretnej dziedziny, a także analizę struktury dyscypliny nauki i jej ewolucję. W późniejszym okresie gwałtownego wzrostu zasobów piśmiennictwa naukowego bazy danych liczyły kilka milionów rekordów – artykułów. Aby sprostać wymogom skalowalności, metodologia rozszerzyła się do algorytmów klastrowania, sieci neuronowych i map samoorganizujących się oraz technik kombinowanych.

[63] W bibliometrii macierze prawdopodobieństwa przedstawiają relacje między publikacjami na podstawie liczby ich cytowań. Model macierzowy opiera się na liczbowej reprezentacjicech dokumentów, co pozwala zmierzyć podobieństwo między nimi.

[64] Samoylenko I. i in., dz. cyt.

W ostatniej dekadzie mapowanie nauki wykonuje się na dużą skalę. Jednym z interesujących przykładów, gdzie wykorzystano indeks naukowych cytowań i repre-zentującą kompletną strukturę aktualnej nauki, włączając nauki przyrodnicze i społeczne jest praca *Mapping the backbone of Science*[65]. Kolekcja danych liczyła ponad milion dokumentów i około siedmiu tysięcy czasopism. Do wygenerowania grafu użyto ośmiu miar podobieństwa pomiędzy czasopismami, różniące się dokładnością, skalowalnością algorytmu oraz czytelnością układu czyli stopniem klasteryzacji. Dzięki temu, że badane czasopisma są przypisane do konkretnych dyscyplin, można zrozumieć organizację aktualnej nauki na poziomie agregacyjnym.

Najnowsze mapy wynikające z badań *KDViz* regularnie są publikowane na portalu *Places@Spaces* w sekcji *Domain Maps*[66]. R. Klavans i K. Boyack użyli metryki opartej o bibliografie łączone oraz wektory słów kluczowych ponad siedmiu milionów publikacji oraz szesnastu tysięcy czasopism i materiałów konferencyjnych za okres 2001-2005[67] (p. Rysunek 16). Za przestrzeń docelową wybrano powierzchnię sfery, po czym na potrzeby prezentacji dane odwzorowano metodą walcową równokątną[68]. Zbieżność z metodologią niniejszej rozprawy doktorskiej jest przypadkowa, natomiast jest dowodem słuszności przedstawionej strategii. Na podstawie zmian powiązań pomiędzy poszczególnymi obszarami danych R. Klavans i K. Boyack podjęli się prognozowania struktury nauki w najbliższej przyszłości.



	matematyka, fizyka		biotechnologia		nauki o mózgu
	chemia		nauki o Ziemi		nauki o zdrowiu
	nauki komputerowe		biologia		nauki społeczne
	inżynieria		nauki medyczne		nauki humanistyczne

Rysunek 16. Dwuwymiarowa mapa nauki zrobiona na podstawie mapowania bibliografii łączonych i słów kluczowych ponad 7 milionów publikacji za okres 2001-2005

Źródło: R. Klavans, K. Boyack. *Maps of Science: Forecasting Large Trends in Science* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.scimaps.org/dev/big_thumb.php?map_id=164.

[65] K. W. Boyack, i in. dz. cyt..

[66] *Browse map. Overview* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.scimaps.org/browse/.

[67] R. Klavans, K. Boyack: *Maps of Science: Forecasting Large Trends in Science* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.scimaps.org/dev/big_thumb.php?map_id=164.

[68] *Odwzorowanie walcowe równokątne* – jeden z zrzutów kartograficznych, zwany też *odwzorowaniem Mercatora*. Południkom i równoleżnikom odpowiadają odcinki, kąty między nimi są zachowane, przy czym najmniejsze zniekształcenia powstają blisko równika, największe – na biegunach... Por. *Odwzorowanie walcowe równokątne*. W: *Wikipedia. Wolna Encyklopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Odwzorowanie_ walcowe_r%C3% B3wnok%C4%85tne.

Widzimy, że wizualizacja domen wiedzy sprowadza się głownie do mapowania literatury naukowej w wybranej przestrzeni semantycznej bądź tematycznej. W procesie tworzenia reprezentacji domen wiedzy najczęściej wyróżniamy następujące etapy:

- wybór odpowiedniego do postawionego problemu źródła danych;
- wybór jednostek analizy oraz ekstrakcja niezbędnych danych;
- wybór właściwej miary podobieństwa (w literaturze są opisywane osiem podstawowych miar[69, 70];
- tworzenie graficznego układu danych przy użyciu wybranych algorytmów;
- eksploracja wygenerowanej mapy w odpowiedzi na pierwotne zapytania.

Informację semantyczno-strukturalną wydobywa się także przy pomocy wspólnie występujących sekwencji słów w dwóch lub więcej artykułach (**co-words**), słów kluczowych (**co-keywords**) lub deskryptorów (**co-descriptors**). Na alternatywne jednostki analizy wybiera się między innymi: kategorie czasopism[71], autorów, słowa kluczowe, źródła itp. i w zależności od potrzeb użytkowników generowane są mapy o różnym stopniu zawartości tematycznych treści naukowych. W mapowaniu dyscyplin naukowych zaadoptowane zostały współczesne technologie grafiki komputerowej. Na wspomnianym portalu *Places@Spaces* swoje artystyczne wizje naukowego świata zamieszczają również graficy. Rysunek 17 ilustruje hipotetyczny model ewolucji nauki, wykonany przez D. Zellera[72], uniwersum nauki zostało przestawione w kształcie meteoru, a obszary badawcze – jako tunele ewoluujące od środka.

Zainteresowanie tą tematyką może przebiegać równolegle z chęcią zrozumienia jak funkcjonują firmy naukowo-technologiczne i jak nimi zarządzać. Inicjatywa mapowania współczesnej wiedzy podejmowana jest również w sektorze komercyjnym przy wykorzystaniu danych bibliometryczno-biznesowych. Celem takich projektów jest też upublicznianie wiedzy naukowej i technologii informacyjno-komunikacyjnych. Aktualne badania KDViz uwzględniają zapotrzebowanie na uproszczone mapy domen o wystarczającej zawartości informacyjnej dla wszystkich niespecjalistów danej dziedziny. Obok projektów wizualizacji dyscyplin naukowych, w niekonwencjonalny sposób przedstawiane są również aktualne trendy w ruchu i aktywności użytkowników w sieci, na przykład: Wikipedia, błogi, technologie WWW. To tylko udowadnia obserwowany rozwój wszelkich form interakcji miedzy użytkownikami w sieci Web 2.0. Pierwsza semantyczna mapa angielskiej Wikipedii (Rysunek 18) wygenerowana została przy użyciu metryki, rejestrującej wspólne kategorie artykułów[73]. Widoczna tu jest logiczna kompozycja klastrów, a mapa udowodniła sprawne zarządzanie strukturą kategorii tematycznych, pomimo dużego zróżnicowania autorów: zarówno botów tak i ludzi.

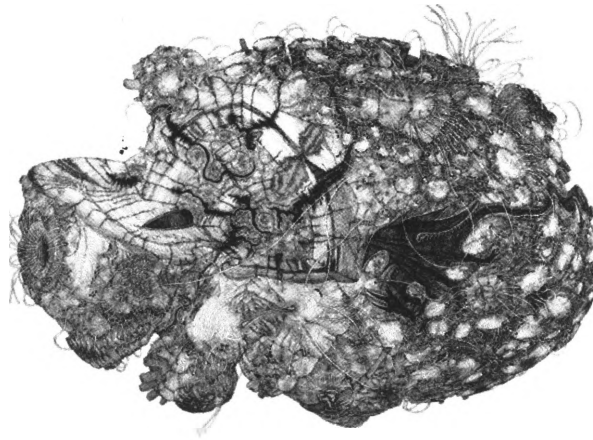
[69] K. W. Boyack, dz. cyt.

[70] R. Klavans, K.W. Boyack: *Identyfying a better measure of relatedness for mapping science.* Journal of the American Society for the Information Science and Technology 2005. Vol. 57, nr 2, s. 251-263.

[71] F. Moya-Anegón i in., dz. cyt..

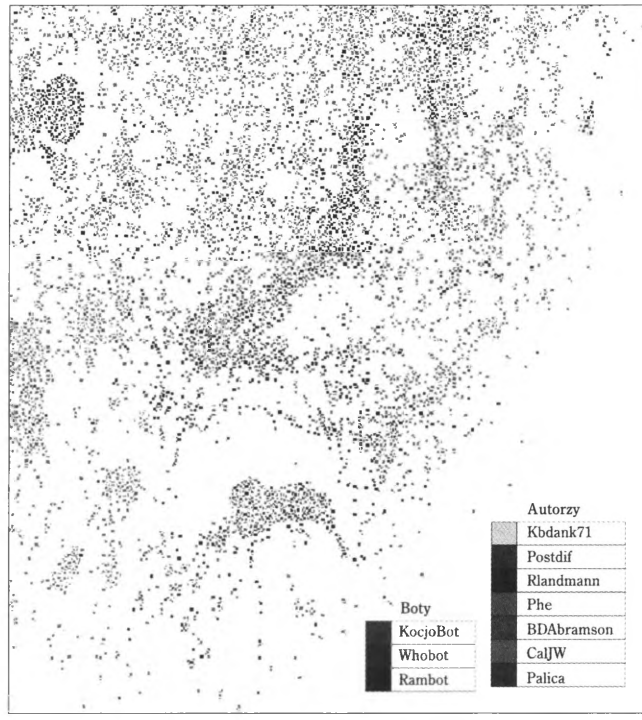
[72] D. Zeller: *Hypothetical Model of the Evolution of Science* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja.2009]. Dostępny w World Wide Web: http://www.scimaps.org/dev/map_detail.php? map_id=163.

[73] T. Holloway, M. Bozicevic, K. Börner: *Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors.* Wyd. specjalne: Understanding Complex Systems. Complexity 2007. Vol. 12, nr 3, s. 30-40.



Rysunek 17. Hipotetyczny model ewolucji nauki oczyma grafika. Uniwersum nauki zostało przestawione w kształcie meteoru. Każdego roku pojawia się nowa warstwa; niebieski kolor wskazuje na nowe dyscypliny, brązowe – na ustabilizowane; każdy „trąbkopodobny" obszar badawczy ewoluuje od środka.

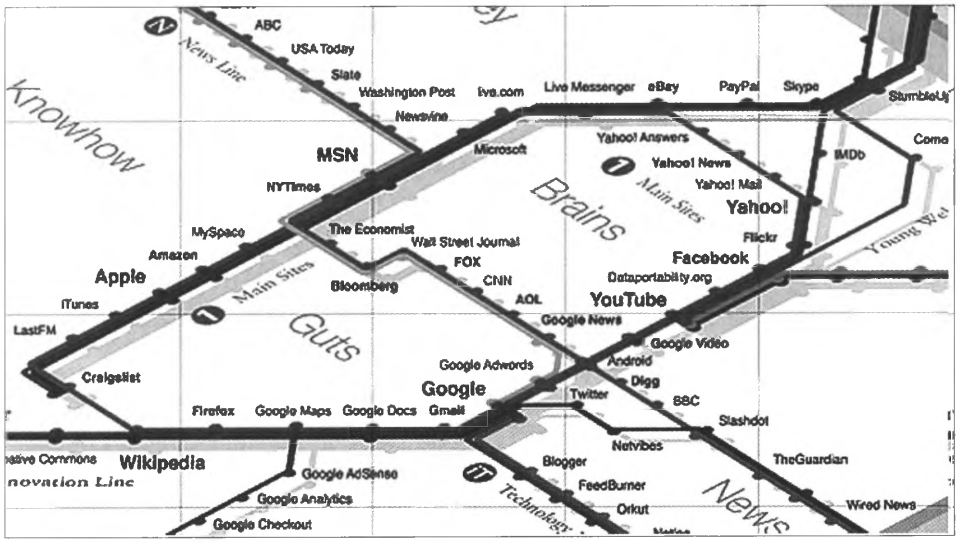
Źródło: D. Zeller. *Hypothetical Model of the Evolution of Science* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.scimaps.org/ dev/map_detail.php?map_id=163.



Rysunek 18. Wizualizacja rodzaju autorstwa artykułów angielskiej Wikipedii

Źródło: : T. Holloway, M. Bozicevic, K. Börner: *Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors*. Wyd. specjalne: *Understanding Complex Systems*. Complexity 2007, Vol. 12, nr 3, s. 38.

Japońska aplikacja pod nazwą *Trendmap*[74] przedstawia kilkaset najbardziej znaczących serwisów i portali internetowych poprzez zmapowanie ich na strukturę metra w Tokio – Rysunek 19. Dodatkowo obiekty (strony) są uporządkowane według kategorii, bliskości semantycznej oraz popularności.



Rysunek 19. Prezentacja modnych serwisów sieciowych poprzez zmapowanie ich na strukturę topologiczną metra w Tokio w aplikacji *Trendmap*

Źródło: O. Reichenstein. *Trend Map 2008. What's new?* [on-line]. Information Architects Japan [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://informationarchitects.jp/ trendmap3-countdown-sneak-peak/.

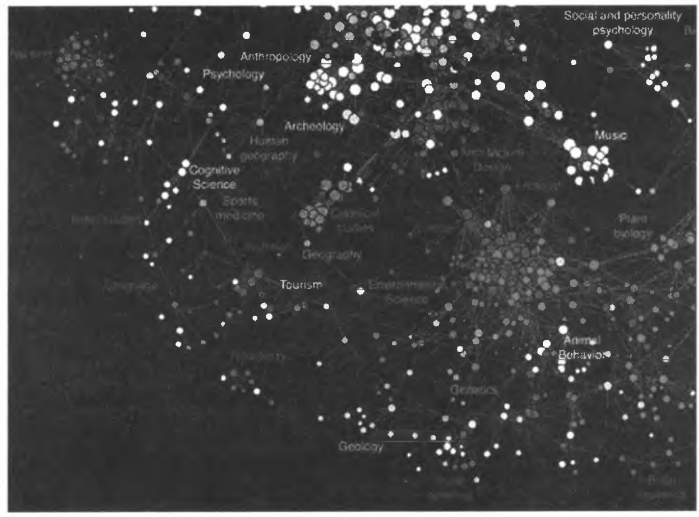
W Internecie nietrudno jest trafić na serwisy sieciowe[75], kierujące do przykładów specjalistycznych aplikacji wizualizacyjnych. Autorzy prześcigają się w ekscentrycznym nazewnictwie: „atlas nauki", cyberatlas, cybergeografia, atlas cyberprzestrzeni[76], naukogramy (*scientogram*[77], *scientograph*), mapy naukowych paradygmatów itp.

[74] O. Reichenstein: *Trend Map 2008. What's new?* [on-line]. Information Architects Japan [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://informationarchitects.jp/trendmap3-countdown-sneak-peak/.

[75] *1100+ examples of information visualization* [on-line]. Infovis.info 2008 [dostęp 19. maja 2009]. Dostępny w World Wide Web: http://www.infovis.info/index.php?cmd=search&words=science&mode= normal. http://www.cs.umd.edu/class/spring2005/cmsc838s/viz4all/viz4all_a.html; *Viz4All – visualization survey* [on-line]. University of Maryland. College Park [dostep 19 maja 2009]. Dostępny w World Wide Web: http://www.cs.umd.edu/class/spring2005/cmsc838s/viz4all/viz4all _a.html.

[76] M. Dodge, R. Kitchin: *An Atlas of Cyberspace* [on-line]. Manchester, 2007 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/ atlas/atlas.html.

[77] B. Vargas-Quesada: *Domain analysis by means of the visualization of maps of vast scientific domains* W: *Proceedings of I International Conference on Multidisciplinary Information Sciences and Technologies, Mérida (Spain), 25th-28th October, 2006* [on-line]. *E-LIS. E-prints in Library and Information Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://eprints.rclis.org/8170/1/Domain_analysis_by_means_of_the_visualization_of_maps_of_vast_scientific_domains.pdf.



Rysunek 20. Fragment mapy nauki, wygenerowanej na podstawie danych o logach użytkowników w serwisach specjalistycznych (kolorowe kółka reprezentują czasopisma, poklasyfikowane według klasyfikacji Art & Architecture Thesaurus; podpisy – ich grupy tematyczne).

Źródło: J. Bollen i in. *Clickstream Data Yields High-Resolution Maps of Science*. PLoS ONE [on-line] 2009, no. 4(3) [dostęp 19 maja 2009] Dostępny w World Wide Web: http://www.plosone.org/ article/info:doi/10.1371/journal.pone.0004803.

Rozwiązania dynamicznych naukogramów pozwalają na obserwowanie ewolucji badań w ramach konkretnej dyscypliny. W wyniku eksploracji możliwa jest analiza trendów w rozwoju przyszłych kierunków badań w dowolnym regionie świata oraz co jest istotne – ich wzajemne porównanie. Wzory powstające na globalnych mapach dyscyplin i specjalności badawczych potwierdzają wzrastającą interdyscyplinarność nauki. Przypomnieć tu należy, iż według Kuhna, twórcy koncepcji paradygmatycznej nauki, typowi naukowcy nie są „obiektywnymi i niezależnymi myślicielami, lecz konserwatystami", gdyż stosują wiedzę zgodnie z dyktatem wyuczonej przez nich teorii. W związku z tym można takie globalne mapy wykorzystać do weryfikacji stopnia twórczości i innowacyjności w podejmowaniu decyzji danej grupy autorów. W poszukiwaniu nowych rozwiązań problemów badawczych analiza piśmiennictwa naukowego nie powinna ograniczać się do tradycyjnie określonego zakresu danej dziedziny, ani do przetestowanych i uznanych w naukometrii i webometrii metodologii. Warto tu, jako przykład innowacyjnego podejścia, zacytować pracę z roku 2009[78], gdzie do mapowania informacji pochodzących z sieciowych portali wykorzystano statystykę logów użytkowników – Rysunek 20. Studiowanie źródeł z tematycznie odległych nauk może przyczynić się zarówno do odnalezienia zaskakujących odpowiedzi, jak i do poszerzenia własnych horyzontów.

Widzimy zatem, iż mapowanie nauki nie jest nową procedurą bibliometryczną: powstała ona w latach siedemdziesiątych XX w. i intensywnie się rozwijała. Przyczynili się do tego głównie pomysłodawcy sprawdzonych i dzisiaj uważanych za podsta-

[78] J. Bollen i in: *Clickstream Data Yields High-Resolution Maps of Science*. PLoS ONE [on-line] 2009, Vol. 4, no. 3 [dostęp 19 maja 2009] Dostępny w World Wide Web: http://www.plosone.org/ article/info:doi/10.1371/journal.pone.0004803.

wowe, metodologii w tej dziedzinie, jak np.: E. Garfield, H. Small, B. Griffith, I. Marshakova, B. Hjørland oraz H. White. Lecz do dalszych postępów w mapowaniu nauki potrzebne było podejście zorientowane na użytkownika. Postępy w technologiach informatycznych i badaniach kognitywnych wyznaczyły nowoczesne reguły projektowania interfejsów programów komputerowych. Wizualizacja (a także estetyka prezentacji) informacji stała się obecnie najistotniejszym ogniwem, decydującym o właściwym postrzeganiu i rozumieniu danych dla wszystkich poziomów końcowego użytkownika. Wizualizacja domen wiedzy jako nowe/odnowione podejście w mapowaniu nauki powstało na progu 3-go tysiąclecia. Znajdujący się w czołówce badaczy *KDViz*: Ch. Chen i K. Börner, podkreślają wieloperspektywiczność zastosowań metod wizualizacji: w wyszukiwaniu informacji, monitorowaniu trendów w nauce i technologiach, polityce finansowania badań i konkretnych naukowców.

Przyszłość nauki rysuje się na podstawie ścisłej współpracy ekspertów dziedzinowych w kwestiach teoretycznego i praktycznego wykorzystania narzędzi wizualizacyjnych w badaniach interdyscyplinarnych. Naukowcy mogą „spojrzeć z lotu ptaka na panoramę nauki", aby jej użyć do eksploracji obszarów zainteresowań, do komunikowania wyników i do anonsowania odkryć naukowych. Początkujący badacze i studenci znajdą tu punkt startowy do weryfikacji własnego mentalnego obrazu świata naukowego. Podjęta tematyka nabiera szczególnej wagi w momencie dyskusji, na ile proponowana metoda może przydać się w wizualizacji nauk komputerowych i im pokrewnych w rozdziale podsumowującym.

Rozdział 2

PRZEDMIOT BADAŃ – KLASYFIKACJA NAUK KOMPUTEROWYCH CCS

2.1. Dyscyplina: Nauki Komputerowe i jej pokrewne

Na początku tego rozdziału przedstawiony zostanie zakres tematyczny pojęcia *Computer Science (CS)*, a mianowicie: informatyka, nauki komputerowe, nauki obliczeniowe, technologia informacyjna. W języku polskim *Computer Science* najczęściej utożsamiany jest informatyką, jednakże nie jest jego dokładnym odpowiednikiem.

Termin *Computing* (powinno się odseparować go od innego terminu *computation*, co dokładnie znaczy obliczenie) figuruje w nazwie badanej klasyfikacji, należy zbadać jego odmianę w języku polskim a także interpretację w środowisku naukowym. Ten imiesłów, wywodzący się bezpośrednio od *computer* i pierwotnie oznaczający działania wykonywane za pomocą maszyn liczących, po raz pierwszy został użyty przez J. von Neumanna[1] w 1945 r. w dokumentacji maszyny *EDVAC*[2] – dla określenia zautomatyzowanych systemów komputerowych (*Automatic Computing System*). Rozważając, czy informatyka powinna być uznana za dyscyplinę naukową, P. J. Dening opisuje, jak zmienia się naukowa identyfikacja informatyki w okresie

[1] John von Neumann – zasłużony matematyk, inżynier chemik, fizyk i informatyk. W historii informatyki znany jest przede wszystkim jako autor architektury komputerów, nazwanej jego imieniem, składającej się w zasadzie z jednostek: obliczeniowej – procesora oraz przechowującej dane – pamięci. Por. *John von Neumann*. W: *Wikipedia. Wolna Encyklopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/John_von_Neumann.

[2] EDVAC (*Electronic Discrete Variable Automatic Computer*), kolejna wersja po ENIAC komputera elektronicznego, zbudowanego w kwietniu 1952 r.

ostatniego półwiecza[3]. Do początku lat dziewięćdziesiątych ubiegłego wieku, a wtedy właśnie powstały pierwsze edycje klasyfikacji *CS*, termin *computing* był używany dla określenia wspólnego obszaru domen matematyki, nauk komputerowych oraz inżynierii. Współcześnie, jak podkreśla P. J. Denning, nauki obliczeniowe (ang. *Computing Science*), matematyka, inżynieria, grafika oraz ich wzajemne kombinacje łączone są w ramach bardziej nowoczesnego pojęcia: nauk komputerowych (ang. *Computer Science*). Europejskim synonimem nauk komputerowych (dawniej *computing*) jest wyraz Informatyka (*Informatics*). W języku polskim termin ten zaproponował w październiku 1968 r. Romuald Marczyński na ogólnopolskiej konferencji poświęconej „maszynom matematycznym" na wzór wyrazów: francuskiego *informatique* i niemieckiego *Informatik*[4]. Natomiast ogólna definicja informatyki pozostaje na razie aktualna: zajmuje się ona procesami projektowania, konstrukcji, oceny, wykorzystania i konserwacji systemów zapamiętywania i przesyłania danych. A procesy te dotyczą sprzętu, oprogramowania, aspektów ludzkich i organizacyjnych. W tym miejscu zacytuję najbardziej aktualną, kompletną i przejrzystą definicję z *Wikipedii*[5]: *Informatyka – dziedzina nauki i techniki zajmująca się przetwarzaniem informacji – w tym technologiami przetwarzania informacji oraz technologiami wytwarzania systemów przetwarzających informacje.* Tu warto jeszcze wspomnieć o pojęciu, bardzo zbliżonym do informatyki – *cybernetyka*. Termin ten w krajach byłego bloku wschodniego traktowany był jako równoważny do *informatyka*. Cybernetyka zajmuje się systemami sterowania oraz związanym z tym przetwarzaniem i przekazywaniem informacji. Na stronie Amerykańskiego Towarzystwa Cybernetycznego (*American Society for Cybernetics*[6]) znajdziemy opis historyczny, definicje tej nauki, bibliografię przedmiotu oraz przydatne odnośniki.

Dwoistość natury i zasięgu informatyki jest przedmiotem wieloletnich sporów naukowców, gdzie należy ją umieścić według skali „nauki ścisłe – inżynieria". Obok pytania: czym jest informatyka, bardziej – nauką, inżynierią czy technologią, Denning postuluje, żeby informatykę potraktować też jako rodzaj sztuki. Teza ta jest tym bardziej zasadna, iż programowanie, projektowanie, inżynieria sprzętu i oprogramowania powstawały i rozwijały się na bazie ludzkich idei, kreatywności oraz wyczucia estetycznego twórców. Sztuka komputerowa obecnie objawia się w zastosowaniach grafiki komputerowej, projektowania, ilustrowania, fotografii, animacji, muzyki, gier i rozrywki.

Następny cytat może całą polemikę skierować na zupełnie odmienny nurt: *Computer science is no more about computers than astronomy is about telescopes* (E. Dijkstra)[7]. Powyższe trudności określenia informatyki jako nauki powodują, iż jest ona w różny sposób interpretowana na zewnątrz i wewnątrz społeczności informatyków. Ten problem, jak również układanie programu nauczania informaty-

[3] P.J. Dening: *Is Computer Science Science?* Communications of the ACM 2005, Vol. 48, nr 4, s. 27-31.

[4] *Informatyka*. W: *Wikipedia* [on-line] [dostęp 19.05.2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Informatyka.

[5] Tamże.

[6] *American Society for Cybernetics*. Portal Amerykańskiego Towarzystwa Cybernetycznego [on-line]. [dostęp 19.05.2009]. Dostępny w World Wide Web: http://www.asc-cybernetics.org/.

[7] E. Dijkstra (1930-2002) – duński naukowiec, pionier informatyki teoretycznej. W roku w 1972 otrzymał Nagrodę Turinga za wkład w języki programowania. Archiwum jego tekstów można znaleźć na stronie: http://www.cs.utexas.edu/users/EWD/.

ki, zdefiniowania zawodu informatyka są aktualnie dyskutowane na licznych forach informatycznych[8].

Informatyka w strukturach klasyfikacyjnych często pozycjonowana jest w naukach matematycznych i częściowo również w naukach technicznych (patrz następny podrozdział). Wynika to stąd, iż historycznie przyjęło się dzielić informatykę na teorię (wywodzącą się z matematyki) i praktykę (powiązaną z naukami technicznymi). Te dwa aspekty informatyki: obliczenia i maszyny liczące, matematyka i elektronika, software i hardware – funkcjonują nierozłącznie i nie sposób ustalić, który jest decydujący. Jak zaznacza profesor W. Duch w swojej książce[9], podział ten jest widoczny na uczelniach w sferach edukacji oraz badań naukowych. Informatyka teoretyczna wykładana na uniwersytetach związana jest z matematyką, natomiast informatyka techniczna uprawiana na politechnikach przechodzi w mikroelektronikę. Jednocześnie W. Duch zaznacza, iż nauki komputerowe korzystają z matematyki stosowanej.

Od lat osiemdziesiątych XX w. zaczęto intensywnie poszukiwać formalnych opisu i przedstawienia rozwijającej się dyscypliny nauki komputerowe. Aby określić dyscyplinę i jej pochodne, autorzy pracy[10] użyli dwuwymiarowej macierzy definiującej nauki komputerowe. Jeden z wymiarów reprezentował obszary procesów naukowych takich jak teoria, abstrakcja i projektowanie, wyłaniające się z pokrewieństwa informatyki z matematyką, naukami przyrodniczymi i inżynierią. Wymiar drugi tworzył zbiór dziewięciu działów przedmiotowych, ustalonych według takich kryteriów jak jednostka przedmiotowa, solidne podstawy teoretyczne, znaczący poziom abstrakcji, projektowanie i implementacja:

1. Algorytmy i struktury danych,
2. Języki programowania,
3. Architektura,
4. Obliczenie numeryczne i symboliczne,
5. Systemy operacyjne,
6. Metodologia i inżynieria oprogramowania,
7. Bazy danych i systemy wyszukiwania informacji,
8. Sztuczna inteligencja i robotyka,
9. Komunikacja człowiek – komputer.

Gwałtowne wzrost liczby dyscyplin obliczeniowych i ich łączny wpływ na społeczeństwo postawiły wymóg określenia podziału ich tożsamości. W międzynarodowym raporcie *Computing Curricula 2004*[11], przygotowanym przez *AIS* (*Association for Information Systems*), *AITP* (*Association for Information Technology Professionals*) i *IEEE-CS* (*Computer Society of the Institute for Electrical and Electronic Engi-*

[8] M. Adamski. *Informatyka – nauka, sztuka, czy rzemiosło?* [on-line]. Uniwersytet Zielonogórski – Miesięcznik Społeczności Akademickiej, 2002 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.uz.zgora.pl/wydawnictwo/miesiecznik11-2002/17.pdf.

[9] W. Duch. *Fascynujący Świat Komputerów*. Poznań: Wydawnictwo NAKOM, 1997, R.1-2.

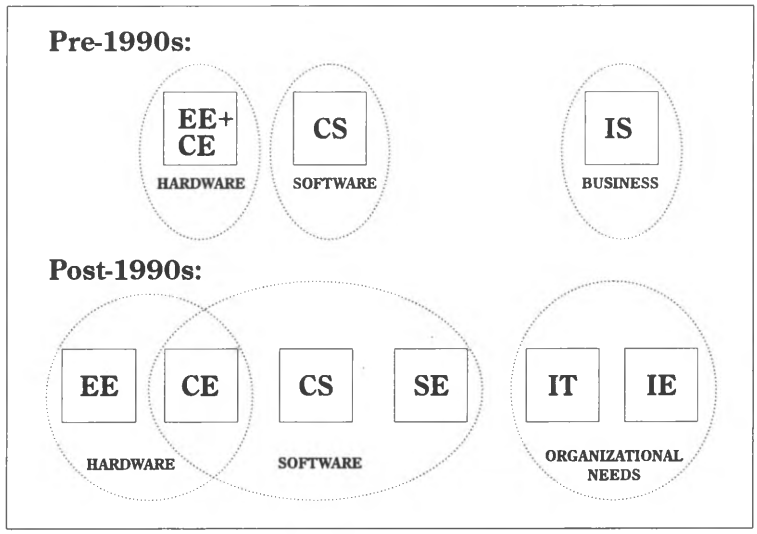
[10] P. J. Denning i in. *Computing as a Discipline*. Communication of the ACM [on-line] 1990, Vol. 32, no. 1 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://cs.gmu.edu/cne/pjd/GP/ CompDisc.pdf.

[11] *Computing Curricula 2004. Overview Report including A Guide to Undergraduate Degree Programs in Computing. A cooperative project of ACM, AIS, IEEE-CS* [on-line]. Shanghai Jiao Tong University. School Of Software [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://se.sjtu.edu.cn/sites/ se/gb/CCSE/CCCS-040601-Overview-Strawman-Rev4.pdf.

neers) nauki obliczeniowe figurują jako rodzina dyscyplin składająca się z inżynierii komputerowej, nauk komputerowych, systemów informacyjnych, technologii informacyjnej (brakującego elementu w poprzednich raportach) i inżynierii oprogramowania. Raport ten zawiera zalecenia, przeznaczone dla programów nauczania na poziomie inżynierskim i stanowi podstawę wiedzy dla kierunku *Computer Science*.

Przed okresem lat dziewięćdziesiątych ubiegłego wieku inżynieria komputerowa figurowała jako specjalność inżynierii elektrycznej – wystarczy zajrzeć do indeksów uczelni politechnicznych tamtych lat. W miarę rozwoju technologii obliczeniowych i komunikacyjnych inżynieria komputerowa, bazująca na technologii mikroprocesorów stała się samodzielna dyscypliną (p. Rysunek 21). Rozpowszechnienie w latach dziewięćdziesiątych XX w. urządzeń elektrycznych i mechanicznych, wykorzystujących elementy elektroniki, które działają na zasadzie logiki cyfrowej, utrwaliły status inżynierii komputerowej jako znaczącej dyscypliny technicznej. W tradycyjnym rozumowaniu inżynieria komputerowa jest to dyscyplina o projektowaniu i konstrukcji komputerów oraz systemów na nich opartych. Aktualnie, dominującym punktem koncentracji inżynierii komputerowej są systemy „wbudowane" (*embedded*), czyli rozwój urządzeń z wbudowanym na stałe oprogramowaniem, np., telefony komórkowe, cyfrowe odtwarzacze audio, kamery cyfrowe, systemy alarmowe, aparaty rentgenowskie, laserowe narzędzia chirurgiczne, tomografy itp.

Zakres nauk komputerowych wyznacza kierunek rozwoju robotyki, cyfrowej analizy obrazów (*Computer Vision*), systemów inteligentnych, bioinformatyki i innych pasjonujących obszarów. Czym się zajmują specjaliści w tej dziedzinie? Do głównych ich zadań należą projektowanie i implementacja oprogramowania, wynalezie-



Rysunek 21. Schemat organizacji działów nauk komputerowych przed- i w latach dziewięćdziesiątych XX w.
Legenda: EE – Electrical Engineering; CE – Computer Engineering; CS – Computer Science;
IS – Information Systems; IT – Information Technology, SE – Software Engineering

Źródło: *Computing Curricula 2004. Overview Report including A Guide to Undergraduate Degree Programs in Computing. A cooperative project of ACM, AIS, IEEE-CS* [on-line]. Shanghai Jiao Tong University. School Of Software [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://se.sjtu.edu.cn/sites/se/gb/CCSE/CCCS-040601-Overview-Strawman-Rev4.pdf.

nie nowych sposobów zastosowania komputerów (rozwój technologii sieciowych, baz danych i interfejsu człowiek-komputer przyczynił się do powstania sieci WWW), odkrywanie efektywnych metod rozwiązywania problemów komputerowych. Aktualna problematyka naukowców z tej dziedziny skupia się na metodach sztucznej inteligencji, zarządzaniu wiedzą (ang. *Knowledge Management*) i rozszyfrowywaniu kodu DNA.

Nauki komputerowe, jako dyscyplina akademicka na amerykańskich uczelniach pojawiły się w latach siedemdziesiątych XX w. Pierwotnie tej decyzji towarzyszyły ataki krytyki, kiedy próbowano zakwalifikować nauki komputerowe jako zawodową specjalność profesji technicznych lub jako pseudo dyscyplinę dla programistów. Kiedy w latach dziewięćdziesiątych XX w. *Computer Science* rozwinęła pola badań wiedzy i zastosowań w obszarze od teorii do praktyki to spory ucichły.

Inżynieria oprogramowania (termin wywodzi się z konferencji NATO w roku 1968) sprzed lat dziewięćdziesiątych ubiegłego wieku rozwijała się w ramach dyscypliny inżynieria komputerowa. Po tym okresie złożonych programów komputerowych już nie tworzyły pojedyncze osoby, lecz zespoły programistów, projektujące osobne moduły, które stanowiły jedną całość. Inżynieria oprogramowania zaczyna rozwijać się jako samodzielna dyscyplina od kiedy zauważono, iż niezbędne jest usprawnienie metod tworzenia dobrego i niezawodnego oprogramowania. Specyfiką inżynierii oprogramowania w odróżnieniu od innych dyscyplin inżynieryjnych jest nieempiryczna natura oprogramowania oraz dyskretny charakter operacji software'owych. A zatem rozwija się ona w oparciu o integrację zasad i praw matematyki oraz inżynierii doświadczalnej, opierającej się na artefaktach fizycznych.

Kluczową rolę w dynamice i perspektywach rozwoju dyscypliny *systemy informacyjne* odgrywają informacja oraz technologia, udoskonalająca narzędzia wytwarzania, przetwarzania i rozpowszechniania informacji. Eksperci od systemów informacyjnych skupiają się wokół rozwiązań integracji technologii informacyjnej i procesów biznesowych w celu sprostania informacyjnym wymogom przedsiębiorstw różnego profilu. Systemy informacyjne istniały już w latach sześćdziesiątych ubiegłego stulecia, kiedy specjalizowały się w takich potrzebach obliczeniowych, jak systemy rachunkowe i płacowe. Z końcem lat dziewięćdziesiątych XX w. posługiwanie się komputerami osobistymi, jako narzędziem pracy obserwuje się nie tylko wśród specjalistów technicznych, lecz innych pracowników na wszystkich szczeblach organizacji do przetwarzania i wymiany pomiędzy przedsiębiorstwami rośnie; pojawiają się problemy z zarządzaniem informacją i efektywność zarządzania procesami informacyjnymi staje się kwestią kluczową. Na etapie rozwoju systemów sieciowych, technologie komputerowe przejmują rolę środków komunikacji oraz współpracy pomiędzy organizacjami. To przyczynia się do polepszenia wydajności oraz problemów w infrastrukturze informatycznej takich jak np., uzależnienie od nowych miejsc pracy.

Technologia informacyjna (TI) wykreowała się, jako dyscyplina w późnych latach dziewięćdziesiątych ubiegłego wieku. Nie tylko w Polsce wyrażenie to figuruje w podwójnym znaczeniu. Według raportu[12] szersza treść TI obejmuje sferę wszel-

[12] Tamże.

kich obliczeń, a węższa – zagadnienia, odpowiadające wymaganiom technologicznym w branżach biznesowej, administracyjnej, medycznej i edukacyjnej. Zgodnie z treścią poprzedniego akapitu, TI jest obustronnym uzupełnieniem systemów informacyjnych, przy czym TI kładzie nacisk bardziej na technologię, niż na przekazywaną za jej pomocą informację. Sprawne funkcjonowanie każdego rodzaju organizacji uzależnione jest zarówno od systemów informacyjnych (które mają być właściwie dopasowane do specyfiki przedsiębiorstwa), jak i potencjału merytorycznego pracowników w branży technologii informacyjnych, co sprzyja umocnieniu pozycji tej dyscypliny. Wyzwaniem TI jest przygotowanie specjalistów, którzy są odpowiedzialni za wybór produktów hardware'owych i software'owych stosownie do profilu i zadań organizacji, za ich integrację z potrzebami infrastruktury organizacji, za instalację, obsługę i konserwację aplikacji komputerowych. Przykładowo, do obowiązków takich pracowników należą: instalacja, bezpieczeństwo i konserwacja sieci, administrowanie bazami danych i sieci, projektowanie stron WWW, rozwój aplikacji multimedialnych, instalacja modułów komunikacji, oraz logistyczne zarządzanie zasobami technologicznymi.

Nauki komputerowe w klasyfikacjach

Pierwsze klasyfikacje uniwersalne powstają w końcu XIX wieku w odpowiedzi na problemy organizacji zasobów bibliotecznych w warunkach szybko rozwijających nauk oraz rosnącej ilości drukowanych książek. Do klasyfikacji uniwersalnych zalicza się: klasyfikacja dziesiętna Dewey'a (*DDC – Dewey Decimal Classification*), uniwersalna klasyfikacja dziesiętna (*UDC – Universal Decimal Classfication*) oraz klasyfikacja opracowana przez Bibliotekę Kongresu USA (*LCC - Library of Congress Classfica-tion*).

Większość bibliotek narodowych używa klasyfikacji Dewey'a. *DDC* jest skomponowana numerycznie: cała wiedza jest dzielona na 10 głównych klas, które podlegają następnym podziałom specyfik przedmiotowych. Preferowanym systemem dla bibliotek uniwersyteckich i naukowych jest *LCC*, który łączy kodowanie alfabetyczne ze schematem numerycznym[13]. Polska Biblioteka Narodowa korzysta z klasyfikacji *UDC* oraz języka haseł przedmiotowych[14]. Obecnie wszystkie uniwersalne klasyfikacje są dostępne w postaci, rozumianej przez maszyny, na przykład w formatach: *XML*, *XHTML*, *RDF*. Pieczę nad rozwojem i dystrybucją *DDC* w formacie *MARC* (*Machine-Readable Cataloguing*)[15] zajmuje się organizacja *OCLC*[16]. Na łamach serwisu *OCLC* można zapoznać się z obecnymi badaniami. Zespół badawczy tworzą specjaliści bibliotekoznawstwa, lingwistyki

[13] *Library of Congress Classification* [on-line]. The Library of Congress [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.loc.gov/catdir/cpso/lcc.html.

[14] J. Sadowska, T. Turowska: *Języki Informacyjno-Wyszukiwawcze. Katalogi rzeczowe.* Warszawa: CUKB, 1990, s. 21-53.

[15] Format MARC jest międzynarodową umową, określającą sposób opisu danych bibliograficznych; cały opis jest podzielony na strefy, pola i podpola. Por. *Konwersja danych marc bn ===> MARC21* [on-line]. Warszawa: Biblioteka Narodowa, 2009 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://mak.bn.org.pl/wykaz5.htm.

[16] *Online Computer Library Center* [on-line]. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.oclc.org.

komputerowej, ekonomii, nauk komputerowych oraz pokrewnych. W obszarze organizacji wiedzy oraz jej inteligentnego zarządzania naukowcy szukają efektywnych i niezawodnych metod automatycznej klasyfikacji zasobów sieciowych oraz elektronicznych dokumentów z zastosowaniem schematów podstawowych klasyfikacji[17].

Zestawienie głównych klas działów komputerowychw różnych systemach klasyfikacyjnych **Tabela 2.**

ACM	LCC
A. General Literature B. Hardware C. Computer Systems Organization D. Software E. Data F. Theory of Computation G. Mathematics of Computing H. Information Systems I. Computing Methodologies J. Computer Applications K. Computing Milieux	Cybernetics Information theory Instruments and machines Calculating machines Electronic computers, computer systems Computer programming Computer software Computer security Computer algorithms Computer architecture Computer simulation Computer engineering, computer hardware
DDC	**UDC**
000 Computer science, information & general works 004 Data processing & Computer science 005 Computer programming, programs & data 006 Special computer methods 007 Not assigned or no longer used 008 Not assigned or no longer used 009 Not assigned or no longer used	004 Computer Science and Technology 004.2 Computer architecture 004.3 Computer hardware 004.4 Software 004.5 Human-computer interaction 004.6 Data 004.7 Computer communication 004.8 Artificial intelligence 004.9 Application-oriented computer-based techniques Cybernetics Q350-390 Information theory QA71-90 Instruments and machines QA75-76.95 Calculating machines QA75.5-76.95 Electronic computers. Computer science QA76.75-76.765 Computer software

Źródło: Opracowanie własne.

W projektach badawczych nad automatami klasyfikacyjnymi poruszane są następujące problemy[18]:

- wykorzystanie i rola technologii semantycznych, np. map tematycznych (ang. *topic maps)*;
- przybliżenie wyników automatycznej klasyfikacji do klasyfikacji wykonanej przez człowieka;

[17] B. Sosińska-Kalata: *Klasyfikacja*. Warszawa: SBP, 2002. s. 230-232.

[18] K. Golub: *Automated subject classification of textual Web pages based on a controlled vocabulary*. New Review of Hypermedia and Multimedia 2006, vol. 12, no. 1, s. 11-27.

- możliwość wyłonienia lub zdefiniowania metadanych w bazie wyników;
- zastosowanie automatów jako podstawowe narzędzie dla webmasterów lub w trakcie interakcji komputer-użytkownik.

Klasyfikacje uniwersalne cechują się również wolną asymilacją do nowych obszarów badań i zainteresowań. Według badaczy, dynamika pojęć i słowników większości cyfrowych hurtowni danych wyprzedza kontrolowaną przez człowieka indeksację zasobów. Klasyfikacje według geograficznego obszaru użytkowania można podzielić na cztery główne typy: uniwersalne, narodowe, przedmiotowe i własne (ang. *home-made*)[19]. Większość klasyfikacji przedmiotowo-specjalistycznych jest tworzona z uwzględnieniem konkretnych grup użytkowników. Zawierają one uporządkowane zbiory specjalne i/lub najważniejsze czasopisma i bibliografie, istotne dla danej dyscypliny naukowej. Mają one zapewnić rozbudowaną strukturę i terminologię, które byłyby bardziej aktualne w porównaniu z klasyfikacjami uniwersalnymi. Serwisy informacyjne, specjalizujące się w takich działach, jak medycyna, inżynieria czy rolnictwo opierają się na klasyfikacjach przedmiotowych lub ich kombinacji z uniwersalnymi schematami. Tabela 2 zestawia listy klas odnoszących się do nauk komputerowych w uniwersalnych systemach klasyfikacyjnych. Jeden rzut okiem wystarczy na uchwycenie różnorodności kategoryzacji przedmiotowej oraz odnotowanie ich rozbieżności w odniesieniu do klas podstawowych schematu klasyfikacji przedmiotowej *ACM*. Klasyfikacje przedmiotowe mają też minusy[20], a mianowicie:

- Powiązanie z serwisami specjalistycznymi z innych dziedzin jest utrudnione; do wymiany zasobów potrzebna jest aplikacja do konwersji.
- Jeżeli baza użytkowników jest mała, to użytkownikom serwisów innych dziedzin trudno jest zorientować się w strukturze klasyfikacji.
- Kolekcje zasobów specjalistycznych mogą włączać tematy z pogranicza lub interdyscyplinarne, których nie da się właściwie zaklasyfikować wewnątrz danego schematu.

W następnym podrozdziale zostanie przeanalizowana najważniejsza klasyfikacja przedmiotowa w dziedzinie informatyki, stworzona przez *Association for Computing Machinery*. Tamże będzie miało miejsce kilkakrotne odwoływanie się do zamieszczonej powyżej Tabeli 2 w celu porównania obiektów analizy. Wyniki badań niniejszej pracy, opisane w rozdziale 3 wykażą dobre i złe strony zastosowanej struktury klasyfikacji. Ponadto dzięki dokonanej wizualizacji zaproponowano nowatorską strategię ewaluacyjną.

[19] B. Sosińska-Kalata. dz. cyt., s. 23.

[20] *DESIRE Information Gateways Handbook* [on-line]. DESIRE (Development of a European Service for Information on Research and Education), 1998-2000 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.desire.org/handbook/.

2.2. Schemat klasyfikacji CCS ACM

a) Przesłanki historyczne

Do najpopularniejszej klasyfikacji piśmiennictwa w dziedzinie informatycznej należy system **CCS (Computing Classification System)**, stworzony i rozwijany przez *Association for Computing Machinery* (*ACM*). Najstarsze z towarzystw informatycz-nych – **ACM** zostało założone w 1947 roku i specjalizuje się w doskonaleniu wiedzy i umiejętności specjalistów i studentów technologii informatycznej z całego świata[21]. System *ACM CCS* uważany jest na świecie za standard w identyfikacji i kategoryzacji literatury komputerowej oraz działalności badawczej w zakresie informatyki. W Polsce system ten także zyskał akceptację kręgów naukowych i władz uczelni. Pierwotna wersja *CCS* została zaimplementowana w roku 1964, w latach 1982, 1983, 1987, 1991 i 1998 ukazały się kolejne uaktualnione wersje[22]. Wpis *Valid through 2009* w nagłówku ostatniej wersji świadczy o tym, iż dokonuje się w niej zmian na bieżąco. Do lat dziewięćdziesiątych XX w. *CCS* był nazywany *Computing Reviews Classification System* (*CRCS*), dlatego że stąd zapożyczono taksonomiczny schemat klasyfikacji[23].

Podobne znaczenie w klasyfikacji działów fizyki pełni schemat *PACS* (*Physics and Astronomy Clasification Scheme*), używany w *Physical Review* od roku 1975, a w matematyce – *MSC* (*Mathematical Subject Classification*). W latach 2000-2003 biblioteki narodowe państw europejskich wzięły udział w projektach na rzecz badań procesów mapowania schematów klasyfikacyjnych. Biblioteka Uniwersytecka w Regensburgu w ramach projektu *CARMEN* (*Content Analysis, Retrieval* and *MetaData: Effective Networking*) opracowała metodologię zintegrowanego wyszukiwania przedmiotowego rozproszonych zbiorów metadanych z uwzględnieniem koncepcyjnie zróżnicowanych tezaurusów i klasyfikacji[24]. Koncepcja opierała się na spostrzeżeniu, że użytkownik, oswojony z klasyfikacją lokalnego systemu informacyjnego (biblioteki) ma problemy z wyszukiwaniem przedmiotowym w zdalnych bibliotekach lub bazach referencyjnych. Jednym z celów było ustalenie prototypu zgodności podkategorii (ang. *cross concordances*) pomiędzy klasyfikacjami *DDC* i lokalną *RVK*[25] w obszarze przedmiotów specjalistycznych, m.in. fizyki i matematyki. Badania dotyczyły między innymi tego, jak pokrywanie się terminów technicznych

[21] *Association for Computing Machinery* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/.

[22] *ACM Computing Classification System* [on-line]. Association for Computing Machinery [dostęp 19 maja 2009]. Dostępny w World Wide Web:http://www.acm.org/about/class/.

[23] N. Coulter i in. *Computing Classification System 1998: Current Status and Future Maintenance Report of the CCS Update Committee* [on-line]. New York: ACM, 1998 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/about/class/ccsup.pdf.

[24] *WP12: Cross concordances of classifications and thesauri* [on-line]. *CARMEN: Content Analysis, Retrieval and MetaData: Effective Networking* [Universitätsbibliothek Regensburg] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.mathematik.uni-osnabrueck.de/projects/carmen/ index.en.shtml.

[25] *RVK – Regenburger Verbund Klassifikation.*

klasyfikacji specjalistycznych (*MSC* i *PACS*) może wpłynąć na implementację wyszukiwania strukturalnego w koncepcji równoległych drzew klasyfikacji.

Kolejnym znaczącym, w rozwiązaniach mapowań klasyfikacji, był projekt *RENARDUS*, zrealizowany w latach 2001-2002 i sfinansowany przez Komisję Europejską, jako część programu na rzecz rozwoju technologii oraz społeczeństwa informacyjnego[26]. W jego rezultacie opracowano narzędzia, służące do przeglądania i wyszukiwania przedmiotowego w zasobach rozproszonych bibliotek europejskich w oparciu o podstawową klasyfikację *DDC*.

Do pionierskich rozwiązań konwersji do poziomu klasyfikacji uniwersalnej zalicza się serwis sieciowy o kontrolowanym dostępie (ang. *gateway*) do wyszukiwania i przeglądania niemieckich zasobów internetowych w projekcie *GERHARD* (*German Harvest Automatem Retrieval and Directory*)[27]. Dokumenty, automatycznie gromadzone, konwertowane były zgodnie z klasyfikacją *UDC* z zastosowaniem algorytmów lingwistycznych i następnie indeksowane. W rezultacie program generuje drzewo przedmiotowe ułatwiające użytkownikom proces wyszukiwania.

Zmiany kolejnych wersji klasyfikacji *CCS* zostały wprowadzone pod redakcją N. Coultera, przewodniczącego komitetu uaktualnienia klasyfikacji. Znalazły się w nim takie osoby, jak były główny redaktor E. Sammet oraz A. Ralston, które wpłynęły bezpośrednio na rozwój poprzednich wersji klasyfikacji *Computer Review*. Członkowie komitetu, odpowiedzialni za zmiany, postawili sobie za cel stworzenie takiego systemu klasyfikacji[28], który po pierwsze, odzwierciadlałby aktualny stan rozwoju nauk komputerowych, po drugie, zawierałby proponowane przez redaktorów korekty i po trzecie, byłby oparty na mechanizmie, pozwalającym na łatwą rozbudowę i jednocześnie zachowywał główną strukturę drzewa. Idee, jakimi kierowali się członkowie komitetu w rozwijaniu systemu klasyfikacji *CCS*, opierały się na następujących założeniach:

- Rdzeniem systemu klasyfikacyjnego ma być drzewo klasyfikacji, które najlepiej nadawałoby się do reprezentowania hierarchicznej struktury formatów publikacji.
- Drzewo klasyfikacji ograniczone jest do 3 poziomów, aby móc opisać dokładnie zasadniczą strukturę dyscypliny.
- Deskryptor przedmiotowy, figurujący jako nienumerowany poziom czwarty, ma zapewnić procesy powstawania i rozbudowy nowych działów. Według pierwotnej koncepcji, określniki przedmiotowe miały podlegać częstym zmianom, lecz w praktyce trudno jest usunąć przestarzałe określniki ze względu na istniejące relacje z oryginałami sklasyfikowanych prac. Deskryptory przedmiotowe zatem pozostały jako permanentna część drzewa klasyfikacyjnego.

[26] T. Koch, H. Neuroth, M. Day. *Renardus: Cross-browsing European subject gateways via a common classification system (DDC).* W: *Subject Retrieval in a Network Environment: Papers Presented at an IFLA Satellite Meeting Sponsored by the IFLA Section on Classification and Indexing and IFLA Section of Information Technology, Dublin, Ohio, USA, 14-16 August 2001* [on-line]. Dublin, Ohio: OCLC, 2001 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.ukoln.ac.uk/metadata/renardus/ papers/ifla-satellite/ifla-satellite.pdf.

[27] B. Rieger, A. Kleber, E. von Maur. *Metadata-Based Integration of Qualitative and Quantitative Information Resources Approaching Knowledge Management* [on-line]. ECIS (European Conference on Information Systems) resources [London School of Economics and Political Science. Department of Information Systems] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://is2.lse.ac.uk/asp/ aspecis/20000115.pdf.

[28] N. Coulter i in., dz. cyt.

- Schemat klasyfikacji nauk komputerowych w całości został zaczerpnięty z taksonomii nauk komputerowych i inżynierii, opracowanej w 1980 r. przez Komitet Taksonomii AFIPS[29]. Późniejsze wersje były rozwijane przez specjalistów w różnych działach nauk komputerowych ze znaczącym wkładem pracy zarządu redaktorskiego pisma *Computing Reviews*.

Próba opisu historycznego, podziału nauk komputerowych zmusza do wyróżnienia głównych kierunków, których ewolucja doprowadziła do powstania współczesnej postaci nauk komputerowych. Kierunki tego rozwoju zostały zdeterminowane przez podstawowe sprawności: rachunkową, technologiczną i analityczną. Kierunki te odpowiadają semantycznym pojęciom podmiotowym: **obliczenia**, **maszyny** i **dane**, określanym jednoznacznie na każdym etapie rozwoju nauki.

W odniesieniu do ery komputerów, obliczenia oznaczają działalność natury technicznej wymagającej obecności komputera. Zatem, obliczenia są wykonywane przy projektowaniu i budowie systemów komputerowych, w przetwarzaniu i zarządzaniu różnego rodzaju informacją w badaniach naukowych z wykorzystywaniem komputera, w tworzeniu i eksploatacji mediów komunikacyjnych itp.

Historycznie pierwotne abstrakcyjne pojęcie obliczeń ewaluowało od pojęcia liczb naturalnych do operacji określanych abstrakcyjnymi działaniami określanymi na ciałach liczb zespolonych. Bez nich dzisiaj nie mogłyby się rozwijać np. nowoczesne technologie przetwarzania obrazu. Maszyny począwszy od abakusa czy maszyny różnicowej, a na technologiach **cloud computing**[30] skończywszy, rozwijały się zawsze zgodnie z postępem teoretycznych podstaw algorytmiki obliczeniowej. Technologie nadawały również asumpt swoistego sprzężenia zwrotnego dostarczając podstaw technologicznych do wykonywania obliczeń nie opracowanych jeszcze dokładnie przez algorytmikę obliczeniową – dzisiaj obserwujemy przykład takich zmian w rozwoju nanotechnologii w połączeniu z konstruowaniem komputera kwantowego.

Wreszcie dane – to zupełnie osobny temat, raczej struktura autonomiczna. Do niej dostosowujemy zarówno procesy obliczeniowe, jak i architektury platform obliczeniowych. Pierwotne dane numeryczne, zastąpiły obecnie zbiory danych o strukturach semantycznych. Wartości czysto numeryczne reprezentują więc często ukryte wzajemne powiązania – domenę dynamicznie się rozwijającej dziedziny analizy danych (ang. *Data Mining*). Analiza danych wielowymiarowych w ogólnych przestrzeniach topologicznych implikuje z kolei konieczność powstania nowych technologii wizualizacyjnych opartych często na możliwościach percepcyjnych człowieka. A te zagadnienia zaczynają wkraczać w domenę nauk psychologicznych.

---

[29] *Taxonomy of Computer Science and Engineering*. Ed. by A. Ralston. Arlington, USA: AFIPS (American Federation of Information Processing Societies) Press, 1982, s. 5-20.

[30] Cloud Computing (czyli „przetwarzanie w chmurze") – nowy sposób korzystania z technologii informatycznych. Określa zarówno FORMĘ DOSTĘPNOŚCI usługi przez szeroko rozumiany Internet, jak i nastawienie użytkownika (brak potrzeb ze strony użytkownika w zakresie posiadania dogłębnej wiedzy i kontroli dotyczących zasobu). Jednym z rozwiązań CC jest model *Pay-As-You-Go*, tzn.: użytkownik płaci tylko za te zasoby, które wykorzystywał. Por. *Cloud Computing*. W: *Wikipedia (English)* [on-line]. [dostęp 19.05.2009]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/Cloud_computing.

Opisane powyżej ogólne zależności implikują więc znacznie szerszy horyzont działalności niż pierwotnie określony. Mimo prostych definicji określających obliczenia jako działalność natury technicznej wymagającej obecności komputera, otrzymujemy w wyniku wzajemnej interakcji cały system zależności, wskazujący zarówno metody projektowania jak i implementacji systemów komputerowych, przetwarzanie i zarządzanie informacją oraz liczne inne drogi jej wykorzystania.

b) Rozwój schematu klasyfikacji CCS

Pierwsza odmiana klasyfikacji *ACM* z 1964 r. składa się z siedmiu klas głównych, zilustrowanych na Rysunku 22.

1. *General Topics and Education*	(Tematyka ogólna i edukacja)
2. *Computing Milieu*	(Środowisko obliczeniowe)
3. *Applications*	(Aplikacje)
4. *Programming*	(Programowanie)
5. *Mathematics of Computation*	(Matematyka i obliczenia)
6. *Design and Construction*	(Budowa i projektowanie)
7. *Analog Computers*	(Komputery analogowe)

Rysunek 22. Klasy główne klasyfikacji *CCS* w wersji z 1964 r.

Źródło: Na podst. *The 1964 Computing Reviews Classification System* [on-line]. Association for Computing Machinery [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/about/class/.

Przede wszystkim rzucają się w oczy definitywne różnice w terminologii i podziale pomiędzy ówczesną klasyfikacją a teraźniejszą (p. niżej), które można rozważyć w oparciu o 40-letnią historię rewolucji technologicznej. Na liście tematów nie ma jeszcze sekcji *hardware* i *software*; aktualnie podział taki figuruje w każdej nowoczesnej klasyfikacji lub tematycznych indeksach sieciowych serwisów. Epoka komputerów osobistych nadejdzie dopiero za 15 lat, w owym momencie tematyka sprzętu odpowiada zainteresowaniom wąskiego grona specjalistów. Kolejność pozycji na liście wskazuje na to, iż edukacji informatycznej (pod numerem 1) przypisywano ważną rolę w dziedzinie nauk komputerowych. Aktualnie, edukacja informatyczna jest problematyką osób związanych bezpośrednio z dydaktyką: nauczycieli, wykładowców, instruktorów i szkoleniowców (ang. *trainers*) oraz, co jest ważną cechą, ludzi kształcących się samodzielnie. Polskie konferencje informatyczno-dydaktyczne (np. „Informatyka w szkole") są ściśle odseparowane w zakresie tematycznym od biznesowych.

Dla współczesnych uczniów, którzy istnienie komputerów kojarzą z technologiami układów scalonych i systemem binarnym, ciekawostką historyczną bę-

dzie sekcja „Komputery analogowe". Komputer analogowy – jest to komputer przetwarzający sygnał ciągły (analogowy) przeważnie elektryczny. Wyróżniają go dwie podstawowe charakterystyki: 1) wykonywanie operacji w sposób równoległy – w jednym czasie jest w stanie wykonać wiele operacji; 2) posługiwanie się ciągłymi zmiennymi, a nie dyskretnymi, jak w przypadku cyfrowego. Komputery analogowe mają długa historię, która skończyła się w momencie skonstruowania pierwszego mikroprocesora (1968). Ze względu na łatwość i intuicyjność programowania komputerów analogowych, próbowano je naśladować w specjalizowanych komputerach cyfrowych np. cyfrowym analizator różnicowym *JAGA* oraz językach programowania i programach. Pierwszym polskim elektronicznym komputerem analogowym był ARR (Analizator Równań Różniczkowych)[31] Leona Łukaszewicza z 1953 r.

System *CCS ACM* składa się z dwóch głównych części: pierwszą, opisującą kategorię, jest 3-poziomowe numerowane drzewo z nienumerowanymi określnikami, druga zawiera listę terminów ogólnych (ang. *General Terms*). Każdy z listy 16. terminów ogólnych[32] (przykładowo: algorytmy, teoria, pomiary, języki, czynniki ludzkie itp.) wskazuje wątek tematyczny i nadaje się do uzupełnienia opisu dowolnego działu. W celu większej precyzji została wprowadzona lista ukrytych deskryptorów przedmiotowych (ang. *Implicite Subject Descriptors)* oraz nazw własnych *(Proper Nouns)*[33], które określają nazwy konkretnych produktów, systemów, języków oraz nazwiska eminentnych osób w informatyce. Dla przykładu, „C++" jest ukrytym deskryptorem przedmiotowym dla klasyfikacji: „D.3.2 Oprogramowanie. Języki programistyczne. Klasyfikacje języków programistycznych"; kolejny ukryty określnik „Bill Gates" odpowiada podklasie „K.2 Środowisko obliczeniowe. Historia Obliczeń". Lista ukrytych deskryptorów jest bardzo długa i dynamicznie rozrasta się w miarę dopisywania nowych nazw i nazwisk.

Schemat klasyfikacji kategorii jest oparty o 4-poziomową hierarchię taksonomiczną o jedenastu stałych klasach poziomu wyższego. Na każdym z pozostałych poziomów indeksowane pozycje mogą być dodawane, usuwane (ang. *retired*) albo korygowane (ang. *revised*). Etykieta oznacza, że termin ten nie będzie używany w procesach indeksowania, lecz figurować będzie jako deskryptor dla wcześniej przeindeksowanych zasobów. Według koncepcji autorów niższe poziomy mogą się krzyżować, a rozrost (lub zanik) gałęzi drzewa jest możliwy na niższych poziomach. Takim sposobem można łatwiej przystosować się do szybkich zmian w strukturze nauk komputerowych.

W Tabeli 3 została umieszczona tabela chronologicznych zmian w schemacie klasyfikacji *CCS* w czterech kolejnych wersjach włączając ostatnią (aktualną).

---

[31] Por. *Analizator Równań Różniczkowych*. W: *Wikipedia. Wolna Encyklopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/ARR.

[32] *The ACM Computing Classification System [1998 Version]. Valid through 2009* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/class/1998/.

[33] *List of Implicit Subject Descriptors in ACM CCS* [on-line]. *The ACM Portal* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://portal.acm.org/lookup/ccsnoun.cfm.

Historyczny przegląd zmian klasyfikacji *CCS* **Tabela 3.**

Rok	Poziom	Il. Pozycji dodanych	Il. Pozycji usuniętych	Il. Pozycji przemianow.	Il. Pozycji krzyżujących się
1983	2	2	0	0	2
	3	6	0	0	1
	4	16	0	1	
1987	2	0	0	0	0
	3	2	0	0	0
	4	16	0	0	
1991	2	1	0	1	0
	3	17	0	2	3
	4	103	4	8	
1998	2	1	0	3	0
	3	23	13	9	19
	4	246	171	32	

Źródło: N. Coulter i in. *Computing Classification System 1998: Current Status and Future Maintenance Report of the CCS Update Committee* [on-line]. New York: ACM, 1998 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/about/class/ccsup.pdf.

Zmiany te demonstrują przyśpieszoną w latach dziewięćdziesiątych XX w. w porównaniu z poprzednimi dekadami ewolucję nauk komputerowych. Klasyfikacja *CCS* wykorzystywana jest przede wszystkim w dostępnej na portalu stowarzyszenia bibliotece cyfrowej *ACM*[34], która zawiera przyzwoitą kolekcję abstraktów i tekstów publikacji (licząca 1.4 miliona stron tekstu) czasopism oraz materiałów konferencyjnych *ACM*.

c) Analiza aktualnej klasyfikacji CCS

Nazwy jedenastu klas głównych klasyfikacji *CCS* pokrywają się z głównymi kategoriami tematycznymi Taksonomii nauk komputerowych i inżynierii, opracowanej w latach osiemdziesiątych ubiegłego wieku[35]. Na Rysunku 23 przytoczona została ich lista:

A. *General Literature*	(Literatura ogólna)
B. *Hardware*	(Sprzęt)
C. *Computer Systems Organization*	(Organizacja systemów komputerowych)
D. *Software*	(Oprogramowanie)
E. *Data*	(Dane)
F. *Theory of Computation*	(Teoria obliczeń)
G. *Mathematics of Computing*	(Matematyczne metody obliczeń)
H. *Information Systems*	(Systemy informacyjne)
I. *Computing Methodologies*	(Metodologie obliczeniowe)
J. *Computer Applications*	(Aplikacje komputerowe)
K. *Computing Milieux*	(Środowisko komputerowe)

Rysunek 23. Klasy główne klasyfikacji *CCS* w wersji z 1998 r.

Źródło: Na podst. *The 1998 Computing Reviews Classification System* [on-line]. Association for Computing Machinery [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/about/class/ccs98-html.

[34] *The ACM Digital Library* [on-line]. *The ACM Portal* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://portal.acm.org/dl.cfm.

[35] A. Ralston, dz. cyt., s. 5-8.

Klasom głównym przyporządkowane są litery od **A** do **K** – czyli liczba klas na poziomie podstawowym wynosi 11, poziomy wyższe oznakowane są za pomocą alfanumerycznej kombinacji. Wykaz podklas dla klas pierwszego i drugiego poziomów rozpoczyna się od *General* i kończy się na *Miscellaneous*, dwóch kategorii o uniwersalnym przeznaczeniu. Podklasa „Ogólne" używana jest dla klasyfikacji zasobów o szerokim zakresie tematycznym mieszczącym się w kategorii klasy nadrzędnej. Można nią się posłużyć również dla określenia wszechstronnych zagadnień pokrewnych (równorzędnych dla danej podklasy) klas. Przykładowo, do podklasy „K.7.0 Ogólne" klasy „K.7 Zawód informatyka" zostaną zaklasyfikowane artykuły na tematy związane z profesją informatyk. Kategorii Ogólne można również użyć do materiałów z następujących grup tematycznych: „K.7.1 Zatrudnienie", „K.7.2 Organizacje" oraz „K.7.3 Testowanie, certyfikaty i licencje".

Dwie z głównych klas **B** i **D** dotyczą ściśle sprzętu i oprogramowania. Te dwa słowa pochodzenia obcego: *hardware* i *software* weszły w powszechne użycie na tyle, że nie drażnią już purystów językowych. *Hardware* – jest to fizyczna część komputera. Ogólnie hardwarem nazywa się sprzęt komputerowy w odróżnieniu od software, który jest zestawem instrukcji przeznaczonych do wykonania przez komputer. Podział ten, z upływem czasu coraz bardziej się zaciera, ponieważ współcześnie wiele elementów zestawu komputerowego ma już zaimplementowane oprogramowanie, stanowiące jego integralną część. Zapoznanie się z etymologią tych wyrazów pozwoli na zorientowanie się, który z nich został użyty pierwotnie w znaczeniu technicznym. Wyraz *software* narodził się w 1851 r. dla określenia fabryk wełny lub bawełny, a także relatywnie szybko psujących się dóbr konsumenckich (*soft* + *ware*). Jako neologizm o znaczeniu informatycznym, termin z odniesieniem do hardware zaczęto stosować od 1960 r. Wyraz hardware, pochodzący z 1515 r. (*hard* + *ware*) zyskał sens fizycznych komponentów komputera w 1947 r.

Podklasy pod nazwami „sprzęt" i „oprogramowanie" istnieją w klasyfikacjach *UDC* oraz *LCC*. W klasyfikacji *LCC* działy, związane tematyką komputerową są „porozrzucane" po kilku głównych klasach (p. Tabela 2). Na przykład, „Inżynieria komputerowa i sprzęt" należy do klasy **T** – Technologie, a „Oprogramowanie" – do klasy **Q** – Nauka, podklasy QA – Matematyka. Takich „przeskoków" można uniknąć, kiedy posługujemy się klasyfikacjami specjalistycznymi.

W klasie **C** o systemach komputerowych dostępne są działy, odnoszące się do architektury, projektowania sieci komputerowych oraz systemów rozproszonych. Ponadto są zamieszczone tu informacje o architekturach procesorów równoległych, pojedynczych i mnogich strumieni danych (multiprocesorów), jak również kategorie omawiające zagadnienia implementacji systemów komputerowych.

Naukowym poszukiwaniom wydajnych sposobów przechowywania i przetwarzania danych poświęcona jest odrębna klasa pod nazwą „E. Dane", gdzie są zamieszczone działy o strukturach, reprezentacjach, kodowaniu danych, a także teoria informacji, która powtarza się w niektórych podklasach klasy **H**.

Kolejne klasy **F** i **G** odnoszą się do bieżących problemów informatyki teoretycznej. W kategorie klasy „F. Teoria obliczeń" włączono: teorię obliczeń, analizę algorytmów, złożoność obliczeniową, logikę matematyczną oraz metody i języki formalne stosowane w projektowaniu i weryfikacji oprogramowania lub systemów

informatycznych. Pokrewna tematyka z zakresu zastosowań matematyki w informatyce zawarta jest w sąsiedniej klasie **G** o matematycznych metodach obliczeń: obejmuje ona analizę numeryczną, teorię grafów, teorię prawdopodobieństwa i statystykę oraz oprogramowanie służące wymienionym zagadnieniom.

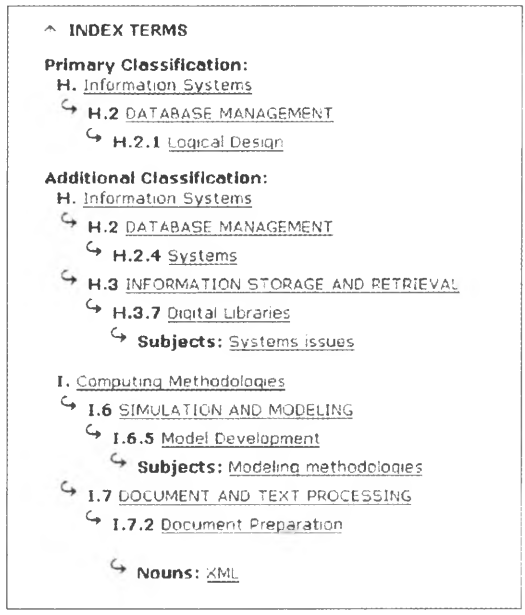
Klasa „H. Systemy informacyjne" obejmuje zagadnienia, związane z przechowywaniem, zarządzeniem i wyszukiwaniem informacji oraz reprezentacją i zastosowaniem systemów informacyjnych. System informacyjny w szerszym znaczeniu oznacza system ludzi, danych oraz komponentów technologicznych ukierunkowany na procesy przetwarzania informacji. Kluczowymi w dzisiejszych systemach informacyjnych są zastosowane technologie, głównymi funkcjami których są między innymi obróbka dużych baz danych, wykonanie złożonych obliczeń oraz kontrola wykonywanych procesów. Informacja i wiedza są obecnie traktowane jako czynniki współzależne. W literaturze informatycznej można natknąć się na przypadki zawężenia sensu systemu informacyjnego do systemu, opartego na sprzęcie i oprogramowaniu, z pominięciem zasobów ludzkich, elementów organizacyjnych, baz wiedzy itp. – czyli systemu informatycznego.

Zakres klasy **H** pokrywa się tematycznie ze współczesną problematyką informacji naukowej. Tak więc, zagadnienia informacji naukowej figurują na głównym poziomie drzewa *CCS*. Badanie procesów informacyjnych zachodzi z wykorzystaniem gotowych produktów informatycznych (np., oprogramowania, sieci, maszyn), liczne przykłady z życia zawodowego dowodzą, iż granica pomiędzy informatyką a informacją jest rozmyta. Nieprzypadkowo te dwie dziedziny: zarówno komputery, jak i informacja naukowa są połączone w obrębie jednej z głównych klas 000 klasyfikacji *DDC* (p. Tabela 2). Kilka z podklas klasy **H** odnosi się bezpośrednio do aktualnych kierunków badań naukowych w zakresie bibliotek cyfrowych i sieciowych serwisów informacyjnych.

Klasa „I. Metodologie obliczeniowe" obok typowej treści, ukierunkowanej na rozwiązywanie i analizę problemów przy pomocy technik informatycznych odwołuje się do grafiki komputerowej, przetwarzania i rozpoznawania obrazów, przetwarzania tekstu, symulacji i modelowania, a także sztucznej inteligencji. Sztuczna inteligencja w schemacie *CCS* pojawia się jako podklasa I.2 klasy **I**. Kategoriami równorzędnymi do niej są między innymi: „Grafika komputerowa", „Symulacja i modelowanie", „Przetwarzanie dokumentów i tekstu", „Rozpoznawanie obrazów". Podkategoriami podklasy „Sztuczna inteligencja" są: „Aplikacje i systemy eksperckie"; „Programowanie automatów"; „Dedukcja i dowody teoretyczne" (ang. *Theorem Proving*); „Metody i formalizmy reprezentacji wiedzy"; „Języki programistyczne i oprogramowanie"; „Uczenie się"; „Przetwarzanie języków naturalnych"; „Rozwiązywanie problemów"; „Wyszukiwanie i metody kontroli"; „Robotyka"; „Systemy percepcji wizualnych" (ang. *Vision and Scene Understanding*); „Rozproszona sztuczna inteligencja". Kategoria „Aplikacje i systemy eksperckie" dzieli się z kolei na kartografię, gry, automatyzację przemysłową (ang. *industrial automation*), prawo, medycynę i naukę, interfejsy języków naturalnych, automatyzację biura. Taki podział nie odzwierciedla dobrze aktualnych kierunków rozwoju tej kategorii. Kartografia na przykład mogłaby należeć do podklasy „Aplikacje geograficzne", która powinna być dodana do klasy „Medycyna i nauka".

Na bazie aktualnego drzewa *CCS*[36] można poznawać nowe trendy w nauce oraz zestawienia i relacje między nimi. Dla przykładu, tematyka sieci komputerowych zawarta jest w katalogu C. *Computer System Organization*, a zainteresowani problematyką sztucznej inteligencji, przetwarzaniem i analizą obrazów cyfrowych, modelowaniem i symulacją, rozpoznawaniem obrazów powinni odwołać się do działu I. *Computing Methodologies*. Nauki przyrodnicze, medyczne, ich pochodne, wszystkie integracyjne z metodami komputerowymi należą do klasy zastosowań informatyki: J. *Computer Applications*. Tematykę związana z historią komputerów i obliczeń, edukacją informatyczną, a również poruszająca aktualne aspekty społeczne i etyczne informatyki (na przykład problem legalności użytkowania zasobów sieci P2P) zrzesza klasa ostatnia: „K. Środowisko obliczeniowe".

Sukces tak dobrze zorganizowanego schematu polega na współpracy z autorami publikacji, którzy najlepiej są rozeznani w tematyce swoich prac. Oprócz klasyfikacji podstawowej należy podać pomocniczą, zlokalizowaną w innej klasie bądź podklasie na drzewie klasyfikacyjnym. Zachęca się autorów do opisania kategorii i słów kluczowych w wysyłanym dokumencie. Pomocną jest lista rzeczowników ukrytych deskryptorów przedmiotowych. Na podstawie podanych danych można określić główne i dodatkowe drzewa taksonomiczne klasyfikacji *CCS*. Rysunek 24 ilustruje wygenerowane schematy do przykładowego artykułu z 2006 r. pod tytułem *XML Schema and Objectrelational Modeling of Multimedia Digital Libraries*.



Rysunek 24. Klasyfikacje głowna oraz dodatkowe przykładowego artykułu pod tytułem *XML Schema and Object-relational Modeling of Multimedia Digital Libraries*

Źródło: opracowanie własne.

[36] *The ACM Computing Classification System*, dz. cyt.

Wymienione zostały tu klasyfikacje: główna o symbolu H.2.1 oraz dodatkowe: H.2.4, H.3.7, I.6.5, I.7.2 wraz z deskryptorami tematycznymi na niektórych poziomach oraz nazwą własną (*XML*). Ten przykład drzewa taksonomicznego ilustruje krzyżowanie się dwóch klas głównych (**H, I**) oraz ich podklas.

Powyższa kategoryzacja wykorzystywana jest nie tylko do artykułów naukowych. Osoby, które zdecydują się na członkowstwo w *ACM* przy wypełnianiu formularza muszą określić rodzaj swojej działalności oraz wybrać obszar zainteresowań w zakresie technologii informatycznych. Kategorie tematyczne w elektronicznym formularzu pokrywają się ze strukturą *CCS*, dlatego czytelnicy i wydawcy materiałów czasopism *ACM* mają ułatwione zadanie poruszania się w gąszczu tematów nauk komputerowych.

d) Ontologia w klasyfikacji CCS

Aktualne schematy klasyfikacyjne obiektów (tematów badań w kontekście tego rozdziału) lub pojęć w opracowaniach naukowych budowane są w oparciu o ontologie – sposoby formalizacji wiedzy.

Ontologia zajmuje się odkrywaniem i opisywaniem „tego co jest" – w rzeczywistości, w umysłach ludzi i zapisanego w postaci różnych symboli[37]. Termin ten pojawia się w kontekście informatycznym już w roku 1967 i dotyczy modelowania danych, ale dopiero w dobie nadmiaru informacji i konieczności jej wymiany oraz maszynowego przetwarzania informacji zyskał szersze zainteresowanie. W ostatnich dwóch dekadach wysiłki naukowców w sferze automatyzacji procesów organizacji wiedzy koncentrują się wokół projektowania **systemów eksperckich** (systemów z bazą wiedzy). Są to zestawy programów, wspomagające korzystanie z wiedzy i ułatwiające podejmowanie decyzji. Głównym ich celem jest wspomaganie lub zastępowanie ludzkich ekspertów w danej dziedzinie. Inżynieria wiedzy (ang. *Knowledge Engineering*) powiązana jest z takimi obszarami jak bazy danych, przetwarzanie danych (ang. *data mining*), systemy eksperckie, systemy podejmowania decyzji (ang. *decision suport systems*) i systemy informacji geograficznej (*GIS*). Inżynieria wiedzy wspierana jest badaniami w dziedzinie logiki; zawartość wiedzy w systemach zautomatyzowanych konstruowana jest w oparciu o pojmowanie ludzkiego myślenia i wnioskowania. Węższą domeną inżynierii wiedzy jest **Inżynieria Ontologiczna** (ang. *Ontological Engineering*)[38], nowa subdyscyplina powstała w ramach sztucznej inteligencji.

Jednoznaczny przekaz wiedzy na temat ontologii nauk komputerowych wykorzystuje hierarchizację, czyli umiejscowienie określonej klasy w hierarchicznej strukturze drzewa; klasa posiada cechy dziedziczone z klas nadrzędnych. W literaturze angielskiej dla nazwania przykładu klasy używa się terminu *instance* – **instancja**. W języku programowania obiektowego każdy obiekt klasy dziedziczący cechy tej

[37] Por. *Ontologia*. W: *Wikipedia. Free Encycklopedia* [on-line]. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Ontologia.

[38] B. Sosińska-Kalata: *Struktury klasyfikacyjne w organizacji zasobów informacyjnych Internetu* [on-line]. W: *Multimedialne i sieciowe systemy informacyjne. Materiały konferencyjne*. Pod red. Cz. Danilowicza. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2002 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.zsi.pwr.wroc.pl/zsi/missi2002/pdf/s403.pdf.

klasy, nazywany jest instancją tej klasy. Mózg ludzki jest w naturalny sposób najlepiej przystosowany do obiektowej analizy rzeczywistości przy przetwarzaniu informacji. Arystoteles, przedstawiciel racjonalnego poznawania rzeczywistości, wprowadził pojęcia formy (odpowiednika platońskiej idei) i materii[39]. Formie (lub idei) w programowaniu obiektowym odpowiada klasa, materii – instancja – obiekt. W analizie informacji płynącej z otaczającego nas świata dążymy do sklasyfikowania występujących w rzeczywistości obiektów w grupy – klasy. Czynimy to na podstawie wspólnych ich cech (wygląd, przeznaczenie, pochodzenie itp.) lub funkcji i zachowań (co obiekt może wykonać?). Obiekty w trakcie poznawania coraz większej ich ilości grupujemy w klasy i definiujemy klasy nadrzędne. Atutem programowania obiektowego jest wykorzystywanie w najlepszy sposób właściwości naszego mózgu – panowanie nad większymi strukturami daje bardziej obiektywny wgląd na rzeczywistość i pełniej ją naśladuje. Zatem, programy lepiej naśladujące procesy rzeczywistości, będą lepiej z nią współpracowały. Realizacja graficznych projektów w programach 3D zaczyna się od budowania modeli światów wirtualnych, następnie tworzone są scenerie środowiska, później trzywymiarowe obiekty architektoniczne itp. Które wpisywane są autonomiczne obiekty 3D. Zasada definiowania instancji (ang. *instantiation*) wywodzi się z koncepcji Platona, zakładającej, że jeśli istnieje cecha, powinien istnieć i nośnik tej cechy.

Treści ontologiczne nie są odwzorowaniem taksonomii obiektów, lecz stwarzają formalne przesłanki wedle których takowe mogą być budowane. Jeden z postulatów ontologii głosi: „Naturalne jest istnienie wielu ontologii – uznanie braku możliwości stworzenia jednej ogólnej ontologii". Badacze semantyki lansują różne ontologie, podejścia do ich budowania, standardy itp. W inżynierii ontologicznej można spotkać się z przybliżeniem modelowania wieloperspektywicznego (ang. *Multi-Perspective Modelling*)[40]. Według autorów, pojedyncze ontologie nie są w stanie przekazać pełnego opisu badanego obiektu, a zatem kompletna reprezentacja pojęcia lub obiektu wymaga użycia co najmniej sześciu ontologii[41], zbudowanych według przymiotów (perspektyw): kto, co, jak, gdzie, kiedy i dlaczego, które mogą powtarzać się na różnych poziomach abstrakcji. Treści tych perspektyw w domenie wiedzy są charakteryzowane następująco:

- „co" – opisuje zazwyczaj zasoby określonego rodzaju;
- „jak" – odwołuje się do metod i technik;
- „kto" – inteligentny program – agent, wpływający na zachowanie człowieka (*agent*);
- „gdzie" – wskazuje powiązania zewnętrzne;
- „kiedy" – włącza kontrolę i ograniczenie;
- „dlaczego" – wymienia uzasadnienie i cele.

[39] Por. *Arystoteles*. W: *Wikipedia. The Free Encyclopedia.* [on-line]. [dostęp 19.05.2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Arystoteles.

[40] J. Kingston: *Ontologies, Multi-Perspective Modelling and Knowledge Auditing* [on-line]. CEUR Works Workshop Proceedings [RWTH Aachen University. Informatik 5] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-48/kingston.pdf.

[41] Tamże.

J. Kingston w kolejnej pracy[42] szczegółowo analizuje możliwości zastosowania teorii przybliżenia multiperspektywicznego do struktury klasyfikacyjnej *CCS ACM*. Testy zostały przeprowadzone na prośbę pracownika Biblioteki Sztucznej Inteligencji Uniwersytetu w Edynburgu. Biblioteka ta przez lata korzystała z klasyfikacji *ACM* wraz z rozszerzonym działem sztucznej inteligencji.

Wzrost zainteresowania tematyką zarządzania wiedzą ze strony organizacji komercyjnych i badawczych spowodował napływ publikacji i materiałów o tej tematyce i klasyfikacja według schematu *ACM* już nie dawała satysfakcjonujących rezultatów. W rozszerzonej wersji klasyfikacji *CCS* dla działu „I.2 Sztuczna Inteligencja"[43], do podklas tej klasy trzeciego poziomu zostały dodane „Modelowanie kognitywne oraz psychologiczne studia sztucznej inteligencji" oraz „Zagadnienia społeczne i filozoficzne", a podklasa „Rozproszona sztuczna inteligencja" zastąpiona została przez „Specjalistyczne architektury sztucznej inteligencji". Podklasa skierowana na problematykę aplikacji – „Aplikacje i systemy eksperckie", dzielona jest na 19 subkategorii (7 z nich – są zaproponowane przez *ACM*), a te ostatnie – na liczne podkategorie niższego poziomu.

Klasyfikacja *ACM* według analizy wieloperspektywicznej[44] wykorzystuje trzy z wymienionych perspektyw wiedzy. Kategorie klasyfikacji odpowiadają trzem różnym poziomom abstrakcji (Tabela 4). Niektóre dotyczą samego komputera oraz jego „wnętrza" (hardware, software, organizacja systemów komputerowych, dane, systemy informacyjne), inne komputer traktują jako samodzielne pojęcie w kontekście aplikacji obliczeniowych (metodologie obliczeniowe, aplikacje komputerowe, środowisko obliczeniowe). Trzeci poziom objawia się w dwóch teoretycznych kategoriach: Teoria obliczeń oraz Matematyczne metody obliczeń, które zapewniają zastosowanie podstawowych technik w organizowaniu systemów komputerowych, danych i systemów informacyjnych.

Kategorie najwyższego poziomu klasyfikacji *CCS*, pogrupowane zgodnie z przybliżeniem wieloperspektywicznego modelowania **Tabela 4.**

	What	How	Why	When	Where	Who
Computer applications	Computer applications	Computing Methodologies	Computer milieux			
What goes inside computer	Hardware, Software	Computer Systems Organization, Data, Information Systems				
Theoretical Level		Theory of Computation, Mathematics of Computing				

Źródło: J. Kigston. *Ontology, Knowledge Management, Knowledge Engineering and the ACM Classification Scheme* [on-line]. University of Edinburgh. School of Informatics [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.inf.ed.ac.uk/publications/ online/0169.pdf.

[42] J. Kigston: *Ontology, Knowledge Management, Knowledge Engineering and the ACM Classification Scheme* [on-line]. University of Edinburgh. School of Informatics [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.inf.ed.ac.uk/publications/online/0169.pdf.

[43] Tamże.

[44] Tamże.

W oparciu o konkretne przykłady J. Kigston udowadnia, że klasyfikacja *ACM* oraz powiązane rozszerzenie do klasyfikacji sztucznej inteligencji (*Scientific Datalink's Extention*) zbudowane są zgodnie z dwiema lub trzema zasadami strukturyzacji. Charakteryzują się one strukturą taksonomiczną, bazującej na perspektywie wiedzy „Co" (podkategoria jest tym, w czym znajduje zastosowanie kategoria nadrzędna, innymi słowy są to zasoby w ogólnym tego słowa znaczeniu) oraz perspektywie „Jak" (tj. techniki i metody w celu uzyskania kategorii nadrzędnej). Artykuł potwierdza założenie modelu wielo-perspektywicznych ontologii oraz wskazuje pozycje w klasyfikacji do poprawnego zaklasyfikowania następujących działów: zarządzanie wiedzą, inżynieria wiedzy oraz przyswajanie wiedzy.

Klasyfikacja *CCS* jest jednym z głównych narzędzi adnotacji i wyszukiwania publikacji w zasobach naukowych o tematyce informatycznej takich jak np. biblioteka cyfrowa *ACM*. Powyższa praca oraz kilka innych z ostatnich lat demonstrują próby naukowców (między innymi jako użytkowników klasyfikacji) modernizacji schematu klasyfikacji za pomocą nowoczesnych metod inżynierii wiedzy[45]. Innym celem badań jest znalezienie odpowiedniego modelu ontologii klasyfikacji, który byłby kompatybilny z dynamiką zmian nauk komputerowych.

e) Aktualizacja klasyfikacji CCS

Komitet aktualizacji *CCS* konsekwentnie pracuje nad zmianami w schemacie klasyfikacyjnym *CCS*, czego dowodzą etykiety: *New*, *Revised* i *No longer used* pojawiające się w odpowiednich kategoriach drzewa[46]. Struktura dyscyplin komputerowych zmieniała się błyskawicznie w ciągu ostatnich dwóch dekad. W dokumentach oficjalnych *ACM* w tym kontekście nadużywa się przymiotnika *dramatic*, podkreślając duży wymiar problemu z aktualizacją taksonomii, terminologii i definicji. Pierwotne założenia przy tworzeniu klasyfikacji sprawiają, iż korekta jest możliwa na trzech poziomach taksonomii *CCS* (wyłączając klasy główne) poprzez skreślenie, konsolidację i przemianowanie węzłów (ang. *nodes*) w strukturze drzewa.

Nadal poważnym dylematem, jest koherentna identyfikacja wyższych poziomów klasyfikacji w różnych dokumentach związanych z przedmiotem nauki komputerowe. Na najwyższym poziomie nie były uwzględnione nowo powstałe gałęzie takie, jak nauki obliczeniowe i interakcja człowiek-komputer. Z drugiej strony, klasa główna „E. Dane" na miarę współczesnych zagadnień stała się pozycją nierelewantną; charakteryzuje się długim czasem indeksacji oraz ubogim zestawem podklas. Właściwie, dane wszelkiego formatu funkcjonują wewnątrz jakiejś struktury organizacyjnej (bazy danych, macierze, strumienie itp.), rzadko występują pojedyncze obiekty – dane.

[45] B. Mirkin, S. Nascimento, L. M. Pereira: *Representing a Computer Science Research Organization on the ACM Computing Classification System* [on-line]. *CEUR Workshop Proceedings* [RWTH Aachen University. Informatik 5] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-354/p19.pdf.

[46] *The ACM Computing Classification System (1998)* [on-line]. Association for Computing Machinery [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/about/class/ccs98-html.

Przyjrzyjmy się, jak zorganizowana jest treść w czwartej edycji *Encyklopedii nauk komputerowych* (*Encyclopedia od Computer Science*) z 2000 r. – niekwestionowanego autorytetu w literaturze komputerowej i technicznej[47]. Książka mieści zwięzłe objaśnienia najnowszych technologii informatycznych wraz z przykładami ich zastosowania praktycznego. Naświetlone są tu również historyczne motywy pojawienia się przełomowego punktu w rozwoju Nauk komputerowych i technologii. Poruszanie bogatego zestawu zagadnień: perspektywy historyczne, aktualna wiedza i przewidywane trendy w bliskiej przyszłości lokuje encyklopedię na półce klasyki referencji naukowych.
Rozdziały encyklopedii są pogrupowane następująco:

- Hardware,
- Software,
- Computer Systems,
- Information and Data,
- Mathematics of Computing,
- Theory of Computation,
- Methodologies,
- Applications,
- Computing Milieux.

Łatwo tu jest zauważyć bliską analogię do organizacji klas głównych klasyfikacji *CCS*. W konstrukcji taksonomii nauk komputerowych od wielu lat bierze udział ten sam skład osób (głównym redaktorem encyklopedii jest A. Ralston, który także tworzył *Taxonomy Computer Science and Engineering*[48] – widzimy tu konsekwentne trzymanie się zasady historycznej ciągłości w konstruowaniu schematu.

Autorzy systemu *CCS* i osoby (mogą to być kompetentni użytkownicy serwisu) biorące udział jego aktualizacji przy opracowaniu zmian doceniają znaczenie historycznych ciągłości zarówno w procesach wyszukiwania informacji, tak i dla poprawnego funkcjonowania obecnego oprogramowania. Każde unowocześnienie struktury wymaga włączenia elementów mapowania z zachowaniem archiwalnej integralności oraz ergonomiczności; struktura taka jednocześnie powinna być przejrzystą dla przeciętnego użytkownika, wyszukującego zasoby. W raporcie Komitetu aktualizacji *CCS* z 1998 r.[49] przytoczona została wizja przyszłego systemu *CCS* na 2018 r. Ma być to dynamiczny system, odzwierciedlający rzeczywistość aktualnych lat, jak również korzystający z porad systemów eksperckich, dotyczących każdego przekroju obszaru nauk komputerowych, a także opierać się o techniki automatyzacji zachodzących zmian.

[47] *Concise Encyclopedia of Computer Science...*
[48] *Taxonomy of Computer Science and Engineering...*
[49] N. Coulter i in., dz. cyt.

2.3 Systematyka przedmiotu w serwisach sieciowych

Coraz więcej zasobów literaturowych, również naukowego przeznaczenia zamieszcza się w internecie, w specjalizowanych serwisach sieciowych. Udostępniane są one często na zasadach praw *Creative Commons*[50] albo innych form wolnego dostępu. Funkcje pomocniczą, jeśli nie główną, w wyszukiwaniu i eksploracji zamieszczonych w takich systemach dokumentów pełnią katalogi systematyczne z określonym poziomem hierarchii. Różnorodność używanych schematów jest proporcjonalna do ilości rozpatrywanych serwisów – taką mieszaninę klasycznych i amatorskich klasyfikacji obserwuje się na dzień dzisiejszy. Problemem manualnego zarządzania strukturą hierarchiczną takich katalogów jest potrzeba ciągłej aktualizacji. Natomiast jeśli redaktorzy (bądź oprogramowanie) serwisów w trosce o ich aktualność będą stosować zbyt częste zmiany w drzewie kategorii, to zmienna lokalizacja dokumentów i spowolnienie działania systemu z powodów powtarzających się indeksacji może zniechęcić użytkowników do korzystania z takich stron, nie mówiąc już o niewygodach takiego podejścia.

W obliczu rozrastającego ruchu Web 2.0 społeczności aktywnych użytkowników wnoszą również swój wkład w systematykę treści zasobów sieciowych. W Tabeli 5. przedstawiona jest organizacja głównych działów przedmiotu nauki komputerowe w serwisie *Wikipedia*. Prezentuje ona według mnie bardzo logicznie ułożoną strukturę, składająca się z 12 kategorii. Podkategorie tematyczne znajdują się w prawej kolumnie. Należy zwrócić uwagę, że na poziomie głównym nie znalazło się miejsca dla kategorii „sprzęt" i „oprogramowanie". Zapewne uległy one tematycznemu rozdrobnieniu i zostały zaadoptowane w innych działach. Dwie pierwsze kategorie zawierają zagadnienia informatyki teoretycznej: „Teoria obliczeń" oraz „Struktury danych". Językom programowania słusznie przeznaczono osobny dział. Niezwykle aktualna jest kategoria „Typy obliczeń" wraz z podkategoriami: obliczenia *Grid* i *Cloud*, klastrowe, rozproszone, równoległe. Architektury komputerów, systemów informatycznych oraz informacyjnych zostały włączone do jednego działu. Użytkowanie sieci komputerowych połączono z zagadnieniami telekomunikacji. W kategorii „Bazy danych" zawarte są odnośniki do teorii baz danych oraz tematów dużych repozytoriów danych – *Data Mining*, *GIS*, *OLAP*. Popularność grafiki komputerowej jest niezaprzeczalna – dlatego też znajduje się w osobnej kategorii. Wyodrębniony w tym schemacie jest również nowy dział w nauce – „Obliczenia naukowe" (ang. *scientific computing*) poszukujący modeli matematycznych oraz rozwiązań technicznych problemów naukowo-inżynieryjnych. Sztuczna inteligencja powinna zostać samodzielną kategorią w naukach komputerowych, co się zgadza w danym schemacie.

[50] *Creative Commons* (CC) – licencja działająca na zasadzie „pewne prawa zastrzeżone". Granice dozwolonego użytku są szersze i wyraźniejsze niż te wytyczone na zasadzie „wszelkie prawa zastrzeżone". CC szanuje prawo twórców do określenia stopnia, w jakim chcą się dzielić swoją twórczością z innymi. Jednocześnie zachęca do tworzenia wspólnej kultury, której elementy mogą być swobodnie wymieniane i zmieniane. Por. *Udostępniaj swoją twórczość na jasnych, przyjaznych zasadach – na licencjach CC* [on-line]. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://creativecommons.pl/.

Organizacja głównych działów nauk komputerowych w Wikipedii. **Tabela 5.**

Działy	Podkategorie
Theory of computation	*Automata · Computability · Computational complexity · Quantum Computing*
Algorithms and Data structures	*Analysis of algorithms · Algorithm design · Computational geometry · Interface*
Programming languages	*Theory · Compilers / Parsers / Interpreters · Programming paradigms (Declarative · Imperative · Logic · Procedural) · SDLC · Software Distribution*
Types of Computation	*Cloud computing · Cluster Computing · Distributed computing · Grid computing · Parallel computing*
System architecture	*Computer architecture · Computer organization · Operating systems · Management information system · Information systems*
Telecomm & Networking	*Broadcasting · Network topology · OSI model · Cryptography · World Wide Web · Semantic Web · Internetworking · PSTN / SONET · IEEE 802*
Security	*Intelligence · Encryption · Protocols · Spam · VPN · Online predator · Identity theft · Internet privacy · Trusted Computing · Advertising Ethics · Computer forensics · Computer surveillance · DoD*
Databases	*Database theory · Data mining · Data modeling · OLAP · Geographic information system*
Computer graphics	*CGI · Visualization · Image processing*
Scientific computing	*Artificial life · Bioinformatics · Cognitive Science · Computational chemistry · Computational neuroscience · Computational physics · Numerical algorithms · Symbolic mathematics*
Artificial intelligence	*Automated reasoning · Computational linguistics · Computer vision · Evolutionary computation · Machine learning · Natural language processing · Robotics · Cybernetics*
BCI / HCI / MMI	*Computer accessibility · User interfaces · Wearable computing · Ubiquitous computing · Mixed reality*

Źródło: *Computer Science*. W: *Wikipedia. The Free Encyclopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/Computer_science# Fields_of_computer_science.

Wikipedyści dla własnej i innych wygody założyli w przestrzeni *Wikipedii* projekt pod nazwą *WikiProject_Computers_Structure*[51] mający na celu ujednolicenie różnorodnych struktur wewnętrznych portalu *Wikipedia* odnoszących się do tematu „Informatyka" przy równoległym wsparciu toczonych dyskusji. Pomysł ten narodził się w odpowiedzi na problem taksonomii *Computer Science*, z których żadna nie sprawdziła się w dłuższym okresie czasu. Projekt na razie znajduje się w stadium początkowym, twórcy zdają sobie sprawę z długotrwałości przedsięwzięcia i potrzeby koordynacji pracy organizacji oraz zainteresowanych jednostek. W aspekcie rozwarstwienia tematycznego *Wikipedii* wydzielono trzy główne tematy badań: Komputery, Obliczenia (*Computing*) i Technologię Informacyjną.

Na obecnym etapie wykoncypowano, iż najwygodniej będzie istniejące artykuły informatyczne podzielić według następujących kategorii:

[51] *WikiProject Computing*. W: *Wikipedia. The Free Encyclopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Computers# Structure.

1. History of computing,
2. Computer Science and Programming,
3. Networking and Internet,
4. Server Technology,
5. Databases,
6. Hardware,
7. Computer security,
8. Software,
9. Operating Systems.

W tym podziale uwzględniono częstość odwołań do danego tematu, czyli popularność, jego wagę w aktualnym stanie rozwoju nauk komputerowych oraz perspektywiczność. Po raz pierwszy historię komputerów potraktowano na równi z „klasycznymi" kategoriami – dowodzi to, iż potrzebne są opracowania opisujące historię informatyki i maszyn liczących. Niefortunna nazwa drugiej kategorii rozmywa pojęcie nauk komputerowych, ale najpewniej mają być tu zawarte treści bardziej „naukowe" w odróżnieniu od szeroko rozumianej informatyki stosowanej. Kategoria 7. jest bardzo ważnym zbiorem zagadnień związanych z szacowaniem i kontrolą ryzyka wynikającego z korzystania z komputerów, sieci i przesyłania danych do zdalnych lokalizacji.

Jak widać, edytorzy *Wikipedii* czynią śmiałe przymiarki do stworzenia środowiska do wygodnych eksploracji i przeszukiwania danych. Nasuwa się pytanie czy aktywni są w tych działaniach polscy użytkownicy? Polskie wersje systematyki nauk komputerowych postanowiono przez autorkę niniejszej rozprawy doktorskiej zbadać w najbliższej przyszłości. A niewątpliwie będzie to szerokie pole do przeanalizowania. Autorka skupiła się w tym rozdziale na angielskojęzycznych wersjach systemów klasyfikacyjnych i systematykach sieciowych. Informatycy, również polscy stanowią społeczność dla której język angielski nie jest problemem. Polscy informatycy wysyłając swoje prace naukowe do anglojęzycznych czasopism, kategoryzują swoje prace za pomocą klasyfikacji *ACM*. Z polskim systemem klasyfikacyjnym mają styczność dopiero przy komunikowaniu się z Radą Naukową MNiSW (dawniej KBN).

B. Sosińska-Kalata wnikliwie zbadała problem stosowanych w serwisach internetowych klasyfikacji, zwanych *home-made*. W trakcie studiów i analizy klasyfikacji specjalistycznych spotykamy się z pytaniem, dotyczącym ich celowości i przeznaczenia. Czy możemy schematy klasyfikacji przedmiotowych zastosować do zasobów wyspecjalizowanych serwisów sieciowych? Według spostrzeżeń autorki artykułu: *coraz szerzej jako instrument organizacji zasobów Internetu wykorzystuje się znane klasyfikacje piśmiennictwa, których popularyzację w sieci wspierają instytucje zarządzające ich rozwojem – Biblioteka Kongresu, OCLC, Konsorcjum UKD i British Standards Institution oraz organizacje nadzorujące rozwój klasyfikacji specjalistycznych*[52].

Takie podejście demonstruje wiele zalet. Hierarchiczne struktury ułatwiają nawigację zasobów; użytkownik może intuicyjnie manewrować poprzez poszerzenie lub zawężenie przeszukiwanej strefy. W przeglądaniu tematycznie ułożonych wy-

[52] B. Sosińska-Kalata. *Struktury klasyfikacyjne...*

szukiwanych terminów częściowo znikają problemy z homonimami. Użytkownik ma do czynienie z posegregowaną bazą danych. Ważnym atutem jest uniezależnienie się od języka: niektóre schematy pozwalają na wielojęzykowy dostęp do kolekcji. Użytkownik wpisuje wyrażenie wyszukiwawcze we własnym języku, które zaprogramowany „przełącznik" tłumaczy na język schematu klasyfikacyjnego; w procesie generowania wyników sytuacja się odwraca. Większość uznanych schematów klasyfikacyjnych jest odporna na dezaktualizację dzięki stałemu nadzorowi przez odpowiedzialne za ich rozwój instytucje. Schematy klasyfikacyjne w serwisach sieciowych mogą zawierać odwołania do poklasyfikowanych zasobów innych serwisów, tym samym jest realizowana konwersja pomiędzy klasyfikacjami.

W odpowiedzi na postawiony problem nasuwa się wniosek o potrzebie badań nad metodami klasyfikacji automatycznej. Warto na zakończenie przytoczyć cytat B. Sosińskiej-Kalaty o istocie problemu: *Pytanie, które dzisiaj należy postawić nie dotyczy więc sensu tworzenia struktur klasyfikacyjnych zasobów sieci, ale tego, jak efektywnie te struktury generować, jak je sprawnie aktualizować i jak efektywnie operować nimi w opracowaniu ogromnych, silnie zróżnicowanych i bezustannie zmieniających się zasobów sieci. Jest oczywiste, że metody manualnej klasyfikacji nie mogą rozwiązać problemu. Kwestią kluczową pozostaje zatem wypracowanie efektywnych metod automatycznej klasyfikacji informacyjnej zawartości sieci*[53].

[53] Tamże.

Rozdział 3

OPIS PRAC BADAWCZYCH

3.1. Objaśnienie podstawowej metodyki badań

Przyglądając się dendrogramom (p. Rozdział 1.2) i różnym schematom, przedstawiającym struktury hierarchiczne, widzimy, iż informacja jest tam organizowana w kierunku góra – dół (pionowo). Klasy podstawowe i główne kategorie przekazują swoje właściwości klasom podrzędnym: zasadę tą najłatwiej jest zrozumieć na przykładach genetycznego dziedziczenia cech (genotypu) rodziców przez dzieci. W automatycznej klasteryzacji proces grupowania informacji przebiega w kierunku odwrotnym[1]: obiekty (np. dokumenty) są grupowane początkowo na poziomie najniższym, następnie definiowane są kategorie.

Reprezentowanie hierarchii za pomocą drzewa cechuje się jednym wymiarem i liniowością[2]. Pierwsza cecha oznacza, że położenie dowolnego węzła w takim drzewie da się określić za pomocą jednego parametru, zawierającego np. zakodowaną informację o numerach węzła i jego klas(y) na odpowiednich poziomach hierarchii (przykład poniżej). Dodanie jeszcze jednego wymiaru powoduje zmapowanie struktury hierarchicznej na płaszczyznę, w taki sposób B. Schneiderman wykonał pierwszą implementacje stosując strategię zagnieżdżonych prostokątów oraz J. Stasko za pomocą koncentrycznych pierścieni (Rozdział 1.2, Rysunek 7b). W ostatnim przykładzie korzeń hierarchii (klasa macierzysta) umieszczono w środkowym okręgu, a klasy podrzędne – w kolejnych okręgach zewnętrznych ułożonych miarę zwiększania się odległości od ich wspólnego środka. Położenie węzła w takim kulistym diagramie określa się za pomocą dwóch współrzędnych biegunowych:

[1] J. Gelernter: *Visual Classification with Information Visualization (Infoviz) for Digital Library Collections.* Knowledge Organization 2007, nr 34, s. 128-143.

[2] Zdolność do uzyskiwania wyników pomiaru analitycznego wprost, za pomocą prostych, proporcjonalnych operacji matematycznych.

promienia r oraz wartości kąta nachylenia. Jeśli kolor jest intuicyjnym narzędziem do znakowania podstawowego atrybutu, np. klasy/kategorii głównej, to wymiary płaszczyzny pozwalają dodatkowo wykorzystać wartości pól zagnieżdżonych obszarów (prostokątów, kwadratów, segmentów, łuków) do wskazywania np. rozmiarów lub liczebności obiektów. W przypadku wycinków kołowych należy wziąć poprawkę na to, że pole wycinka jest proporcjonalne do kwadratu promienia koła, na skutek czego te wartości muszą być traktowane jako względne dla każdego poziomu hierarchii. Kołowy diagram ma tą ważną zaletę, iż nie narzuca ograniczenia poziomów hierarchii, strukturę można rozbudowywać w kierunku wzrostu promienia właściwie nieograniczenie.

W trakcie poszukiwania docelowej przestrzeni informacyjnej rozważone były także możliwości systemów wizualizacji hierarchii pod względem przeglądania i wyszukiwania danych. Nawigacja w obrębie drzewa hierarchicznego jest bardzo niewygodna. Wystarczy przypomnieć jak działa dowolny menadżer plików. W danym momencie dostępna jest zawartość tylko konkretnego katalogu, aby przejrzeć inne poziomy, trzeba zmienić pozycję w drzewie katalogowym. Na mapie natomiast mamy wgląd w ogół danych na wszystkich poziomach i tym samym możliwość ich porównania i wykrycia nieprawidłowości. Poza tym, na rozmieszczenie obiektów wpływa jedna współrzędna więcej, niż w przypadku liniowego drzewa, co daje dużo więcej możliwości. Powstał pomysł, aby dla zachowania ciągłości topologicznej, badany schemat klasyfikacji zmapować na zakrzywioną płaszczyznę. Ostatecznie zdecydowano się na sferę, jako powierzchnię o absolutnej symetrii pozwalającej stworzyć ergonomiczną przestrzeń wizualizacyjną. Jest to wygodny i intuicyjny kształt dla systemu percepcyjnego użytkownika w procesach nawigacji i przeglądania obiektów.

W erze grafiki 3D odpowiednikami przestrzennych koncepcji wizualizacji są projekty, wykorzystujące przestrzeń kuli. Warto tu scharakteryzować jeden z nowo-czesnych i wygodnych programów pod nazwą *Walrus* firmy *Caida*[3]. Jest to otwarte narzędzie do interaktywnej wizualizacji dużych, skierowanych grafów w trójwymiarowej przestrzeni. Rysunek 13 (Rozdział 1.2) ilustruje zrzut ekranowy wizualizacji hierarchii struktury katalogowej. Program Walrus miał być wykorzystany w celu renderowania[4] wizualizacji danych w początkowym stadium niniejszych prac. Ze względu na takie ograniczenia jak: brak różnicowania kąta rozpiętości drzewa (jednolity algorytm) oraz wartości poszczególnych krawędzi (te same dla całego drzewa), odstąpiono od tego pomysłu.

W powstałej koncepcji przeniesienia struktury drzewa na płaszczyznę zrodziła się potrzeba określenia topologii, czyli sąsiedztwa wizualizowanych obiektów, które składają się z: klas, podklas wszystkich poziomów klasyfikacji oraz poklasyfikowanych dokumentów. W myśl podstawowej zasady wizualizacji, im większe podobieństwo pomiędzy obiektami względem wcześniej zdefiniowanych cech, tym bliżej siebie są one umieszczone na generowanej mapie. Niepodobne do siebie obiekty

[3] *CAIDA: The Cooperative Association for Internet Data Analysis* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.caida.org.

[4] W programach graficznych 3D komputerowa analiza trójwymiarowej sceny i utworzenie na jej podstawie dwuwymiarowego obrazu wyjściowego w formie statycznej lub w formie animacji.

usytuowane są na dużej odległości. Podstawowym problemem w danej koncepcji stało się wyznaczenie sensownej metryki podobieństwa w oparciu o metadane dokumentów publicznie dostępnych w bibliotece cyfrowej *ACM*.

Struktura klas głównych i podklas *CCS* została poddana ściślejszej analizie, podczas której uwzględniono system notacyjny, przypisany do klas i podklas, składający się z litery głównej klasy, i liczb porządkowych podklas każdego poziomu, np. B.1.8, C.1.2, A.1.m. Jeśli za pomocą takiego zakodowanego indeksu można określić położenie węzła na drzewie, to matematyczne różnice indeksów, obliczone w sposób liniowy (np. odległość pomiędzy A.1.0 oraz B.4.6 wynosi 1.3.6) absolutnie nie scharakteryzują podobieństwa pomiędzy klasami. A zatem za pomocą bezpośredniego wykorzystania metadanych, metodami liniowymi, trudno znaleźć logiczny algorytm na obliczenie bliskości badanych obiektów: klas, podklas oraz dokumentów w nich zawartych.

W trakcie badań stwierdzono, iż wytypowana do wizualizacji, klasyfikacja literatury informatycznej *CCS* z powodów wymienionych w rozdziale 2 nie spełnia żadnego z warunków podziału klasyfikacji logicznej[5]. Występowanie wspólnych dokumentów dla wielu par klas wskazuje na brak rozłączności. Na klasycznym dendrogramie gałęzie – metafory klas/podklas – ze wspólnymi elementami muszą być skrzyżowane, co robi takie zakresy mało czytelne graficznie. Nadmiar sumy zakresów klas podrzędnych względem zawartości klasy głównej (brak adekwatności) świadczy o występowaniu dokumentów na każdym stopniu struktury z największą ich ilością na najniższych poziomach drzewa. Ta właściwość zaburza hierarchię i trudno ją wykoncypować w spójnej formie graficznej. Ponownie dostrzegamy powody przeniesienia struktury klasyfikacyjnej na płaszczyznę. Te dwie cechy, a raczej ich brak potraktowano jako pozytywne, mogące pomóc w znalezieniu metryki, podobieństwa.

Główną ideą w określeniu podobieństwa klas pod względem reprezentowanej tematyki było uwzględnienie liczby ich wspólnych dokumentów. Przyjęto, iż tematyczna bliskość klas będzie proporcjonalna do liczby ich wspólnych dokumentów. **Im bliższe podobieństwo tematyczne pomiędzy klasami (podklasami), tym więcej zawierają one wspólnych dokumentów. I odwrotnie, dokumenty nie powielają się w klasach o odmiennej tematycznie zawartości.** Należałoby tu odwołać się do wcześniej wzmiankowanych wniosków wynikających z terminologii analizy bibliometrycznej (p. Rozdział 1.3): jeśli prefiks 'co' oznacza wspólne występowanie danych, to w danym przypadku można wprowadzić nowe definicje – **„koklasy"** (ang. *co-classes*).

Następstwem takiego nieliniowego podejścia w ustaleniu podobieństwa klas jest policzenie wspólnych dokumentów dla każdej pary klas i podklas i skonstruowanie z unormowanych danych macierzy podobieństwa. Poza bardzo nielicznymi przypadkami, wszystkie klasy, notowane w klasyfikacji podstawowej, występują w klasyfikacjach dodatkowych. A zatem, wyjściowa macierz powinna być symetryczna; a jej wymiar – równy liczbie wszystkich możliwych klas i podklas w schemacie klasyfikacji. Jest istotne, że operujemy nie na wektorach cech obiektów, które zazwyczaj

[5] B. Sosińska-Kalata. *Klasyfikacja...*, s. 21-23.

są poddawane kolejnej obróbce, lecz poprzez założenie zależności podobieństwa klas od liczby wspólnych dokumentów, dzięki czemu otrzymujemy macierz podobieństwa w sposób bezpośredni, wykorzystując dane pomiarowe. Według tej zasady, elementy macierzy dla absolutnie niepodobnych klas przyjmują wartość zero. Elementy diagonalne macierzy natomiast dla zachowania symetrii powinny równać się 1, co uzyskano poprzez normalizację. Ten warunek oznacza, iż każda klasa/podklasa jest tożsama sama ze sobą. Wyjściowa macierz podobieństwa charakteryzowała się mocno dyskretnym rozrzutem wartości: od częstego wystąpienia wartości zerowych po liczby mieszczące się w przedziale [0,1].

Popularną techniką, stosowaną do redukcji wielowymiarowych danych jest skalowanie wielowymiarowe (*MDS*), które rozmieszcza badane obiekty w przestrzeni o zadanej liczbie wymiarów, np. 3 i sprawdza, na ile ta nowa konfiguracja odwzorowuje podobieństwa pomiedzy obiektami[6]. W kolejnym etapie za pomocą MDS przeprowadzono redukcję wymiaru danych do trzech, aby móc zbudować geometryczną reprezentację głównych węzłów klasyfikacji. Otrzymano kartezjańskie współrzędne wykoncypowanej sfery. Przy założeniu, że badane obiekty są punktowe i poddawane działaniu siły odpychającej skierowanej od środka układu – centroida, można odwołać się do praw wywodzących się z fizyki molekularnej. Wykorzystano w tym celu potencjał *Morsa*[7] – często stosowany spektroskopii molekularnej model energii potencjalnej dwuatomowej molekuły, zademonstrowany na Wykresie 1. Potencjał *Morsa* przyjmuje minimalną wartość (czyli stan równowagi molekuły) w odległości r_e pomiędzy cząsteczkami:

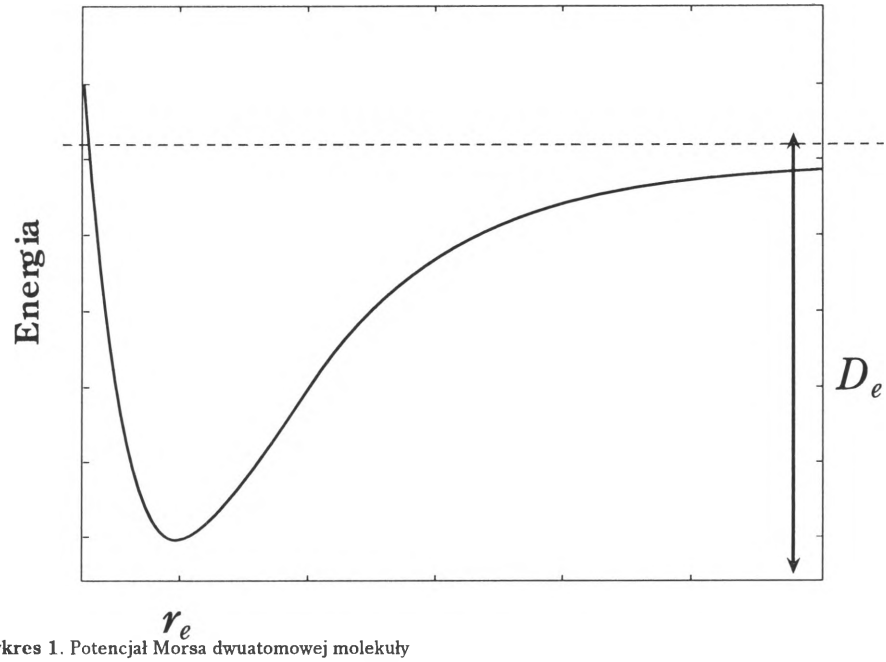
$$E_s(r) = D_e\left[\left(1 - e^{-b(r-r_e)}\right)^2 - 1\right] \tag{1}$$

gdzie wielkość D_e – nazywana głębokością studni potencjału i określa ona energię dysocjacji (rozpadu) atomów, a parametr b kontroluje szerokość studni.

Według powyższej idei po pierwsze „przyciągnięto" obiekty (traktując je jako punktowe) do powierzchni sfery o uśrednionym promieniu, po drugie „zasymulowano" siłę ich wzajemnego odpychania.

[6] *Elektroniczny Podręcznik Statystyki PL* [on-line]. Kraków: StatSoft, 2006 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.statsoft.pl/textbook/stathome.html.

[7] I. G. Kaplan. In *Handbook of Molecular Physics and Quantum Chemistry*. Ed. by S. Wilson. New York: Wiley, 2003, Vol. 3, s. 207.



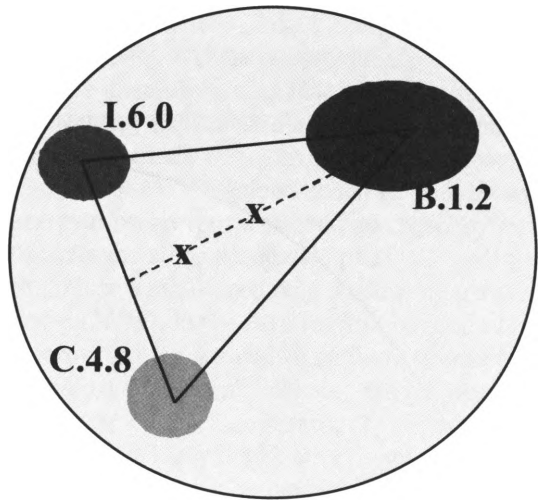
Wykres 1. Potencjał Morsa dwuatomowej molekuły
Źródło: Opracowanie własne.

Jako funkcję korelacji do zbioru danych, zastosowano potencjał *Morsa*, który opisuje energię potencjalną dwóch oddziałujących cząsteczek.

Po rozmieszczeniu węzłów klas i podklas na powierzchni sfery obliczono lokalizację pozostałych licznych obiektów – dokumentów. Jeśli węzły klas i podklas tworzyły trójkąt lub wielokąt, to węzeł dokumentu będzie się znajdował wewnątrz takiej figury. W przypadku płaskiej figury będzie to środek ciężkości. Na Rysunku 25 położenie wynikowego środka ciężkości, zaznaczonego jako „x" będzie zależało od wag klasyfikacji podstawowej i dodatkowych, które ustalono w sposób empiryczny – wyniosło ono 0.6:0.4.

Do kodowania wielowymiarowych danych na sferze użyto trzech atrybutów: barwy – do określenia klasy głównej, jasność koloru – do znakowania poziomu hierarchii klasy i geometryczny rozmiar węzła, który wskazywał liczebność klasy/podklasy. W wizualizacji 11. klas głównych i ich pochodnych (dokumentów włącznie) przydatna okazała się być paleta 12. podstawowych kolorów, uwzględniająca właściwości ludzkiej percepcji[8]: czerwony, zielony, żółty, niebieski, różowy, turkusowy, szary, pomarańczowy, brązowy, czarny, fioletowy, biały (nieużyteczny ze względu na tło aplikacji).

[8] C. Ware, dz. cyt., s. 123-143.



Rysunek 25. Zasada znajdowania pozycji węzła dokumentu ze znanych wartości współrzędnych węzłów klas/podklas. Symbol „X" oznacza położenie wybranych dokumentów.

Źródło: opracowanie własne.

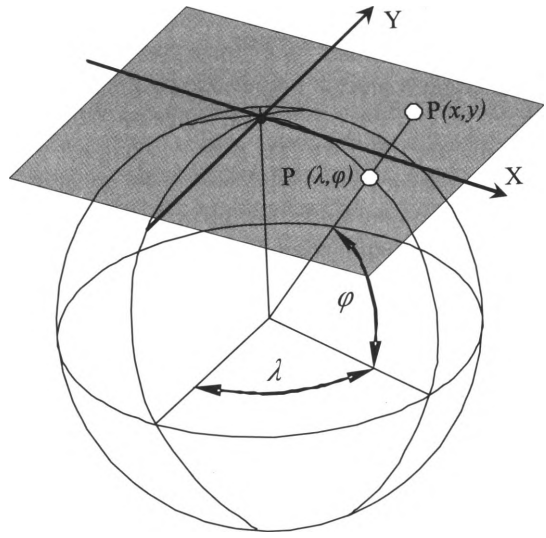
Zestawy wybranych glyfów (p. Rozdział 1.1), kodujących dokumenty układały się w kolorowe klastry. Struktura tych ostatnich na mapie sferycznej odwzorowywała tematyczną organizację zbadanej literatury informatycznej według klasyfikacji *CCS*. W celu przeprowadzenia wygodniejszej analizy i poprzedzającej ją obróbki graficznej posłużono się rzutem kartograficznym. Ze względu na czytelność mapy, wybrano odwzorowanie walcowe równoodległościowe, w którym powierzchnia sfery jest rzutowana na powierzchnię boczną walca i następnie rozwijana do płaszczyzny[9]. Równoodległościowy oznacza, że punkty o równej odległości od równika na sferze, znajdą się w równej odległości od równika na wyjściowej płaszczyźnie. Równikiem nazwiemy największy przekrój, powstały wskutek przecięcia badanej kuli płaszczyzną prostopadłą do hipotetycznej osi obrotu i przechodzącą przez środek kuli (Rysunek 26). Ponieważ odległość od równika jest szerokością geograficzną φ, za pomocą prostych wzorów otrzymujemy współrzędne x i y na płaszczyźnie:

$$x = \alpha\beta(\lambda - \lambda_0) \quad (2)$$

$$y = \alpha\phi, \quad (3)$$

gdzie λ jest długość geograficzna (Rysunek 26), λ_0 – południk (przekrój prostopadły do równika) przechodzący przez środek mapy, α 0 stała skalowania mapy, β – stała determinująca proporcję wymiaru pionowego do poziomego.

[9] *Odwzorowanie walcowe*. W: *Wikipedia. Wolna Encyklopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Odwzorowanie_walcowe. Por. z metodą *odwzorowania Merkatora* na s. 52-53 niniejszej pracy.



Rysunek 26. Współrzędne geograficzne

Źródło: Na podst. D. Małyszko. *Systemy informacji przestrzennej* [on-line]. Politechnika Białostocka 2007. [dostęp 19.05.2009]. Dostępny w World Wide Web: http://aragorn.pb.bialystok.pl/~dmalyszko/GIS_Materialy/SIP_Zajecia/SIP_Odwzorowania.htm

Powstałe mapy punktowych rozproszonych obiektów pomimo wyraźnej kolorowej klasteryzacji nie były pozbawione szumów. Zastosowano algorytmy cyfrowego przetwarzania obrazów; do odszumiania graficznej reprezentacji użyto filtra mediany. Jest to technika nieliniowa, zwykle używana do usuwania szumów i zakłóceń w sygnałach[10]. Filtr ten pobiera wartości jasności pikseli obrazu oryginalnego, sortuje je i wykorzystuje wartość środkową, która jest wpisywana do pola macierzy wynikowej, tak jak pokazano to w poniższym przykładzie:

1. Pobrane zostały 9 wartości z następującego pola macierzy.

2	100	98
55	45	34
25	5	11

2. Następnie je posortowano:

2	5	11	25	[redacted]	45	55	98	100

3. Wartość środkowa – 34 zastąpiła pierwotną macierz 9-iu liczb.

Filtr mediany eliminuje te piksele, dla których wartość jasności znacznie odbiega od wartości pozostałych pikseli w polu (to są tak zwane szpilki). Takim sposobem na mapie usunięto pojedyncze, odległe od większych skupisk punkty, które

[10] W. Malina, M. Smiatacz: *Metody cyfrowego przetwarzania obrazów*. Warszawa: EXIT 2005, s. 70-78.

zaburzały wynikową wizualizację klastrów. Kolejnym filtrem, który został wykorzystany, był filtr wykrywania konturu (ang. *trace contour*). Ten algorytm dokonuje detekcji krzywych składających się z pojedynczych pikseli dookoła fragmentów o wyraź-nym kontraście względem otoczenia, które sprowadza się do koloru tła[11]. Dzięki funkcji wygładzenia możliwy jest płynny gradient intensywności koloru od maksymalnego w obszarach o największej gęstości obiektów do minimalnego o małej, lecz ponadprogowej ilości punktów. Metoda ta pozwala na wychwycenie istotnych wzorów na mapie wizualizacji, a również na śledzenie zależności we właściwościach klastrów.

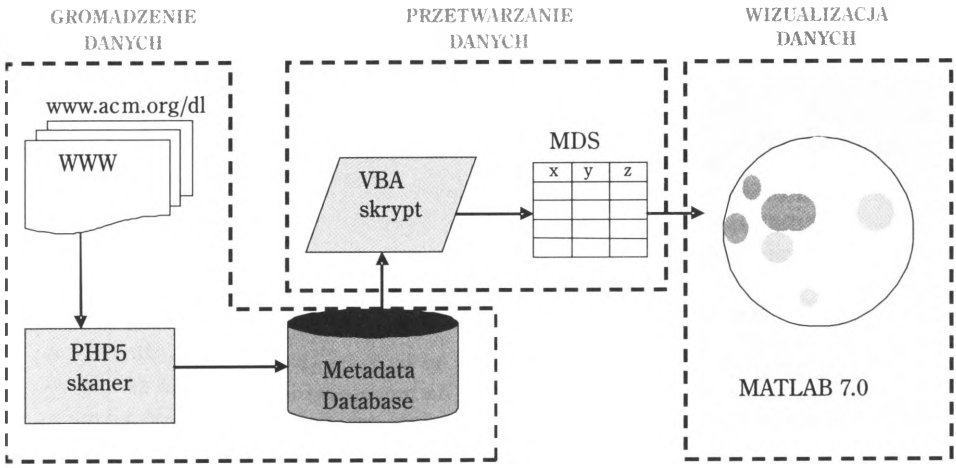
Wykształcone w wyniku obróbki graficznej klastry można dalej analizować pod względem zawartych w nich dokumentów. Tym razem wykorzystano takie metadane, jak deskryptor przedmiotowy i słowa kluczowe. Z obróbki statystycznej tych danych klastrom można przypisać zestaw najbardziej charakterystycznych słów kluczowych. Otrzymamy w ten sposób mapę semantyczną badanego zestawu dokumentów. Ich klasyfikacja odegrała kluczową rolę w pierwotnej wizualizacji (tematycznej). Drugi etap wizualizacji (semantyka) przeprowadzono wykorzystując słowa kluczowe. W poprawne funkcjonowanie biblioteki cyfrowej *ACM* wkład wnoszą zarówno autorzy wysyłanych publikacji, tak i edytorzy serwisu. Na autorach, jako osobach najlepiej orientujące się w treści prac, spoczywa odpowiedzialność za dobór właściwych słów kluczowych. W klasyfikowaniu artykułów biorą udział obie strony, np. redaktorzy dokonują korekty przydzielonych klas. Ten ludzki czynnik przy obopólnej weryfikacji powoduje, iż proponowany sposób wizualizacji wydaje się być bardziej wiarygodny, niż metody oparte na automatycznej klasteryzacji. Pomimo intensywnie rozwijających się technologii sztucznej inteligencji, maszyny wciąż nie są w stanie zastąpić ludzi w zadaniach precyzyjnej kategoryzacji i klasyfikacji dokumentów.

3.2. Przebieg procesu podstawowej analizy danych

a) Narzędzia i implementacja

Całościowy proces kompletowania, analizy oraz wizualizacji danych składał się z trzech etapów, przedstawionych na Rysunku 27, które szczegółowo będą opisane w następnych podrozdziałach.

[11] I. Pitas, A.N. Venetsanopoulos: *Nonlinear digital filters: principles and applications*. Boston, USA: Kluwer Academic Publishers, s. 63-76, 236-237.

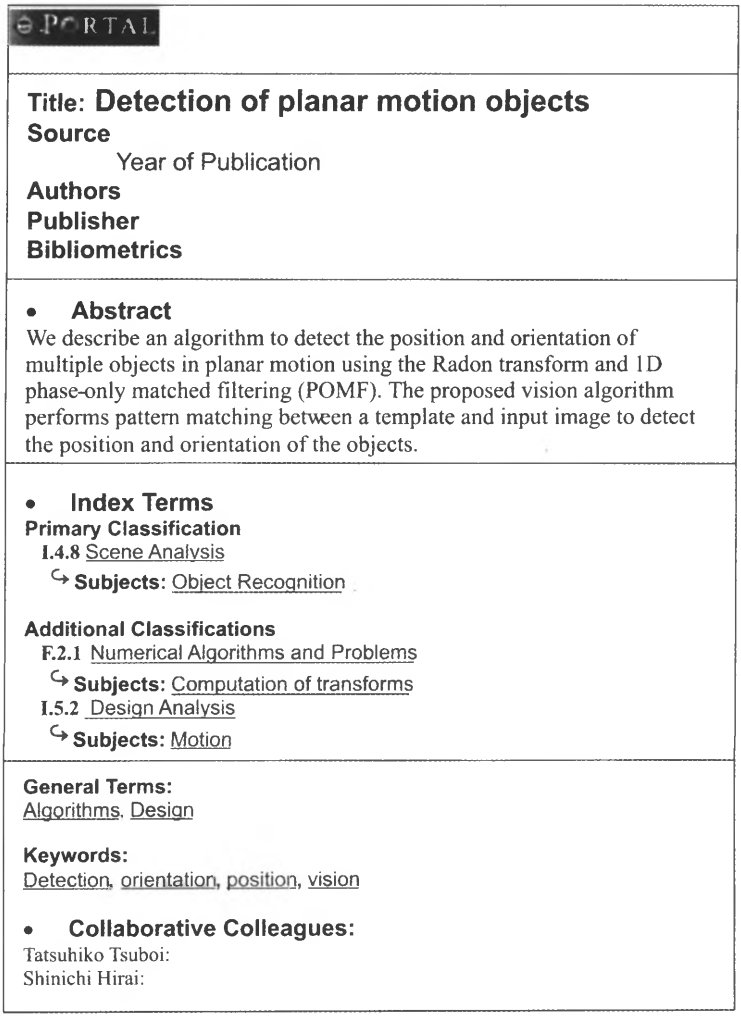


Rysunek 27. Schemat etapów procesu wizualizacji

Źródło: opracowanie własne.

Kompletowanie metadanych dokumentów ze stron biblioteki cyfrowej *ACM* zostało przeprowadzone skanerem sieciowym – aplikacją napisaną w języku PHP 5.0 i uruchomioną na serwerze Apache Web Server. Dane w postaci *ASCII* stanowiły wyjściową bazę metadanych. Użyto programu *Excell 2007* oraz języka *Visual Basic for Application* (*VBA*), aby dokonać ich pierwotnej obróbki, policzyć wszystkie występujące w bazie danych klasy i podklasy.

Plik z gotową macierzą podobieństwa zaimportowano do programu *Matlab* w wersji 7.0. Jest to interaktywne środowisko obliczeniowe przeznaczone do modelowania, wykonywania obliczeń naukowych i inżynierskich, oraz do tworzenia symulacji komputerowych. Ponieważ podstawowym typem danych programu *Matlab* są macierze, było to doskonałe narzędzie do dalszej obróbki danych dwuwymiarowych. Za pomocą skryptów w środowisku *Matlab 7.0* zaprojektowano interaktywną aplikację do wizualizacji wyników.



Rysunek 28. Postać analizowanych dokumentów w bibliotece cyfrowej *ACM*

Źródło: opracowanie własne.

Następne obliczenia takie jak statystyka słów kluczowych wykonane były na metadanych pierwotnych za pomocą języka *VBA*. W programie graficznym Adobe *Photoshop CS3* użyto filtrów graficznych do map kartograficznych. W nim również sporządzone zostały mapy semantyczne poprzez wprowadzenie etykiet do zarysowanych klastrów.

b) Etap kolekcjonowania danych

Stworzenie aplikacji służącej do skanowania danych wymagało dokładnego przestudiowania zawartości dokumentów publicznie dostępnych w bibliotece cyfrowej

ACM. Należy zaznaczyć, iż z kompletnym tekstem publikacji można zapoznać się dopiero po opłaceniu składek członkowskich tego Towarzystwa. Poniżej na Rysunku 28 przedstawiona jest uproszczona postać takiego dokumentu z dostępnymi metadanymi i treścią abstraktu. Za pomocą skanera sieciowego skompletowano informację zawierającą metadane w naturalnej kolejności, która jest prezentowana na Rysunku 38:

- Tytuł,
- Autor,
- Rok publikacji,
- Kody klasyfikacji: podstawowej i dodatkowej,
- Terminy główne,
- Słowa kluczowe,

oraz adres *URL* strony. Jak widać z Rysunku 38, relewantne informacje można uzyskać z wyodrębnionych fragmentów tekstu w dokumencie HTML, a taki właśnie format został zaimplementowany w poindeksowanych dokumentach biblioteki cyfrowej *ACM*. Konieczne jest wtedy użycie odpowiednich funkcji interpretujących łańcuchy tekstowe (np. *STRISTR*, *STRIPOS*, *STRREV*, *EXPLODE*), jak również wyrażeń regularnych[12]. W nielicznych przypadkach, na przykład kompletowania danych o tytule, algorytm dało się uprościć do wyszukiwania odpowiednich znaczników (meta, koloru i innych parametrów czcionki). Proces skanowania pierwotnie nie uwzględniał duplikatów dokumentów, ale w późniejszych etapach zoptymalizowano go poprzez podział zadań. Na początku pobierano linki do filtrowanych według daty publikacji stron. Adresy *URL* skanowanych stron zawierały unikalny identyfikator dokumentu w bazie danych *ACM*, a także informację o klasie startowej w dendrogramie, skąd rozpoczęto wyszukiwanie. Zachowując w adresach jedynie część z identyfikatorem artykułu, można było łatwo pozbyć się powtarzających się rekordów. W kolejnej fazie była odczytywana zawartość stron, unikalnie zidentyfikowanych w bazie danych.

Jako główne jednostki analizy wybrano trzy ostatnie pozycje. Zdecydowana większość dokumentów należała do „ko-klas", czyli krzyżujących klas się na drzewie klasyfikacji. Proces kolekcjonowania danych okazał się być bardzo czasochłonny ze względu na ograniczony dostęp użytkowników do serwera. Różnica czasu – serwer jest ulokowany w Stanach Zjednoczonych – wymagała uruchamiania aplikacji porą nocną. Sposobem na uniknięcie problemu z obróbką ogromnej ilości danych z kolekcji biblioteki cyfrowej było odfiltrowanie publikacji w przedziale rocznym, a mianowicie opublikowanych w 2007 r. Taka próbka dokumentów powinna być wystarczająco reprezentatywna dla literatury informatycznej nagromadzonej od 1999 r., czyli epoki „panowania" *Google*. W następnych fazach eksperymentu zbadano artykuły z lat poprzedzających w cyklach 10-letnich.

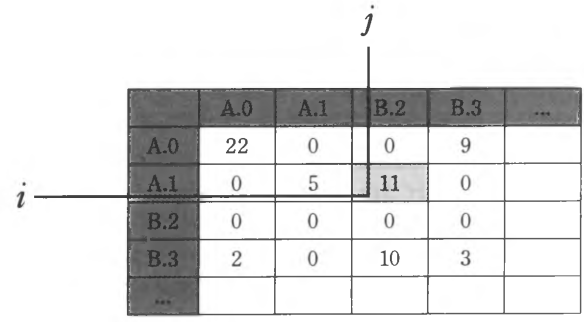
Pierwotna liczba rekordów – artykułów opublikowanych w 2007 r. wynosiła około 64 tys., a po usunięciu duplikatów została zredukowana do 37 543. Pierwszopla-

[12] Wyrażenia regularne (ang. *regular expressions*) – opisują łańcuchy znaków. Za pomocą wyrażeń regularnych można opisywać pewne skomplikowane wzorce wyszukiwania treści dokumentów. Autorka niniejszej rozprawy opracowała autorski program ćwiczeń, poświęcony metodom eksploracji tekstu *Text Mining* przy wykorzystaniu wyrażeń regularnych.

nowym zadaniem niniejszej pracy była optymalna wizualizacja takiej ilości obiektów na powierzchni sfery.

c) Etap przetwarzania danych

Liczba wszystkich możliwych klas i podklas, które pojawiały się w pierwotnej i dodatkowej klasyfikacjach na każdym poziomie determinował ważny parametr – wymiar macierzy podobieństwa. Obliczono, iż ta liczba wynosi 353. Do stworzenia takiego arkusza wymagany był więc program, obsługujący tablice o 16-bitowej adresacji. Skonstruowano zatem macierz $\boldsymbol{S}$, składającą się z 353 wierszy i 353 kolumn, przy czym etykiety wierszy odnosiły się do symboli podstawowej klasyfikacji, a kolumn – klasyfikacji dodatkowych. Element macierzy, leżący na przecięciu wiersza i oraz kolumny j zawiera liczbę dokumentów, występujących jednocześnie w klasyfikacji podstawowej A.1 oraz w klasyfikacji dodatkowej B.2; na Rysunku 29 element $s_{i,j} = 11$



Rysunek 29. Konstruowanie macierzy podobieństwa klas

Źródło: opracowanie własne.

Sumując wprowadzone liczby według wierszy i kolumn otrzymaliśmy całościową ilość dokumentów w bazie danych:

$$N = \sum_{i=1}^{353} \sum_{j=1}^{353} s_{ij} \tag{4}$$

Następnie przeprowadzono normalizację do całkowitej liczby dokumentów dla każdej klasy podstawowej oraz klasy dodatkowej. Współistnienie ko-klas założono w obu kierunkach, dlatego znormalizowane ilości par klas zsumowano i tym samym zrealizowano symetryzację macierzy. Fragment gotowej macierzy jest przedstawiony na Ilustracji 1. W wyniku tych operacji elementy macierzy $s_{i,j}$ dla $i \neq j$ mieściły się w przedziale [0,1], natomiast elementy diagonalne równały się 1. Można tu posłużyć się interpretacją, iż klasa jest tożsama z własnym odpowiednikiem. Im większy jest element macierzy $s_{i,j}$, tym większa jest tematyczna bliskość ko-klas. Jeśli do problemu podobieństwa podejść z innej strony, czyli od budowania macierzy odle-

głości klas d_{ij}, to taka macierz zawierałaby same zera na diagonali oraz wszędzie indziej – liczby większe od 1.

W klasycznych zastosowaniach technika skalowania wielowymiarowego odtwarza czytelną reprezentację obiektów, uporządkowanych w przestrzeni Euklidesowej, tak jak na przykład odległości pomiędzy miastami. W przypadku mierzonej macierzy podobieństwa chodzi natomiast o odległości w przestrzeni semantycznej. Na podstawie wyników oceny bliskości między klasami w przestrzeni 353 – wymiarowej znaleziono ich przestrzenną reprezentację w postaci kartezjańskich współrzędnych *X*, *Y* i *Z*.

	1 A.0	2 A.1	3 A.2	4 A.m	5 B.0	6 B.1	7 B.1.0	8 B.1.1	9 B.1.2	10 B.1.3	11 B.1.4	12 B.1.5	13 B.2	14 B.2.0
A.0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
A.1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
A.2	0,000302	0	1	0	0	0	0	0	0	0	0	0	0	0
A.m	0,00075	0,01882	0,055555	1	0	0	0	0	0	0	0	0	0	0
B.0	0,026585	0,00909	0,013515	0	1	0	0	0	0	0	0	0	0	0
B.1	0,000302	0	0	0	0	1	0	0	0	0	0	0	0	0
B.1.0	0,000302	0	0	0	0	0	1	0	0	0	0	0	0	0
B.1.1	0	0	0	0	0	0,11539	0	1	0	0	0	0	0	0
B.1.2	0	0	0	0	0	0,2	0	0,11539	1	0	0	0	0	0
B.1.3	0	0	0	0	0	0	0	0	0	1	0	0	0	0
B.1.4	0	0	0	0	0	0	0	0	0,2	0	1	0	0	0
B.1.5	0	0	0	0	0	0	0	0	0	0	0	1	0	0
B.2	0	0	0	0	0	0	0	0,03846	0	0	0,125	0	1	0
B.2.0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
B.2.1	0	0	0	0	0	0	0	0	0	0	0,125	0	0,083335	0
B.2.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B.2.4	0	0	0	0	0	0	0	0	0	0	0	0	0,19231	0
B.3	0,00302	0	0	0	0,013515	0	0	0	0	0	0	0	0,25	0
B.3.0	0,000905	0	0	0	0	0	0	0	0	0	0	0	0	0
B.3.1	0,00151	0	0,02	0	0	0	0	0	0	0	0	0	0	0
B.3.2	0,000605	0	0	0	0,013515	0	0	0	0	0	0	0	0,25	0
B.3.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B.3.m	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B.4	0,00272	0	0	0	0,013515	0	0	0,11539	0	0	0	0	0	0
B.4.0	0,002415	0	0	0	0	0	0	0	0	0	0	0	0	0
B.4.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B.4.2	0	0	0	0	0,013515	0	0	0,076925	0	0	0	0	0,03846	0
B.4.3	0,000302	0	0,00532	0	0,00532	0	0	0,076925	0	0	0	0	0	0
B.4.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B.4.m	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B.5.1	0	0	0	0,00075	0	0	0	0	0	0	0	0	0,096155	0,125

Ilustracja 1. Fragment macierzy podobieństwa

Użyto minimalizacji funkcji *Kruskal's Stress*, aby oszacować, na ile dobrze (lub źle) wynikowa konfiguracja odtwarza wejściową macierz odległości (podobieństwa). Obserwowane wartości współczynnika *stress* dla konfiguracji trzy- i dwumiarowej wyniosły odpowiednio 0.25 i 0.3. Im więcej wymiarów, tym lepsze dopasowanie: w przestrzeni 4-wymiarowej wartość *stress* zapewne byłaby mniejsza, lecz wizualizacja w tym przypadku byłaby bardzo skomplikowana.

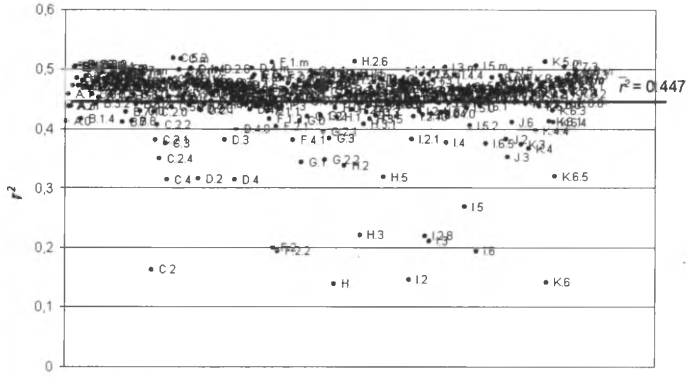
Bardzo charakterystyczne w rozkładzie przestrzennym węzłów klas – nazwijmy je obiektami pomiarowymi – było to, iż zostały one jednakowo „rozrzucone" dookoła początku układu współrzędnych [000]. A zatem, w tym punkcie został umieszczony centroid (środek ciężkości) otrzymanej konfiguracji, który jest średnią współrzędnych obiektów:

$$C_{ijk} = [\Sigma x_i \; \Sigma y_j \; \Sigma z_k]. \tag{5}$$

Niezbędnym krokiem było zestawienie wektorów promienia z obliczeń sumy kwadratów współrzędnych:

$$r^2 = x_i^2 + y_j^2 + z_k^2. \quad (6)$$

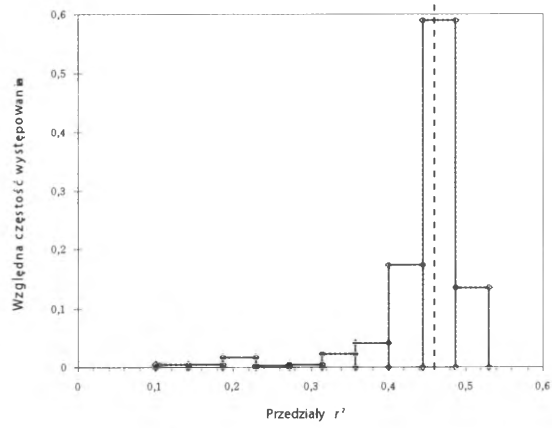
Na Wykresie 2 przedstawiony jest wykres rozrzutu wartości r^2. Dla serii danych naniesiono również etykiety odpowiednich klas i podklas. Wyraźnie widoczna jest akumulacja większości węzłów w pewnej odległości od środka. Zrealizowano zatem rozłożenie obiektów na powierzchni sfery o jednolitym promieniu R.



Wykres 2. Rozkład wartości r^2 dla obiektów zmierzonych za pomocą *MDS*. Czerwona linia wskazuje wynikową wartość kwadratu promienia sfery.

Źródło: Opracowanie własne.

Promień sfery znaleziono metodą najmniejszych kwadratów: R^2=0.447. Policzono również odchylenie standardowe, które wyniosło ^{2} = 0.06. Innymi słowy rozrzut położenia węzłów względem powierzchni sfery wahał się w granicach 27% wartości promienia. Takie duże odchylenie było nie do przyjęcia w toku dalszych rozważań. Kolejny wykres – histogram rozrzutu tych wartości, zaprezentowany na Wykresie 3 – wskazuje na to, iż



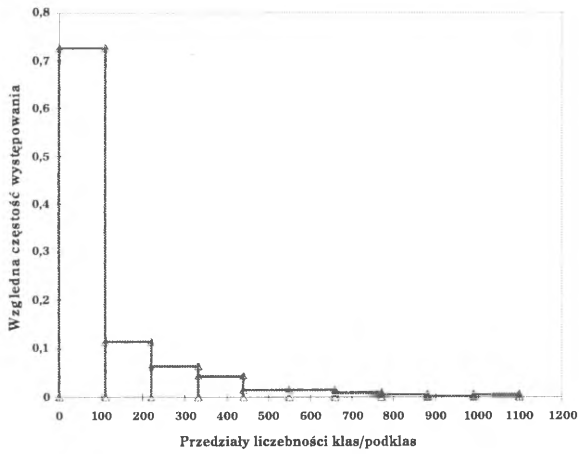
Wykres 3. Histogram rozrzutu wartości r^2. Linia pionowa oznacza wynikową wartość kwadratu promienia sfery.

Źródło: Opracowanie własne.

większa część obiektów zlokalizowana jest bezpośrednio na powierzchni sfery, 36% – z odchyleniem w granicach 0.04 względem R^2 z preferencją lokalizacji wewnątrz kuli. Pojedyncze punkty mające większe wartości odchylania wskazywały klasy, zawierające tylko pojedyncze dokumenty. Ze względu na małą wartość informacyjną zdecydowano się te punkty pominąć, jako zaburzające obraz dopasowania *MDS* i końcowa ilość węzłów wyniosła 347.

Aby „przyciągnąć" punktowe obiekty do powierzchni sfery wprowadzono potencjał *Morsa* (p. wzór 1 oraz Wykres 1). Dla „cząsteczek" o minimalnej energii, czyli w stanie równowagi głębokość studni potencjału powinna równać się promieniu badanej sfery *R*. Parametr *b* dopasowano empirycznie. W wyniku 3-krotnej korelacji danych badawczych z funkcją *Morsa* wszystkie obiekty klas usytuowały się na powierzchni symetrycznej sfery. Dodatkowym efektem działania potencjału było to, iż w przypadkach bardzo bliskiej lokalizacji, „cząsteczki zaczynały się odpychać", korygując swoje współrzędne.

Takim sposobem otrzymano reprezentację 347-iu węzłów klas i podklas na sferze o promieniu *R*. Jak łatwo się domyślić, liczebność dokumentów w klasach i podklasach



Wykres 4. Histogram populacji klas/podklas

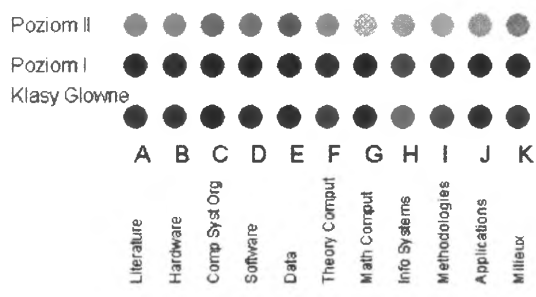
Źródło: Opracowanie własne.

była bardzo zróżnicowano. Histogram na Wykresie 4 przedstawia rozrzut populacji klas. Jak można wywnioskować, zdecydowanie więcej jest klas (ponad 70%) z liczba dokumentów, nie przekraczającej 100. Tylko w trzech klasach w bazie danych zawartych jest ponad tysiąc dokumentów. Takie szacunkowe obliczenia, dotyczące populacji klas przydatne będą w ich dalszej wizualizacji.

d) Etap wizualizacji danych

Do renderowania otrzymanych wyników użyto środowiska programistycznego Matlab, który w wielu pracach[13] wykorzystywany jest do wizualizacji oraz analizy danych w przestrzeni 2D i 3D. Język skryptowy *Matlab* jest językiem wysokiego poziomu. W dużym stopniu jest intuicyjny dla użytkownika, dlatego pozwolił na realizację wizualizacji w czasie znacznie krótszym, niż zajęłoby napisanie pierwotnego kodu w *C* lub *Javie*. Jednak pewne ograniczenia gotowych funkcji renderujących, na przykład stopień mieszania kolorów, sygnalizowały o potrzebie napisania własnej aplikacji, co zaplanowano w przyszłych badaniach.

Jak wspomniano w rozdziale o metodyce badań (Rozdział 3.1), do wizualizacji danych na powierzchni sfery użyto trzech atrybutów: barwy, jasności oraz rozmiaru obiektu graficznego reprezentującego węzeł klasy. Barwa numerycznie określana za pomocą wartości trzech składowych koloru RGB, identyfikowała każdą z 11-tu klas głównych. Wybrano tu paletę 12-tu kolorów, percepcyjnie przyjaznych dla użytkownika[14]. Dla danej klasy trzeba było stopniować jasność koloru na dwóch dodatkowych poziomach. A zatem, pojawiło się 11 x 3 = 33 możliwości kodowania klas i poziomów za pomocą koloru (Rysunek 30): klasy główne przyjmują kolor podstawowy, poziomy I, II – kolory odpowiednio ciemniejsze i jaśniejsze.



Rysunek 30. Paleta kolorów kodowania graficznego klas i poziomów w zastosowanym modelu

Źródło: Opracowanie własne.

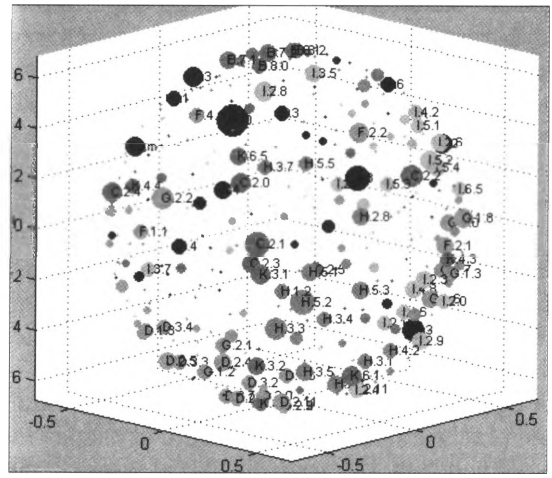
Jako trzeci atrybut zastosowano rozmiar glyfu (p. Rozdział 1.1), najczęściej będący okręgiem, bądź czaszą kuli[15]. Histogram populacji klas (Wykres 4) dowodzi mocno zróżnicowanego (do trzeciego rzędu) rozkładu dokumentów. Dlatego rozmiar glyfu wyznaczany był nie za pomocą liczby dokumentów w klasie, lecz logarytmu tej wartości. Otrzymano zatem rozkład klasyfikacji na powierzchni sfery, zademonstrowany na Rysunku 31. Aplikacja pozwala na podgląd etykiet danych, czyli kodów klas. Ze względu na dużą gęstość obiektów wizualizacji, taka istotna dla

[13] T. Holloway, dz. cyt., s.30-40; N. Shoichiro. *Numerical Analysis and Graphic Visualization with MATLAB.* Upper Saddle River, USA: Prentice Hall, 2002, s. 45-78; Soukup Tom & Davidson Ian. *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. New York, USA:John Wiley & Sons, 2002, s. 203-315.

[14] C. Ware, dz. cyt., s.126.

[15] Czasza – wycinek kuli wyznaczony poprzez płaszczyznę ją przecinającą.

analizy informacja staje się czytelna dopiero przy maksymalizacji okna programu, dlatego zamieszczono powiększoną wersję na Ilustracji 2.



Rysunek 31. Zrzut ekranowy aplikacji do wizualizacji klasyfikacji *CCS* na powierzchni sfery

Źródło: Opracowanie własne.

Kolejnym wyzwaniem było rozmieszczenie dokumentów na tejże sferze. Metoda wykorzystywała otrzymane wyniki wizualizacji klas i opisana została w podrozdziale wyżej. Istotnym zagadnieniem było ustalenie dla dokumentów wag klasyfikacji podstawowej i dodatkowych. Od tego zależało położenie węzła dokumentu wewnątrz figury (odcinek, trójkąt bądź wielokąt), wierzchołki której tworzyły węzły klas (Rysunek 25). Im większa waga klasyfikacji dodatkowych, tym dalej od klasyfikacji podstawowej zlokalizowany będzie dokument. W celu wstępnej oceny rozkładu wybrano trzy wartości relacji wagowych: 0.7:0.3, 0.6:0.4 oraz 0.5:0.5.

Posłużono się modelem figury płaskiej, aby obliczyć współrzędne węzłów dokumentów. Położenie punktów na sferze wyliczono poprzez konwersję z układu współrzędnych kartezjańskich do sferycznych i odwrotnie. Węzły dokumentów odziedziczyły kolor nadrzędnej klasy głównej. Reprezentacja graficzna dokumentów charakteryzowała się dwoma atrybutami: kolorem oraz położeniem na powierzchni sfery. Nie było konieczności różnicować glyfów do znakowania dokumentów osobnych klas. Rozproszone kolorowe punkty dokumentów wykazywały tendencję do gromadzenia się w kolorowe klastry. Taka organizacja pozwalała na weryfikację wyników wizualizacji poprzez przypisanie klastrom kategorii tematycznych klas, z których pochodziły dokumenty.

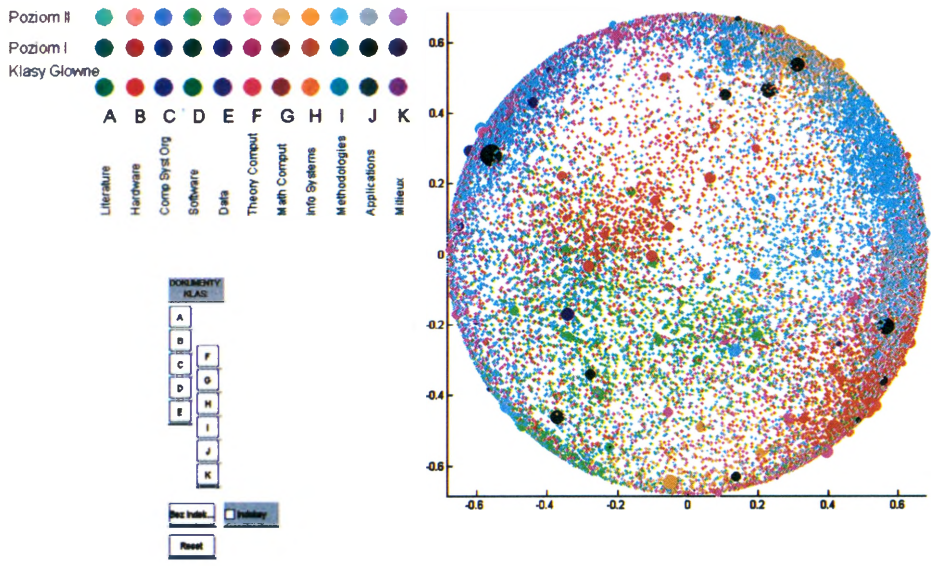
Zanotowano, iż przy relacji wag klasyfikacji podstawowej i dodatkowej 0.7:0.3 węzły dokumentów skupiały się dookoła klasy podstawowej. Przez to tworzyły się „dziury" na powierzchni sfery, a rozkład niewiele się różnił od obrazu wizualizacji klas i podklas. A więc zaniechano dalszego zwiększenia wagi klasy podstawowej i kierunek zmian odwrócono. Z drugiej strony zauważono, że dla połowicznych wag

jednokolorowe klastry, zawierające dokumenty poszczególnych klas znikają, konfiguracja punktów pomiarowych nabiera cech równomiernego rozkładu. Dlatego dalsze zmniejszenie wagi poniżej 0.5 dla klasyfikacji podstawowej nie miało sensu. Zadawalające wyniki rozkładu otrzymano dla relacji wag 0.6:0.4 i wszystkie dalsze obliczenia oraz analizę wyników wykonano dla tych proporcji wag podstawowej i dodatkowych klasyfikacji. Opisana procedurę optymalizacji wag można skonfrontować z otrzymanymi mapami wizualizacji, które zostały załączone na kilku ilustracjach. Na Ilustracji 2 przedstawiona jest wynikowa sfera wizualizacji. Dokumenty – kolorowe punkty wypełniają całą powierzchnię. Większe obiekty wystające spod tej sieci dokumentów – są to klasy i podklasy.

Interaktywny charakter środowiska *Matlab* pozwala na swobodny obrót „kulą wizualizacji" – ułatwia to badanie danych i ich zależności w sposób ciągły.

Natomiast trudno taką kulę zaprezentować na kartce papieru, chociażby w niniejszej pracy. Dlatego następnie zastosowano rzut kartograficzny danych na powierzchni sfery. Kolejnymi ważnymi argumentem „powrotu" do planimetrii są zastosowane metoda oceny rozkładu za pomocą wymiaru fraktalnego (podrozdział niżej) oraz graficzne przetwarzanie obrazów w celu uzyskania map.

Dysponując współrzędnymi sferycznymi dla wszystkich punktów danych, za pomocą wzorów (2) i (3) łatwo jest wyznaczyć współrzędne x i y na płaszczyźnie. Dla wygody przyjęto, iż $\lambda_0 = 0$, oraz $\alpha = \beta = 1$. Ilustracja 3 jest rzutem powierzchni sfery na płaszczyznę. Podstawowa mapa wizualizacji dokumentów, do której będą porównywane kolejne jest przedstawiona na Ilustracji 4. Dla wygody na Ilustracji 5 pokazany jest rozkład dokumentów dla wybranych klas: **A, B, C, D.**



Ilustracja 2. Zrzut ekranowy aplikacji z wynikami wizualizacji

W tym podrozdziale zamieszczono dokładny opis procesu analizy danych, którą należało stosować jako podstawową metodykę przetwarzania danych pierwotnych. W dalszych etapach eksperymentu w zależności od właściwości danych oraz stawianych zadań zastosowano coraz to nowe metody i techniki obróbki.

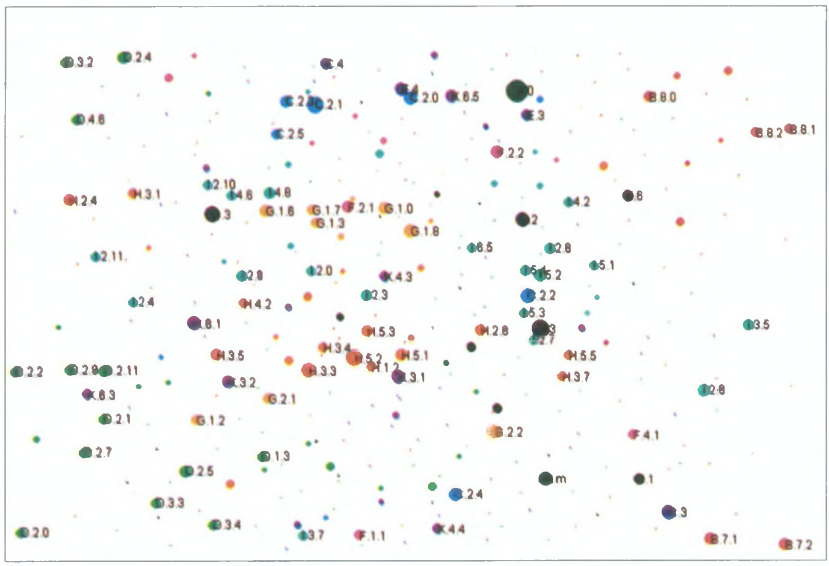
3.3. Interpretacja wyników

a) Powierzchnia sfery

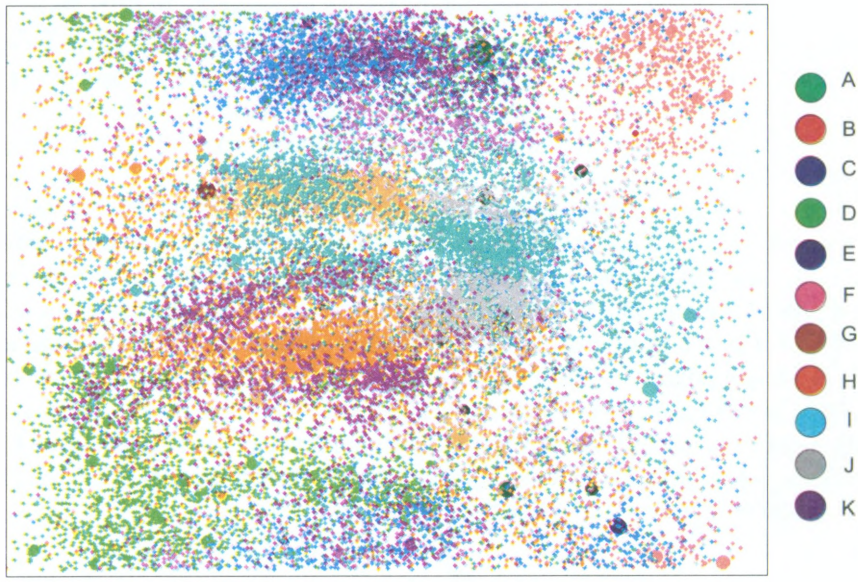
Jak odnotowano wyżej, do wizualizacji 347. klas (6 obiektów zostało odrzuconych w procesie analizy ze względu na znikomą wartość informacyjną – p. Rozdział 3.2.c) na powierzchni sfery użyto trzech atrybutów: barwy, jasności koloru oraz rozmiaru obiektu graficznego reprezentującego węzeł. Barwy, wybrane zgodnie z zasadami percepcji, reprezentowały klasy główne, których klasyfikacja *CCS* liczyła 11. Dla klas niższych poziomów zarezerwowane były te same barwy, lecz z różnym stopniem jasności: poziomu I – kolory ciemniejszy; poziomu II – jaśniejszy (Rysunek 31).

Do renderowania 347. obiektów przetestowano kilka funkcji i ostatecznie wybrano ze względu na mały czas przetwarzania danych najprostszą *PLOT3*. Pozwala ona zdefiniować kolor, rodzaj i rozmiar glyfu. Ten ostatni został obliczany z logarytmu liczby dokumentów w danej klasie/podklasie. Wyniki obciążenia procesora wskazały, że obiektem graficznym może być wypełnione kolorem koło, a nieścisłości spowodowane płaskością tych figur można zaniedbać ze względu na ich skalę w odniesieniu do rozmiaru sfery.

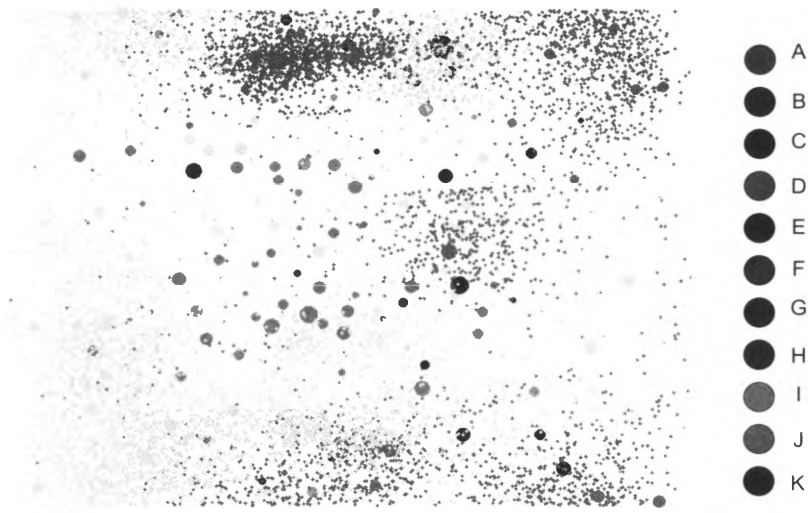
Ilustracja 2 przedstawia otrzymaną wizualizację rozkładu węzłów wszystkich klas klasyfikacji *CCS* na powierzchni sfery. Sfera jest przezroczysta, a zatem widoczne są wszystkie klasy rozrzucone w przestrzeni w danym rzucie obserwacji. Już na pierwszy rzut oka widać, że obraz składa się z mniejszej ilości kolorów, niż 33 (11 kolorów × 3 poziomy). Nie wykorzystanie pełnej palety kolorów tłumaczy się tym, iż klasy/podklasy o śladowej ilości dokumentów są zobrazowane za pomocą punktów praktycznie nie widocznych na rysunku. Na obrazie można zauważyć przewagę kolorów jasnych. A zatem, większość dokumentów była zawarta w podklasach najniższego poziomu. Dodatkowe potwierdzenie dostarczyła opcja podglądu kodów klas, zaimplementowana w aplikacji.



Ilustracja 3. Mapa wizualizacji wszystkich klas (wyświetlone symbole dla klas o dużej ilości dokumentów)



Ilustracja 4. Podstawowa mapa wizualizacji dokumentów (wraz z klasami)



Ilustracja 5. Mapa wizualizacji dokumentów klas A, B, C, D

Obracając sferę można stwierdzić, udało się zachować ciągłość rozmieszczenia klas i podklas. Większość węzłów odnoszących się do jednej klasy usytuowana była blisko siebie na powierzchni, tworząc lokalne ogniska. Takich ośrodków z pewnością było nie jedenaście, lecz więcej. To dowodzi tego, iż tematyczne rozszczepienie klas na podklasy (lub zupełnie nowe, nieznane kategorie) nie było jednolite. Tak na przykład, klasa A. *General Literature* składała się z dwóch dużych węzłów A.0 oraz A.m. Jasnozielone węzły, identyfikujące pochodne klasy D. *Software* skupiały się wokół siebie. Podobnym rozkładem scharakteryzować można klasę B. *Hardware*, liczącą mniej węzłów. „Łańcuszek" kilku ognisk węzłów klasy H. *Information Systems* w jaskrawym pomarańczowym kolorze zdradzał dyskretną naturę zajmującej przestrzeni.

Puste miejsca na sferze wypełniono węzłami dokumentów stosując się do sprawdzonej w rozdziale 3.2 reguły relacji wag klasyfikacji podstawowej i dodatkowych 0.6:0.4. Wynikowa reprezentacja graficzna dokumentów na powierzchni sfery została pokazana na Ilustracji 2. Węzły dokumentów dla łatwiejszej identyfikacji oznakowano w kolorze nadrzędnej klasy głównej.

Przypomnijmy przyjęte w testowanej metodyce założenie, iż powierzchnia sfery kryje więcej możliwości manipulowania wzajemnym rozmieszczeniem obiektów. Na początku przeanalizowano wzory powstałe na skutek mapowania wszystkich dokumentów. Wzory, utworzone z 37 543 kolorowych punktów wykazywały tendencję do skupiania się w kolorowe klastry przy zachowaniu ciągłego wypełnienia całej powierzchni sfery. Uzyskano zatem pocieniowaną za pomocą 11-tu kolorów sferę. Większe kolorowe plamy, wystające spod siatki dokumentów – są klasami i podklasami (Ilustracja 2).

Brak wyraźnie wykształconych pustych miejsc (czyli znikomy stopień lakunarności – p. rozdział 3.4.b) dowodzi słuszności opisywanej metody wizualizacji. Dokład-

niejsza analiza obrazu wykazała, iż kolorowe klastry są nieregularnych kształtów i nie mają wyraźnych granic. Zamiast nich zaobserwowano mieszanie się kolorów, czyli „tematyczne przenikanie" jednej kategorii w drugą. Większość klas z wyjątkiem nielicznych, na przykład klasy E. *Data* są reprezentowane za pomocą więcej niż jeden klaster. Aby dokładnej zbadać wzory klastrów, posłużono się obrazami wizualizacji dokumentów poszczególnych klas z osobna. Według klasycznego dendrogramu *CCS* maksymalnie może być dziesięć (od 0 do 8 oraz m) podklas klasy głównej. W przypadku wizualizacji na sferze struktura klastrowa jest płytsza, lecz aby móc policzyć klastry każdej klasy należy zastosować techniki, wykrywające ukryte krawędzie.

Zauważono, iż dokumenty klas B. *Hardware* oraz D. *Software*, inaczej plamy w kolorach łososiowym i jasnozielonym, zlokalizowane są w maksymalnej odległości od siebie – na przeciwległych biegunach. Łatwo tu o interpretację nawiązującą do tematycznej odrębności pomiędzy tymi klasami, biorąc pod uwagę, iż zagadnienia sprzętu i oprogramowania wykazują obecnie mocno odmienną tematykę. Historycznie rozwijały się one równolegle, lecz zawsze w osobnych kategoriach.

b) Mapy

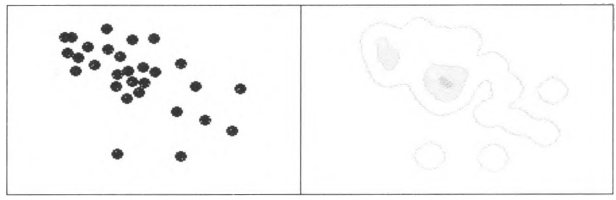
W celu zbadania obszarów krzyżowania się kategorii oraz struktury klastrów poszczególnych klas, wykonano kartograficzny rzut powierzchni sfery. Powstałe mapy pozwoliły na przypisanie klastrom logicznych kategorii tematycznych po poprzedzającej obróbce graficznej. Analiza map na płaszczyźnie jest o tyle wygodna, że nie angażuję ośrodków neuronalnych odpowiedzialnych za precyzyjne ruchy ręką. Wysiłek percepcyjny skupia się na zarazem na kompleksowym wzorze wszystkich kolorowych klastrów, tak i intuicyjnym „przetwarzaniu" ich struktur lokalnych. Ilustracja 4 przedstawia mapę dokumentów wszystkich klas. Przez sieć punktów prześwitują większe obiekty, reprezentujące klasy i podklasy. Dobrze widać, że w pewnych miejscach występuje duże zagęszczenie punktów, przez co obraz może wydawać się przekłamany. Jeśli powiększymy fragment mapy o największej gęstości danych, to wrażenie natłoczenia i tym samym zasłonięcia głębszych warstw kolorów jest błędne. Manipulowanie skalą oraz rozmiarem glyfów umożliwia wykrywanie szczegółów wybranych obszarów. Do uzyskania takiego efektu w przypadku sfery potrzebne są większe zasoby komputera.

Interpretując wzory, należy liczyć się z tym, iż krawędzie mapy są umowne, ponieważ płaszczyzna się zawija. Tak jest na przykład z klastrami dokumentów klasy B. *Hardware*. Skupisko węzłów (kolor zielony) nie urywa się nagle w dole mapy, lecz ma kontynuowane obszary u góry, tam gdzie szerokość geograficzna przyjmuje wartości dodatnie.

Godna uwagi jest także klasa I. *Computing Methodologies*, licząca w strukturze pięć wyraźnych klastrów – kolor turkusowy. Na podstawie stopnia rozproszenia danego zespołu klastrów można zaaprobować wybór wagi klasyfikacji podstawowej i dodatkowych jako relację 0.6:0.4 w procesie obliczania współrzędnych węzłów dokumentów. Mapa na Ilustracji 7 przedstawia wizualizację dla wag 0.5:0.5. Widać na niej, jak zanika struktura kompleksowa klastrów dla klasy I. Z kolei na Ilustracji 8, mapującej dane dla relacji 0.7:0.3 brakuje już informacji o krzyżujących się kategoriach. Chociaż rozłączność jest

pożądaną cechą w klasyfikacji, nadmierne gromadzenie się węzłów dokumentów wokół węzłów klas podstawowych może doprowadzić do mocno niejednorodnego rozkładu z licznymi „dziurami". W zadaniach wizualizacji powinno się unikać takich sytuacji. Przede wszystkim oznacza to, że metoda wizualizacji jest nieefektywna. Mankament tkwi także w tym, iż niejednorodny obraz nie jest czytelny dla użytkownika. Taka wizualizacja klasyfikacji ma brak „perspektyw rozwoju" dla przyszłych kategorii tematycznych albo „narzuca" im nieautentyczną lokalizację.

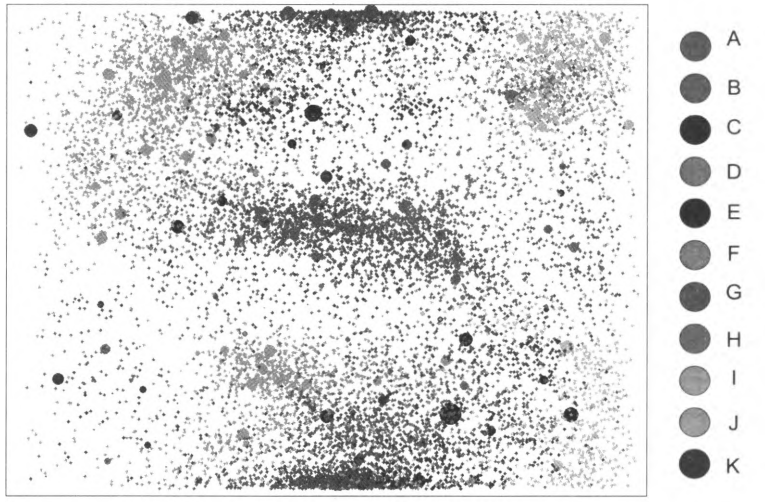
Pokrywanie się warstw kolorów w obrazie wynikowym zmusza nas do zastosowania technik mieszania kolorów. Skupiska kolorowych punktów należałoby rozmyć, aby móc zarejestrować obszary powstałe oraz nałożenie ich barw. W rozdziale 3.1 opisana została metoda zastosowania filtrów graficznych: mediany oraz trasowania konturu. Rysunek 32 ilustruje efekt działania tych filtrów na wygenerowanej próbce danych.



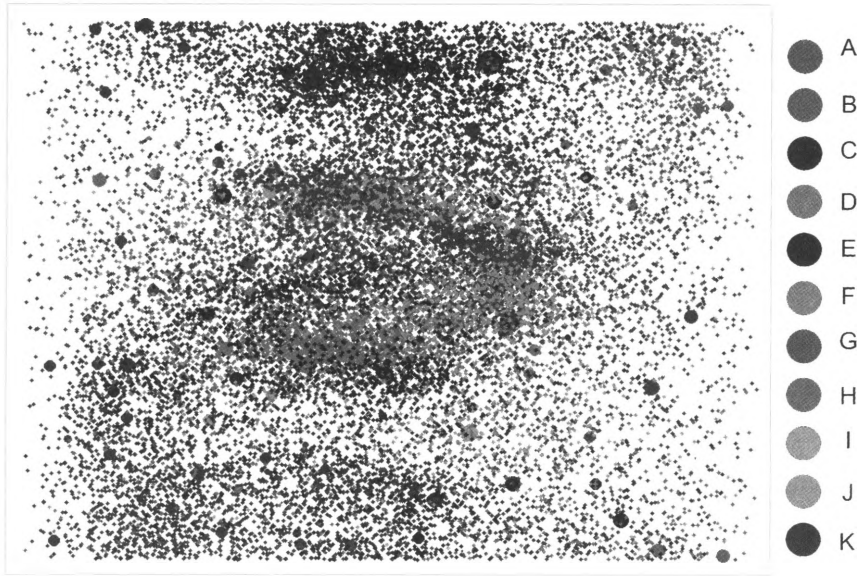
Rysunek 32. Efekt działania wybranych filtrów na próbce danych. Z lewej strony punkty reprezentujące poszczególne artykuły, z prawej – otrzymana na tej podstawie mapa

Źródło: opracowanie własne.

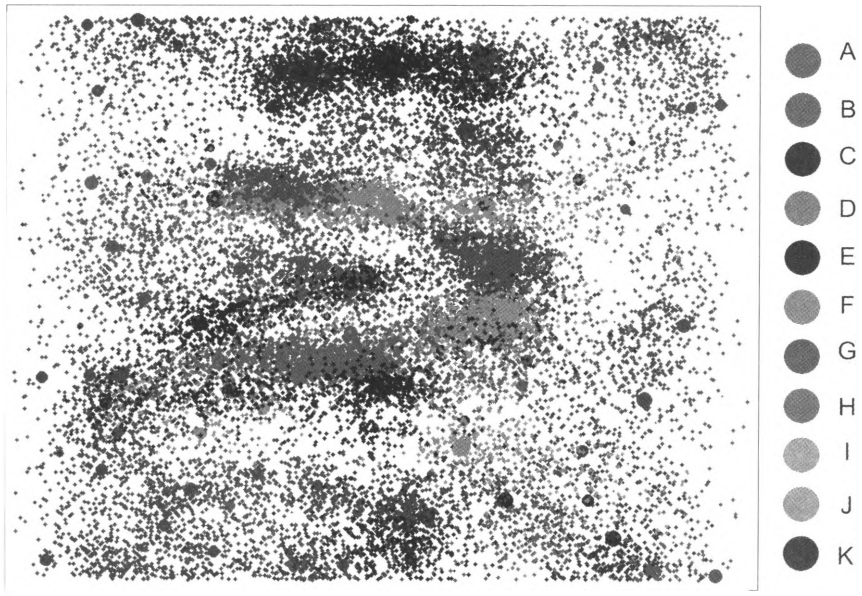
Podobnie jak na mapach topologicznych, intensywność koloru jest stopniowana w zależności od charakterystyki reliefu okolicy. Takimi wyżynami są obszary z największym zgęszczeniem punktów. Im mniejsza gęstość tym lokalnie kolor jest bledszy. Ta metoda kolorowania obrazu włącznie z techniką mieszania kolorów została zaimplementowana w wizualizacji dokumentów kompletnej klasyfikacji.



Ilustracja 6. Mapa wizualizacji zmodyfikowanego zestawu danych (bez klasy I)



Ilustracja 7. Mapa wizualizacji dla relacji klasyfikacji głównej do klasyfikacji dodatkowej 0.5:0.5



Ilustracja 8. Mapa wizualizacji dla relacji klasyfikacji głównej do klasyfikacji dodatkowej 0.7:0.3

Ilustracje 9 i 11 prezentują obrazy wizualizacji po zastosowaniu obróbki graficznej dla wszystkich oraz charakterystycznych kombinacji klas. W wyznaczeniu składowych klastrów na takiej mapie sugerowano się ich lokalną ciągłością albo bliskim położeniem rozproszonych skupisk punktów. Porównując te obrazy z ich wersją podstawowej wizualizacji (Ilustracja 4) widać, że dla niektórych klas zbyt

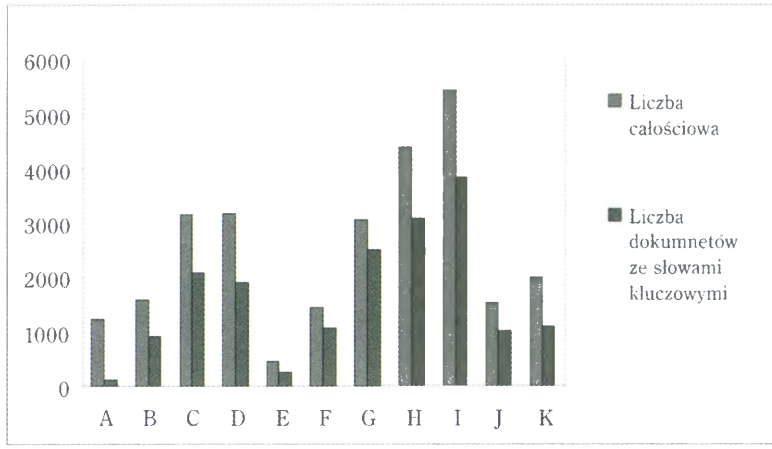
duże rozproszenie punktów przeszkadza w syntezie struktury klastrów. Tak dla dokumentów klasy I. *Computing Methodologies* zostało wyodrębniono aż 6 głównych klastrów i je obrysowano dla ścisłości. Klasa G. *Mathematical Computing* składa się z wyraźnych trzech ognisk z mnimalnym stopniem rozproszenia. Dokumenty klasy głównej E. *Data* organizują tylko jeden klaster; tu nawet nie ma potrzeby użycia filtrów. Świadomość ciągłości mapy okazała się przydatna w identyfikacji górnej i dolnej części klastrów na przykład w przypadku klasy D. *Software*. Klasa A. *General Literature* dzieli się na dwa obszary skupione wokół podklas podstawowych o szerokim zakresie tematycznym: A.0 i A.m, które są bardzo wyraziste na mapie klas. Dla klasy H. *Information Systems* naliczono 4 klastry oraz zanotowano ich dużą niejednorodność: trzy główne – bardzo liczne oraz trzy ledwo zauważalne, zawierające po kilkadziesiąt dokumentów.

Słowa kluczowe

W takich identyfikacji opisanych powyżej skupisk artykułów o różnej wartości informacyjnej był pomocny kolejny etap przetwarzania zmapowanych danych. Tym razem użyto słowa kluczowe. Posługując się numerem rekordu w bazie danych, zgromadzono słowa kluczowe dokumentów, wchodzących w zaznaczone na mapach obszary. Na podstawie rankingu wyselekcjonowano słowa kluczowe, charakterystyczne dla danego klastra. Ilustracja 10 zawiera sekwencje słów z pierwszej dziesiątki listy rankingowej, czyli najczęściej występujących w badanej grupie dokumentów. Na wyniki wpływał fakt, iż nie każdy autor publikacji z kolekcji *ACM* określił słowa kluczowe, a więc spory procent danych – sięgający czasem ponad połowę - trzeba było odrzucić w pierwszej fazie obróbki statystycznej, np. przy normalizacji. Ta grupa „jałowych" danych wpływała na rozkład przestrzenny obiektów, lecz była nieużyteczna na etapie weryfikacji za pomocą słów kluczowych. Wykres 5 dobrze obrazuje wkład wykorzystanych danych klas.

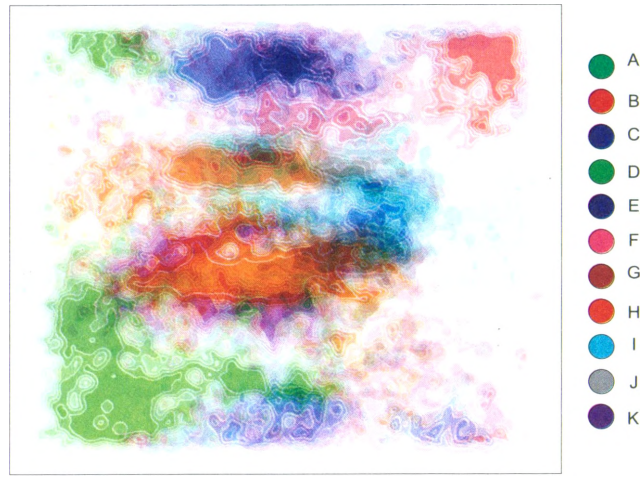
Zauważono poza tym, że nie wszystkie elementy klastrowe jednolicie prezentują wartość analityczną. Nie jest to uzależnione od liczebności klastrów, chociaż istnieje pewien próg liczby dokumentów (od kilkanastu do kilkudziesięciu), poniżej którego trudno jest uzyskanie sensownych wyników. Wartość progową można określić jedynie doświadczalnie. Jeśli skupiska danych nie niosą wartości informacyjnej, to można potraktować je jako nieistotne.

Za drugiej strony, nie zaobserwowano ograniczenia ilości dokumentów od góry. Największy przeanalizowany klaster składał się z 2747 dokumentów – należał on do klasy H. *Information Systems*.



Wykres 5. Statystyka ilości dokumentów zawierających słowa kluczowe. Na wykresie zaznaczono liczbę dokumentów w poszczególnych grupach klasyfikacyjnych
Źródło: opracowanie własne.

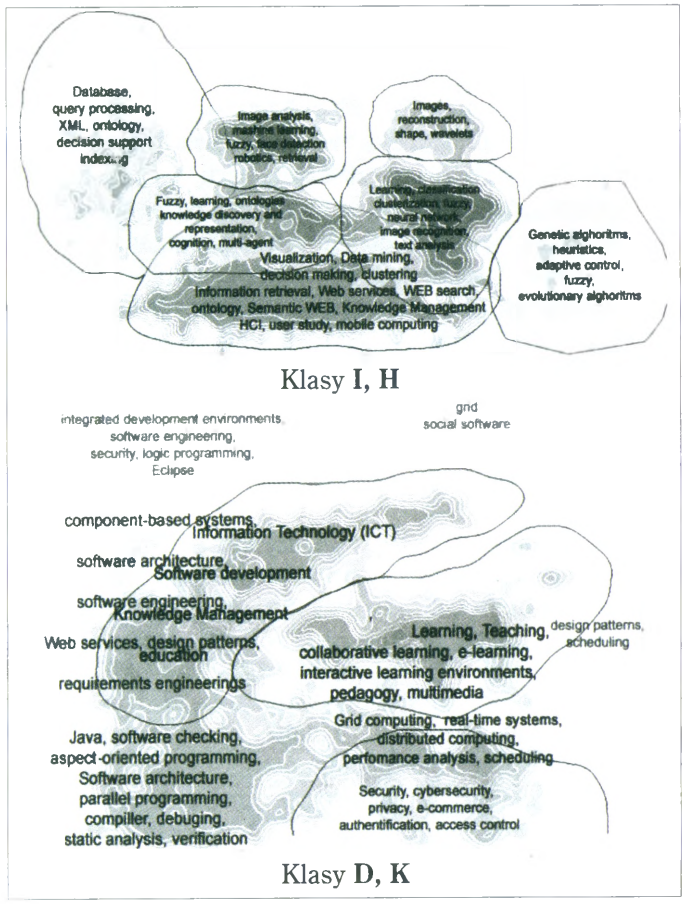
Wyzwaniem było zestawienie słów kluczowych klastrów na wspólnej mapie, którą przedstawia Ilustracja 10. Kolor klasy głównej wskazuje kolor czcionki tekstu. W większej skali taka mapa słów kluczowych jest czytelna i zawiera dużo informacji semantycznej. Sąsiedztwo większości wyrazów jest logicznie uzasadnione. Można zauważyć, że na górnej i dolnej krawędzi (na powierzchni sfery są to pola przylegające), skupia się problematyka związana z bezpieczeństwem funkcjonowania sieci: security, *cybersecurity, privacy, authentification, cryptography*. Obok po lewej stronie zlokalizowany jest obszar, określający tematykę sieci *LAN*[16], sieci bezprzewodowych, sieci Internet, a także temat pochodny – *Grid Computing*, dotyczący zarządzania zasobami rozproszonymi.



Ilustracja 9. Mapa wizualizacji po obróbce graficznej

[16] *Local Area Network* – sieć wewnętrzna, postać sieci komputerowej, ograniczona do jednego pietra, budynku lub kilku sąsiadujących.

Ilustracja 10. Mapa słów kluczowych



Ilustracja 11. Mapy słów kluczowych kombinacji różnych klas

W środowisku sieci Grid kluczowym wymaganiem jest niewątpliwie bezpieczeństwo stąd związek z sąsiednimi obszarami. Z prawej strony rozmieszczone zostały hasła, odwołujące się do sprzętu i zagadnień niskiego poziomu. Grupy wyrazów w kolorze turkusowym, wyznaczające granice klastrów klasy I. Computing Methodologies – rozciągnęły się w środkowej części mapy (odpowiada równikowi na sferze). Dwie grupy ukierunkowane są wyraźnie na uczenie maszynowe, sztuczne sieci neuronowe, robotykę, analizę tekstu, przetwarzanie obrazów, a następna – na zarządzanie wiedzą i ontologie, jeszcze inna wskazuje na prace w zakresie algorytmów genetycznych. Te ostatnie bazują na naturalnym zachowaniu ekspresji genów, dlatego blisko mieszczą się tematy związane z medycyną. To, że tu ponownie pojawia się Grid Computing dowodzi coraz częstszych projektów gridowych w zastosowaniach bimolekularnych i medycznych. Tematycznie uzależnione z nauką o informacji są systemy informacyjne, identyfikowalne z klasą H. Klastry tej klasy wyspecjalizowane są kolejno w eksploracji i wizualizacji danych, semantyce, zarządzaniu wiedzą oraz społecznych aspektach Internetu (Social Computing, Wiki2.0). Klasa K. Milleux, w założeniu autorów klasyfikacji CCS nawiązująca do drugorzędnej problematyki środowiska kompu-terowego zajmuje bardzo rozległą przestrzeń. Słowa kluczowe w kolorze fioletowym determinują tematy związane z technologią informacyjną oraz edukacji na odległość. Zupełnie zrozumiałe, że w bliskim sąsiedztwie znajdują się wyrazy: User study – po jednej stronie i Software development – po drugiej.

Powyższa interpretacja jest logiczna, zawiera ona pierwiastek poznawczy. Dzięki takiej mapie słów kluczowych można rozważać powiązania terminologiczne oraz poznawać ukryty sens pojęć w odniesieniu do najbliższych słów – sąsiadów. Może ona również być wykorzystana do weryfikacji wcześniejszej mapy – kategorii tematy-cznych. Tu należy nadmienić, że przypisywanie artykułów do klasyfikacji podstawowej (i niekiedy dodatkowych) należało do autorów i było kontrolowane przez edytorów serwisu biblioteki cyfrowej *ACM*. Doborem odpowiednich klas dodatkowych zajmowali się głównie redaktorzy serwisu. Słowa kluczowe natomiast określali wyłącznie autorzy publikacji. W wyniku mapowania obszarów słów kluczowych uzyskano transformację wizualizacji tematycznej na mapę semantyczną. Na tych dwóch mapach skonfrontowane zostały dwie koncepcyjno-skojarzeniowe ścieżki, pochodzące z niezależnych źródeł.

c) Nowa klasyfikacja

Na mapie semantycznej poklasteryzowane zostały za pomocą słów kluczowych dokumenty pochodzące z biblioteki cyfrowej *ACM*. Nie należy zapominać o pierwotnej (wyjściowej) klasyfikacji zbadanej grupy dokumentów. Schemat *CCS* zawierał trzy stopnie hierarchii plus nienumerowany deskryptor tematyczny, który okazał się nieużyteczny w procesie analizy danych. Mapując klasy i dokumenty na powierzchnię, informacja o poziomach hierarchii uległa zatarciu. Półautomatyczna klasteryzacja wynikowa sąsiadujących węzłów artykułów o podobnych słowach kluczowych charakteryzuje się tylko jednym poziomem zagnieżdżenia. Równorzędne w nowej strukturze klastry różnią się: liczebnoś-

cią dokumentów, wielkością zajmowanego pola oraz gęstością punktów danych. Najbardziej spójne wyniki selekcji słów kluczowych otrzymano dla klastrów o dużym zagęszczeniu. Jak wspomniano w poprzednim rozdziale, cześć danych o charakterze szumu należało odrzucić w tym klastry o zbyt małej ilości dokumentów.

Logiczny tok wnioskowania świadczy o potrzebie zestawienia dwóch struktur klasyfikacji: pierwotnej i doświadczalnej. Tym samym należało zbadać, na ile kategorie tematyczno-semantyczne pokrywają się w obu drzewach. Można taką metodę potraktować jako ewaluację organizacji kategorii tematycznych wejściowej klasyfikacji. W jakim stopniu odwzorowuje ona rzeczywisty podział współczesnej literatury informatycznej? Czy odpowiada aktualnemu stanowi rozwoju nauk komputerowych? Czy trzypoziomowa hierarchia jest wystarczająca i właściwa?

Tabela A zawiera istotne informacje o strukturze klasteryzacji danych. Oprócz listy słów kluczowych, podane są ilości dokumentów relewantnych (ze słowami kluczowymi) oraz całkowita liczba dokumentów w poszczególnych klastrach. Na podstawie tych danych przeprowadzono numerację klastrów. Trzecia i czwarta kolumny tabeli są wiążące dla pierwotnego schematu, ponieważ wymieniają zawarte w klastrach kategorie tematyczne i deskryptory przedmiotowe (które teraz okazały się być przydatne) na podstawie klasyfikacji głównej. Tabela porównawcza dostarcza wiele informacji o podklasach, deskryptorach i słowach kluczowych w klastrach. Na końcu tabeli załączone zostały schematy wygładzonych klastrów dla każdej klasy głównej. Dzieki temu można znaleźć reguły organizacji dokumentów w każdym klastrze osobno, jak również ich cechy wspólne. Mimo że tabela jest klasycznym, nieco archaicznym środkiem wizualizacji danych, jednak dobrze jest ją wykorzystać do porównania zestawów słów kluczowych poszczególnych klas i klastrów. Jedne listy są nieporównywalnie długie wobec wykazu „tożsamych podklas" klasyfikacji *CCS*, inne odwrotnie zadziwiająco krótkie nawet dla dużej liczby dokumentów. Można też pozliczać hasła i tym samym określić częstość występowania słów kluczowych w klastrach. Dużą pokusą jest przeciwstawienie obu zestawów, lecz nie należy mylić konceptu klasa/podklasa ze słowem kluczowym, wyrwanym z „otoczenia" innych. Ważne jest wyłowienie częstości powtarzania się deskryptorów przedmiotowych w klastrach. Duża częstość oznacza, że kategoria tematyczna nie jest precyzyjna, wskazana jest jej modernizacja. Zanotowano, że w 70% powielające się podklasy w obrębie jednej klasy są etykietowane na oryginalnym drzewie *CCS* jako aktualnie korygowane (ang. *Revised*).

Szczegółowo rozpatrzono dwie losowo wybrane klasy C. *Computer Systems Organizations* oraz H. *Information Systems*. Dwie jednostki analizy jak deskryptory przedmiotowe oraz słowa kluczowe figurowały niezależnie. Wykryte identyczne lub semantycznie podobne wyrazy i wyrażenia wyróżnione zostały pogrubioną czcionką. Klaster *1 góra* należący do klasy C. *Computer Systems Organizations* związany jest z problematyką pracy w sieci komputerowej, zwłaszcza sieci bezprzewodowej oraz sieci o zdecentralizowanej strukturze – *ad hoc networks*. Znacząca część danych pochodziła z publikacji o technologii komórkowej (*mo-*

bile technology). Często powtarzająca się liczba 802.xx oznacza grupę standardów stosowanych w lokalnych oraz miejskich sieciach komputerowych sieciach komputerowych przesyłających dane w systemie pakietowym. Słowo kluczowe *broadcast* występuje jako tryb transmisji danych. Tematy bezpieczeństwa sieci (ang. *security networking*) pojawiają się w obu zbiorach. Wśród deskryptorów klasyfikacji *ACM* nie dało się znaleźć współczesnej ważnej terminologii sieciowej jak na przykład: *LAN, routing, ad hoc, Ethernet, broadcast*. Natomiast występuje nieco przestarzała technologia *ISDN*. Klaster *1 dół* pomimo podobieństwa tematycznego do górnej części dodatkowo odnosi się w słowach kluczowych do tematów obliczeń rozproszonych – *distributed computing* oraz ich współczesnej formy – *grid computing*. Z drugiej strony dopasować można zestaw kategorii: *distributed application, distributed database* a również *network management*. Analogicznie słowo kluczowe *quality of service* odpowiada deskryptorom *reliability* i *serviceability*. Klaster *2* natomiast głownie specjalizuje się w protokołach sieciowych (*network protocols* i *routing)*, klaster *3* – w systemach rozproszonych.

Klastry klasy H. *Information Systems* odwzorowywały szerokie spektrum tematów związanych z informacją naukową. Klaster *1* zakreśla szeroka tematykę badawczą wywodzących ze statystyki i uczenia maszynowego: *data mining, information retrieval, clustering*. Sporo przestrzeni tematycznej poświęconej jest oddziaływaniu człowiek – komputer oraz badaniu zachowań użytkownika: *Human - Computer Interface, Human Factors, User Study*. Istotną cechą jest zaobserwowany brak przedmiotów równoważnych do słów: *ontology* i *knowledge management*. Klaster *2* można zinterpretować jako wynik przeglądu tekstów, dotyczących analizy decyzyjnej (*query processing, decision making*). Deskryptory przedmiotowe opisujące języki baz danych (*SQL, DDL*) odpowiadają słowu kluczowemu *database*. Ponieważ klastry *1* i *3* są położone blisko siebie, ten ostatni przyjmuje cechy pierwszego i jego tematykę podobnie można określić. Dodatkowo pojawia się rozszerzalny język znacznikowy XML odpowiadający wyrażeniu w klasyfikacji pierwotnej: *Schema and subschema*. Także tu występująca podklasa *CCS* H.2.5. *Heterogeneous databases* jest obecnie nieużyteczna, co potwierdza etykieta *Revised*. Klaster *3* można scharakteryzować jako obszar dokumentów o zarządzania wiedzą, ontologiach i semantycznym Webie. Bliskimi odpowiednikami w schemacie *CCS* są deskryptory zamieszczone w podklasie I.2. *Artificial Intelligence*/I.2.4 *Knowledge Representation Formalisms and Methods* klasy I. *Computing Methodologies*. Tym samym da się wywnioskować, że klasyfikacja *CCS* zdecydowanie nie odwzorowuje aktualnego stanu rozwoju nauk komputerowych. Zauważono także brak w drzewie pierwotnym deskryptora przedmiotowego *Visualization* lub mu podobnego, w przeciwieństwie do kolekcji słów kluczowych z częstym jego występowaniem. W takim razie nie da się za pomocą *CCS* precyzyjnie zaklasyfikować dużą ilość artykułów z głównym słowem kluczowym „wizualizacja". Klaster *4* scharakteryzować można jako obszar publikacji o społecznym zastosowaniu technologii informacyjno-komunikacyjnych: *Social Computing*: *Wikipedia, Collaboration, Collaborative learning*.

Za pomocą tych dwóch jednostek analitycznych można wyłowić najistotniejsze cechy tematycznych organizacji obu schematów (klasyfikacji pierwotnej

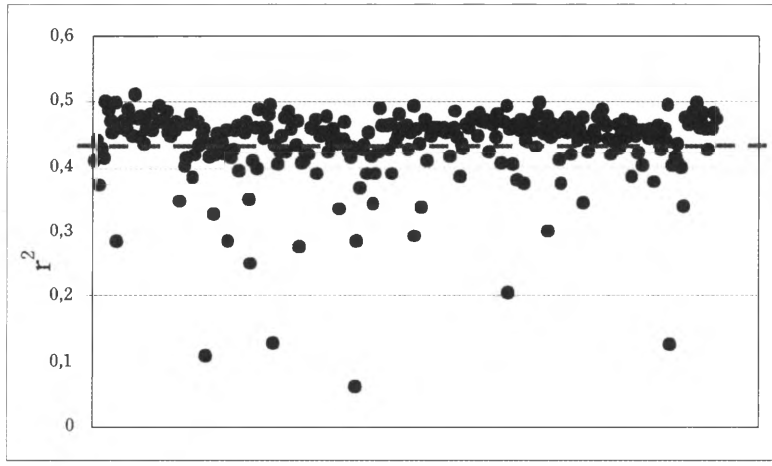
i klasteryzacji wynikowej) i tym samym nadać im właściwe nazwy oraz sensownie porozmieszczać w granicach stałych struktur klas głównych. Z powyższych badań można wywnios-kować, że podstawowym objawem transformacji przestrzeni klasyfikacyjnej na semantyczną jest zredukowanie poziomów hierarchii z trzech do jednego. Hierarchia pierwotnego systemu jest niepotrzebnie zagnieżdżona. Jak pokazują wyniki eksperymentu, jeden poziom wystarczy, aby uzyskać sensowną wizualizację (oraz jej pochodną klasteryzację) dokumentów z zachowaniem ich podobieństwa tematycz-nego.

3.4. Rozszerzone metody obróbki danych

a) Modyfikacja zestawu danych

Podczas analizy wzoru rozkładu elementów klasyfikacji powstało pytanie, jak się zmieni rozłożenie węzłów na sferze przy zmianie danych wejściowych. Zdecydowano się na usunięcie z zestawu danych dokumentów klas, które wykazują największe rozproszenie na całej powierzchni. Tak wytypowano klasę I (*Computing Methodologies*), zawierającą 72 podklasy. Konfiguracja dokumentów charakteryzowała się pięcioma wyraźnie zarysowanymi obszarami, które nakładały się na klastry innych klas, w ten sposób zamazując granice kategorii. Czy usunięcie dokumentów, należących do klasy I. *Computing Methodologies* polepszy rozkład powierzchniowy węzłów?

Tym samym wymiar macierzy podobieństwa został zredukowany do 281. Wykres 7 pokazuje rozkład wartości r^2. Widać duże podobieństwo do poprzedniego rozkładu (Wykres 2). Końcowa wartość kwadratu promienia podobnie jak dla pierwotnych danych wyniosła 0.46. Po usunięciu nieistotnych klas (o dużym odchyleniu) liczba węzłów do rozmieszczenia na sferze wynosiła 279. Przy obliczaniu położenia węzłów dokumentów posłużono się sprawdzona proporcją wag klas 0.6:0.4. Wynikowa kolorowa mapa wizualizacji jest załączona na Ilustracji 6.



Wykres 7. Rozkład wartości r^2 dla zmodyfikowanego zestawu danych (usunięto klasę **I**). Linia przedstawia wartość średnią kwadratu promienia sfery.

Źródło: Opracowanie własne.

b) Kwantyfikacja struktury map wizualizacji

Niejednorodny rozkład elementów klasyfikacji stawia podstawowe pytanie: za pomocą jakiej jednostki można go oszacować? W przypadku jednorodności rozwiązaniem byłaby niewątpliwie gęstość rozmieszczenia obiektów na płaszczyźnie. Zadanie jest niełatwe poprzez innowacyjność metody badawczej, a wyniki wydają się być o tyle ścisłe, o ile nasuwają się intuicyjnie rozwiązania. Pomocne tu okazało się zastosowanie własności **fraktali**[17] jako graficznych obiektów nieliniowych, które stały się bardzo popularne w zastosowaniach nie tylko naukowych, medycznych (diagnostyka), biznesowych (prognozowanie) ale również artystycznych (obrazy cyfrowe). Można znaleźć wiele galerii fraktali, jak również oprogramowania do ich generowania. W znaczeniu popularnym oznaczają one zbiory, wykazujące cechy: skalowalności, samopodobieństwa, rekurencyjności i struktury, której opis w języku geometrii Euklidesowej jest dość skomplikowany. Praktyczną, wykorzystywaną w pracy własnością fraktala jest jego **wymiar fraktalny,** który jest jest mniejszy od wymiaru topologicznego[18].

Z uogólnionym pojęciem wymiaru wiąże się liczba zmiennych (stopnie swobody) w systemach dynamicznych. Przestrzeń Euklidesowa charakteryzuje się trzema wymiarami. Wymiar fraktalny odnosi się do własności „samopodobieństwa" (obrazy ich struktury są takie same w każdej skali) i tak często się nazywa. Jeśli liniowy wymiar fraktali zmniejszamy $1/r$ razy, to na podstawie zależności liczby N samopodobnych kopii w obszarze obiektu:

[17] Termin wprowadzony po raz pierwszy w 1966 r. przez Mandelbrota od łacinskiego słowa *„fractus"*, co oznacza złamany, cząstkowy. Por. *Fraktal*. W: *Wikipedia. Wolna Encyklopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Fraktal.

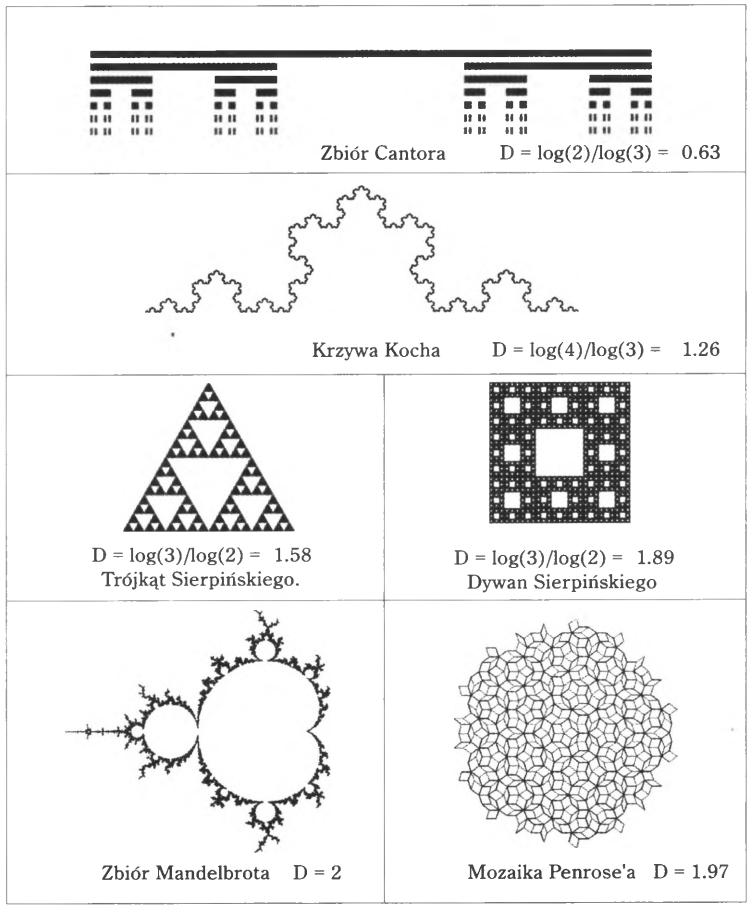
[18] B. Mandelbrot: *Fractal Geometry of Nature*. USA: W. H. Freeman & Co 1982, s. 6-20.

$$N=r^D \tag{7}$$

wymiar fraktalny D może być zdefiniowany w następujący sposób[17]:

$$D = \log(N)/\log(r). \tag{8}$$

Własności wymiaru samopodobieństwa dla fraktali obrazuje następująca zależność: jeżeli w płaskiej figurze geometrycznej (np. kwadracie) dwukrotnie powiększymy boki - jej powierzchnia wzrośnie czterokrotnie, przeprowadzając takie operacje na fraktalu jego powierzchnia zwiększy się mniej niż czterokrotnie. Wyliczenie wymiaru fraktalnego D jest możliwe na podstawie relacji między powierzchnią lub objętością fraktali. Wymiar ten przyjmuje dla fraktala wartości niecałkowite. **Niesie on w sobie bardzo ważną informację, wskazując w jaki sposób fraktal wypełnia przestrzeń**, w której jest osadzony. Wymiary fraktalny figur regularnych, np. linii, kwadratu, sześcianu są takie same jak topologiczne, czyli: 1, 2, 3 odpowiednio. Na Rysunku 33 przytoczono kilka przykładów wymiaru samopodobieństwa znanych fraktali:



Rysunek 33. Wymiary samopodobieństwa popularnych fraktali

Źródło: J. Budrewicz. *Fraktale*. Warszawa: Wydawnictwo Naukowo-Techniczne, 1996. s. 12-20.

Widać, że zbiór Mandelbrota ma wymiar taki sam jak jego wymiar topologiczny, a jest to fraktal na co wskazują jego pozostałe cechy. Następne przykłady – na Rysunku 34, konfiguracje obiektów spotykanych w przyrodzie:

D = 2,66

Brokuł

D = 2,79

Powierzchnia ludzkiego mózgu

D = 1,52

Linia brzegowa Norwegii

D = 2,4

Drzewo

Rysunek 34. Wymiary fraktalne obiektów fizycznych

Źródło: Opracowanie własne na podst. M. Frame, B. Mandelbrot, N. Neger. *Fractal Geometry* [on-line]. Yale University [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://classes.yale.edu/fractals/; *List of fractals by Hausdorff dimension*. W: *Wikipedia. The Free Encyclopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/List_of_fractals_by_ Hausdorff_dimension.

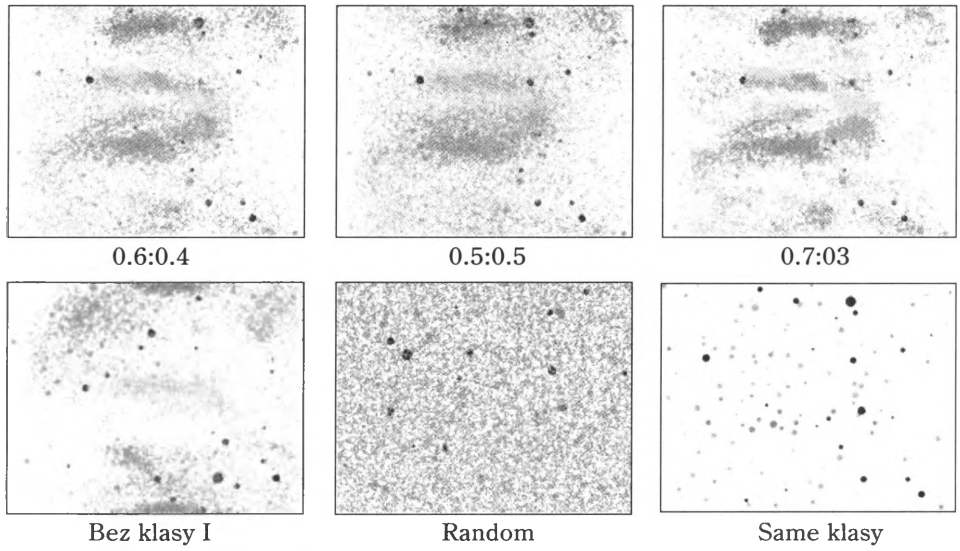
Wiele obiektów i zjawisk spotykanych w przyrodzie może być modelowanych za pomocą geometrii fraktali. Jako przykłady mogą posłużyć: linia brzegowa, zbocza górskie, systemy komórkowe, powierzchnia białek, struktura polimerów, chmury, dyfuzyjnie limitowana agregacja (np. podczas elektrolitycznego wydzielania metali).

Można zauważyć, iż wymiary samopobieństwa większości fraktali zawieraja się w przedziale [1,2], czyli wymiary ich są mniejsze niż figury płaskiej, a większe niż prostej. Tak więc fraktale, których wymiary samopodobieństwa zawierają się w tym przedziale, nie są już prostymi, a jeszcze nie są figurami płaskimi. Obiekty 3D w przyrodzie nie są płaskie i „dążą" do zwiększenia wymiaru fraktalnego powyżej liczby dwa. Im bardziej mają skomplikowaną topologię powierzchni, tym bardziej ta wielkość zbliża się do trzech.

Do policzenia wymiaru fraktalnego dwuwymiarowego rozkładu klasyfikacji z powodzeniem użyto modułu *Fractals* programu *Matlab*. W tym celu dane trzeba było odpowiednio przygotować. Mapy wizualizacji przekonwertowano do skali odcieni szarości. Zakwalifikowane do analizy obrazy przeskalowano do jednakowych rozmiarów z dokładnością do jednego piksela. Dla lepszej precyzji pliki wyeksportowano w trzech formatach: TIF, JPG i BMP. Najkrótszy czas obliczeń wskazywał optymalny format dla całej grupy danych. Struktury fraktalne łatwiej jest porównać jeśli są one jednocześnie dostępne w jednym rzucie obserwacji.

Mały obrazek o niskiej rozdzielczości jest postrzegany jako ostry o wyraźnych krawędziach, mózg ludzki w procesie percepcji uśrednia strukturę tekstury powodując wyraźne odróżnienie różnych treści. Na dużych obrazach proces ten ulega rozmyciu i obraz jawi sie jako „niewyraźny". Znają to zjawisko dokładnie graficy komputerowi – na stronie internetowej zamieszczenie małego obrazka powoduje jego dobrą ostrość mimo iż powiększony byłby postrzegany w gorszy sposób. Dlatego wyniki wizualizacji graficznej w postaci miniatur map różnych konfiguracji, w celach porównawczych umieszczono blisko siebie na Rysunku 35. Pozwala to jednocześnie porównać, w dostępnym polu widzenia, własności struktur teksturowych. W górnym wierszu można zobaczyć jak się zmieniał rozkład przy modyfikacji wag klasyfikacji podstawowej i dodatkowych – ostatecznie wybrano relacje 0.6:0.4 z powodów podanych w rozdziale 3.2.d. Obrazki w pierwszej kolumnie – są to wizualizacje zestawów z różną ilością danych: dla kompletnej klasyfikacji i z usuniętą klasą I. *Computing Methodologies* i pochodnymi dokumentami. Ostatnia mapa ilustruje rozkład węzłów samych klas. Wykonano również mapę dla rozkładu losowego typu Random, który miał posłużyć jako charakterystyka kontrolna. Użyto funkcji *RAND*, jako generatora tablicy liczb losowych, elementy której odpowiadają równomiernemu rozkładowi w przedziale [0,1]. Ten zakres naturalnie można poszerzyć do z góry zdefiniowanych wartości granicznych dla zmiennych. Długość geograficzna mieści się w przedziale $0 \leq \phi < 2\pi$, natomiast szerokość geograficzna może przyjmować wartości – $\frac{1}{2}\pi \leq \theta \leq \frac{1}{2}\pi$. Stąd pierwsza zmienna może się mieścić w przedziale [0, 6.28], druga – [-1.57, 1.57]. Wartości losowe dwóch zmiennych ϕ, θ wygenerowano niezależnie.

Na wszystkich mapach „prześwituje" ten sam rozkład węzłów klas i podklas, na który nakłada się obraz zmapowanych dokumentów. Rozważany w odosobnieniu od węzłów dokumentów (ostatnia mapa na Rysunku 35), wykazuje nieskomplikowaną strukturę fraktalną, dlatego jego wymiar jest najniższy.



Rysunek 35. Zestawienie map wizualizacji po denaturacji dla modyfikowanych zestawów danych.

Źródło: Opracowanie własne.

Porównanie map „na oko" jednak nie zapewnia merytorycznej oceny jakości rozkładu. W ilościowej ewaluacji przydał się opisany wyżej wymiar fraktalny. Otrzymano następujące wartości dla podanych przykładów:

0.6:0.4	0.5:0.5	0.7:03	Bez klasy I	Random	Same klasy
2.24	2.56	2.19	2.19	2.28	1.68

Zmniejszenie D (z 2.24 do 2.19) dla danych z okrojoną klasą **I** jest wytłumaczalne: mniejsza ilość obiektów uprościła mapę. Wyższy wymiar świadczy o większym stopniu kompleksowości struktury. Podobna sytuacja powstała w przypadku zwiększenia wagi klasy podstawowej. Klastry (3 mapa w pierwszym wierszu) zmniejszyły swoją objętość poprzez skupienie się węzłów dokumentów dokoła odpowiednich klas podstawowych. Dla wag połowicznych (środkowa mapa w 1-ym wierszu), klastry się rozmywają i rozkład zbliża się do równomierności, a to jest charakterystyczne dla rozkładu dokumentów *Random*. Trudno tu o wyjaśnienie takiej niekonsekwencji, iż dla pierwszej mapy wymiar się zwiększa i jest maksymalny ze wszystkich zmierzonych – 2.56, a dla losowej dystrybucji – zmniejsza się do 2.28. Natomiast spróbujmy odwołać się do wejściowego układu klasyfikacji – drzewa o wymiarze topologicznym 1. Odpowiedni wymiar fraktalny byłby wartością większą od 1 i mniejszą od 2. Drzewiaste struktury są wyraźnie hierarchiczne. Natomiast mapy wizualizacji po **desaturacji**[19] koloru i w procesie przetwarzania fraktalnego częściowo zgubiły informację o poziomach hierarchii. Dlatego większy wymiar w przypadku równomiernego rozkładu można uzasadnić minimalizacją hierarchii.

[19] Usunięcie w obrazie informacji o kolorze, w wyniku czego zostaną odcienie szarości.

Z drugiej strony można spróbować opisać teksturę fraktalną za pomocą **lakunarności**[20]. Jest to miara wypełnienia fraktalem przestrzeni. Im większa dystrybucja dziur i prześwitów w obrazie fraktalnym, tym większa jego lakunarność. Obliczenie tego parametru jest dosyć skomplikowane, dlatego ograniczymy się do wizualnej estymacji. W wyniku obserwacji, zauważono, iż największą wartością lakunarności charakteryzuje się rozkład klas, najmniejszą – losowy. Usunięcie klasy „zwolniło" przestrzeń z określonej ilości punktów, dało więcej światła w rozkładzie – czyli lakunarność podwyższyła się. Zrozumiałe jest również to, iż im większa waga klasy podstawowej (większy stopień akumulacji dokumentów) lakunarność wzrasta.

O potencjalnym dostosowaniu/podobieństwie (ang. *potencial parallelism*) teorii fraktali i organizacji wiedzy wzmiankowano w pracach z ostatnich lat, rozważających temat w interdyscyplinarnej perspektywie badań struktur złożonych[21]. A. Barát[22] zaznacza, iż naukowcy szukający sposobów opisu formujących się modeli organizacji wiedzy mogą wykorzystywać prawa fizyki, jak na przykład zasadę nieoznaczoności Heisenberga[23], entropię[24], rozkład Boltzmana[25] czy teorię chaosu. Podstawowym jej argumentem, jest to, że prawa fizyki są w przyrodzie powszechne i uniwersalne. Tak jak niezliczoną ilość przykładów fraktali można zobaczyć w naturze (Rysunek 34), tak reprezentacje pojęć, tworzone w naszym mózgu na skutek obserwowania natury i relacji pomiędzy obiektami mogą przyjmować strukturę nieliniową, czyli fraktalną.

W tym rozdziale szczegółowo przedstawiono metodykę oraz przebieg prac badawczych. Pierwszy etap eksperymentu (podstawowy) składał się z kolekcjonowania, przewarzania i wizualizacji badanych obiektów. Przy czym podstawowa analiza danych prowadziła do wizualizacji, która była pierwotnym celem niniejszej pracy. W dodatkowej analizie wykorzystano nowe koncepcje nad badaniem struktury i dynamiki złożonego zestawu danych.

---

[20] B. Mandelbrot, dz. cyt. s. 310-319; R.E. Plotnick, R.H. Gardner. Lacunarity *indices as measures of landscape texture*. Landscape Ecology, 1993. Vol. 8, nr 3, s. 201-211.

[21] Á.H. Barát: *The Structures of Concept And its Connection to Sciences*. W: *Proceedings of IX ISKO Congres Spain Group, New Perspectives for the organization and dissemination of knowledge*. Valencia: UPV, 2009, s. 372-379; Ch. Crowley. *Overview of Complexity* [on-line]. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://wynchar.com/charlie/Complexity/ overviewOfComplexity.html; D. Sperber. *Why Rethink Interdisciplinarity?* [online] *Interdisciplines* 2009. [dostęp 16 maja 2009]. Dostępny w World Wide Web: http://www.interdisciplines.org/interdisciplinarity/papers/1.

[22] Á.H. Barát, dz. cyt.

[23] Zasada nieoznaczoności, odkryta przez Wernera Heisenberga w 1927 roku, mówi iż na poziomie mikroświata zwanym też poziomem kwantowym, nie można z dowolną dokładnością wyznaczyć jednocześnie położenia i pędu cząstki. Istnieje tendencja adaptacji tej zasady również do innych par wielkości, których nie da się jednocześnie zmierzyć z dowolną dokładnością.

[24] Entropia, czyli miara chaosu układu. Zgodnie z drugą zasadą termodynamiki, jeżeli układ termodynamiczny przechodzi od jednego stanu równowagi do drugiego, bez udziału czynników zewnętrznych (a więc spontanicznie), to jego entropia zawsze rośnie.

[25] Rozkład Boltzmana, stosowany w opisie termodynamicznych układów składających się z dużej liczby cząstek. Opisuje sposób obsadzania poziomów energetycznych przez atomy, cząsteczki lub inne cząstki w stanie równowagi termicznej. Prawdopodobieństwo obsadzenia stanu maleje wykładniczo wraz z energią poziomu.

Rozdział 4

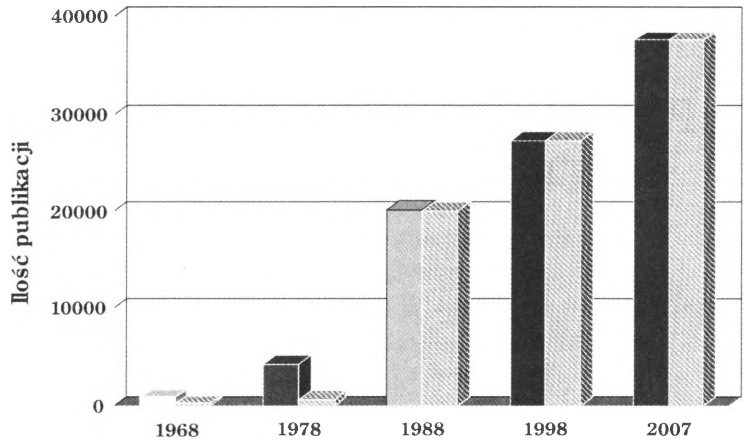
IMPLEMENTACJE PRAKTYCZNE SYSTEMU WIZUALIZACJI

4.1. Charakterystyki czasowe domeny naukowej

Jednym z zaplanowanych zadań było powtórzenie procedur wizualizacji dla okresu poprzedzającego 2007 rok z interwałem co 10 lat. Miało to wykryć wszystkie zasadnicze zmiany w światowej literaturze naukowej w zakresie informatyki na przestrzeni tych lat. Postanowiono zbadać różnice, o ile takie się zarejestruje, w strukturze, hierarchii, terminologii oraz wolumenowe. Na podstawie wyników takich badań przy wsparciu wiedzy z historii rozwoju nauk komputerowych możliwe jest wnioskowanie o ewolucji ich przedmiotowej klasyfikacji, kierunkach integracji z innymi dyscyplinami oraz o trendach w przyszłości.

Jako pierwszy, wybrano rok 1998, ponieważ w końcu tego roku zadebiutowała najpopularniejsza wyszukiwarka internetowa Google, która z pewnością przyczyniła się do pogłębienia problemu destrukturalizacji zasobów sieciowych. Następnie zbadane były okresy 1988, 1978 i na koniec rok 1968. Nie wszystkie artykuły, szczególnie z wcześniejszych lat, poddane zostały klasyfikacji, dlatego wykres słupkowy 8 pokazuje liczby publikacji wymienionych lat wraz z częścią zaklasyfikowanych danych, które były obiektem wizualizacji. Widać, że dopiero od 1988 r. zaczęto masowo praktykować klasyfikowanie prac naukowych, a otrzymany w procesach eksperymentu zbiór danych nabrał cech kompletności. Należy przypomnieć, że w 1982 r. ACM opublikowało zasadniczo nową wersje systemu *CCS*, którą często aktualizowało. Z pewnością schemat klasyfikacji był skutecznie dopasowywany do ówczesnego stanu rozwoju nauk komputerowych, skoro prawie wszystkie prace (99.1%) powstałe w 1988 r. znalazły się na właściwych gałęziach drzewa klasyfikacyjnego.

W związku z powyższym do końcowej analizy ewolucji struktury nauk komputerowych nadają się trzy okresy na osi czasu: 1988 r., 1998 r. oraz 2007 r. Istotnym jest również to, że mała liczba dokumentów (przez to nieliczne wspólne dokumenty w krzyżujących się podklasach) zaniży precyzję naszej metody mapowania. Tabela 6 prezentuje parametry modelu na etapach przetwarzania i wizualizacji danych.



Wykres 8. Ilodci publikacji i ich poklasyfikowana część (słupki z teksturą deseniową) w kolejnych latach

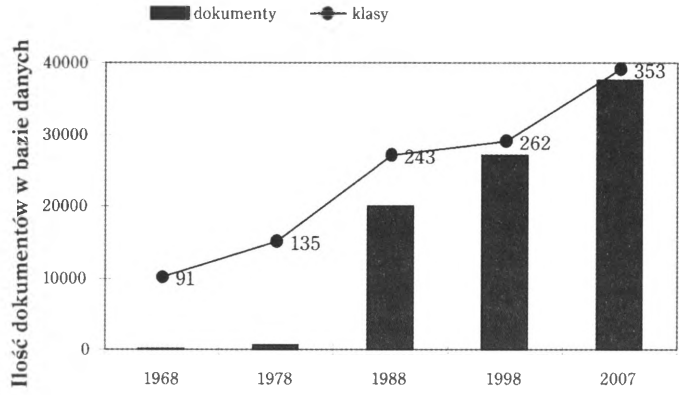
Źródło: Opracowanie własne.

Wyekstrahowana w obliczeniach liczba klas i podklas krokowo ulegała zmianie w każdym kolejnym roku, co ilustruje Wykres 8. Największe przyrosty były zaobserwowane w dziesięcioleciach: poprzedzającym 1988 r. oraz w ostatnim.

Porównanie parametrów wizualizacji dla kolejnych lat **Tabela 6.**

	1968	1978	1988	1998	2007
Il. dokumentów	209	545	19950	27149	37543
Il. (pod)klas	**91**	**135**	**243**	**262**	**353**
Il. obiektów wyeli-minowanych	1	0	1	0	5
Kwadrat promienia sfery	0.441	0.439	0.438	0.439	0.447
Kruskal Stress	0.342	0.335	0.346	0.345	0.343

Źródło: Opracowanie własne.

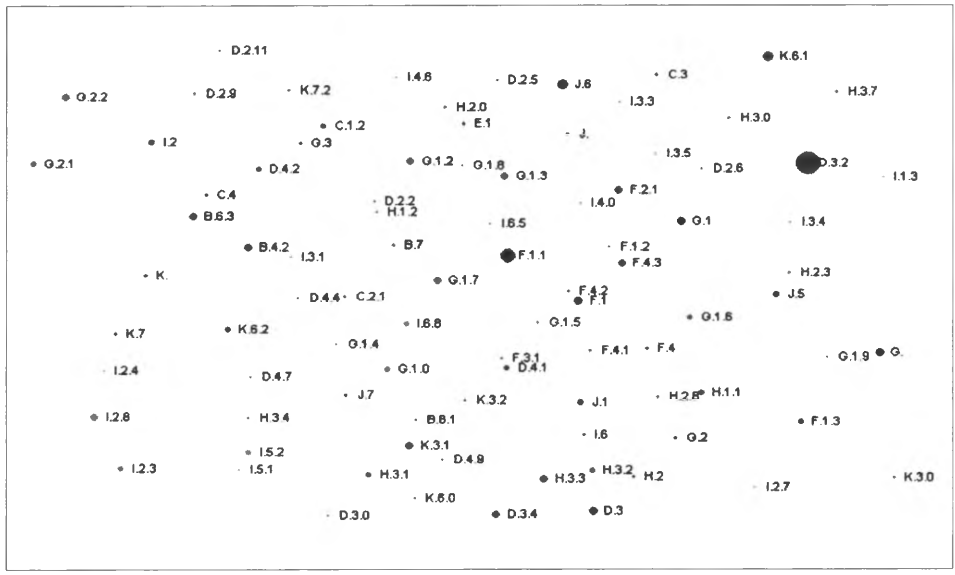


Wykres 9. Zestawienie liczby zbadanych dokumentów i klas w kolejnych latach

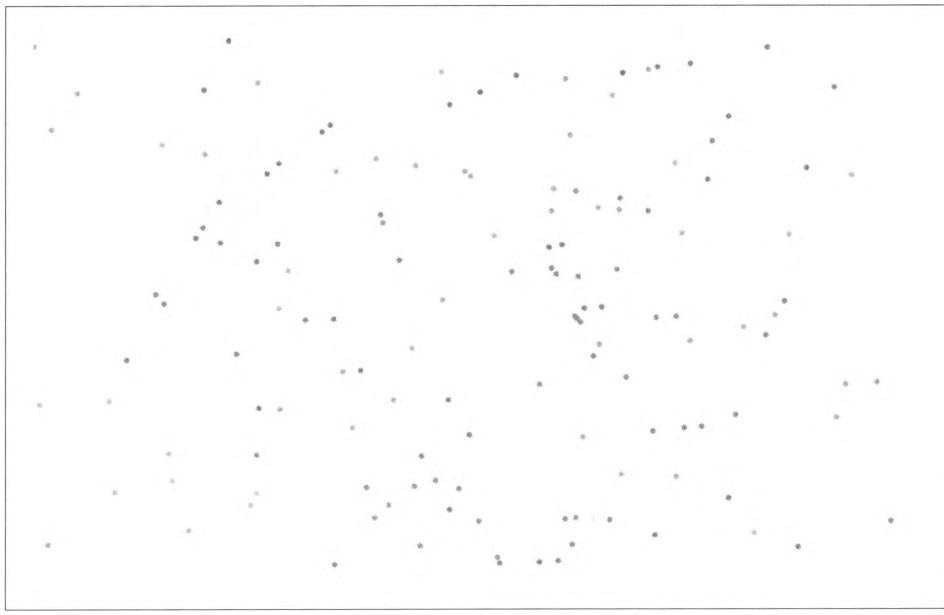
Źródło: Opracowanie własne.

Dla pierwszych dwóch okresów wyniki obliczeń ilości klas niosą niski poziom ufności na skutek małej liczby danych. Trzeba również liczyć się z tym, że mogą występować tak zwane niefunkcjonalne podklasy, których nie sposób wykryć jedynie przy pomocy skanowania metadanych dokumentów. Dane o wszystkich możliwych węzłach, znajdujących się na drzewie klasyfikacyjnym udało się uzyskać tylko dla współczesnego schematu *CCS*. Na chwilę obecną, czyli w 2009 r. liczba klas i podklas wynosi **367**. Stąd tylko 2% klas nie zostało zbadanych.

Wyniki zestawienia przedstawione w tabeli są bardzo podobne, a to jest kolejny powód do przeprowadzenia wiarygodnej interpretacji zmian rozkładu graficznego na mapie. Jakość dopasowania *MDS*, na co wskazuje parametr *stress* - jest taka sama w każdym teście. Dla wszystkich przypadków kwadrat sfery modelu wynosi 0.44 z dokładnością do 0.005. Tak jak dla 2007 r., pojawiały się nieliczne węzły mocno oddalone od wyznaczonego promienia powierzchni sfery, lecz nie kwalifikowały się one do wyeliminowania, ponieważ zawierały bardzo liczne grupy dokumentów. Do stopniowania rozmiarów węzłów klas o bardzo zróżnicowanych pojemnościach zastosowano tak jak pierwotnych procesach przetwarzania danych, funkcję logarytmiczną.

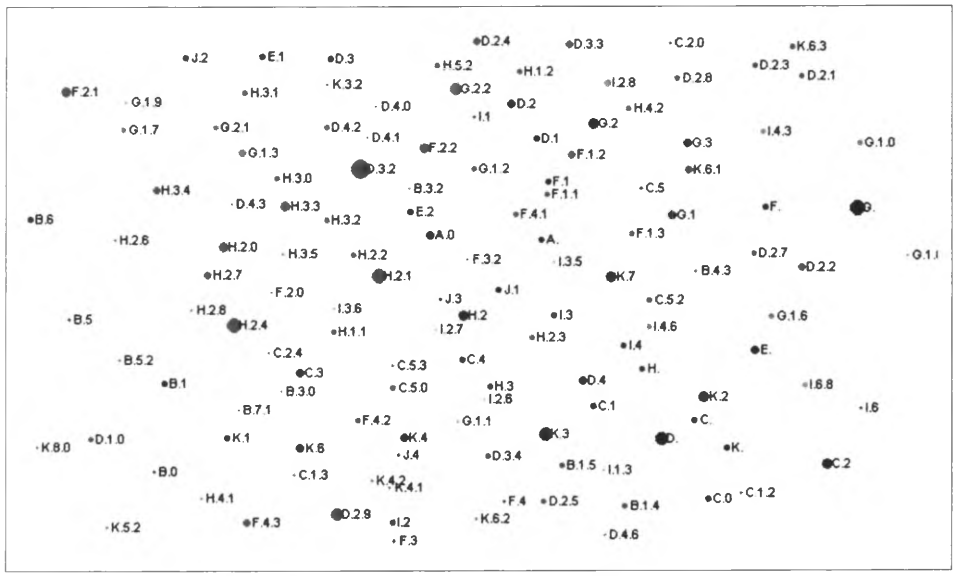


Klasy

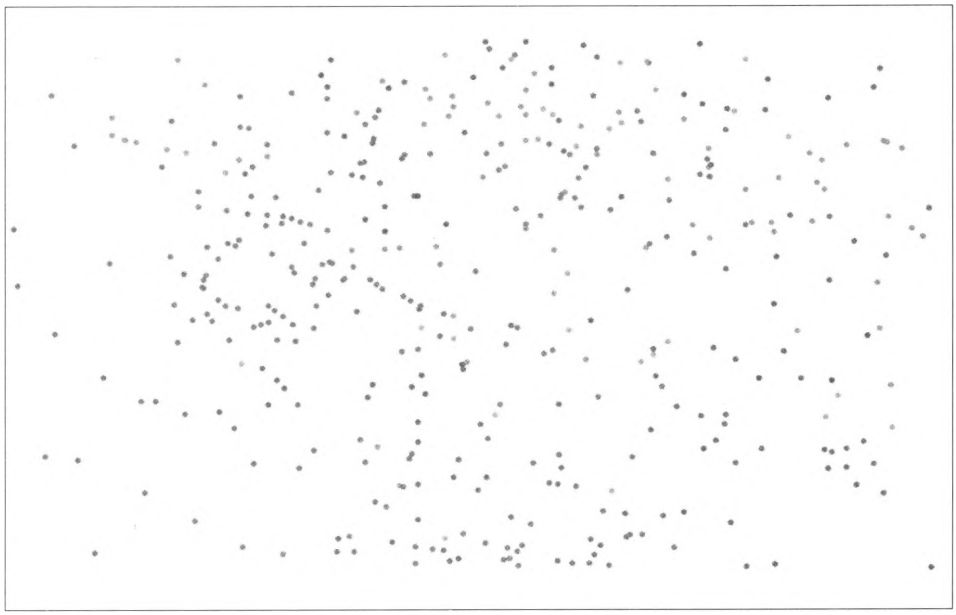


Dokumenty

Ilustracja 12. Mapy klasyfikacji CCS z 1968 r.



Klasy



Dokumenty

Ilustracja 13. Mapy klasyfikacji CCS z 1978 r.

Poniżej rozważymy osobno zestawy map dla każdego roku.

1968 i 1978

Na Ilustracji 12 widać, że większą wartość informacyjną zawiera mapa klas, a nie węzły dokumentów w ilości ponad 200, na tyle rzadko rozsianych po całej powierzchni, że nie ma możliwości ich nałożenia się. Podobnie wygląda rozkład dla następnego roku z listy – 1978 (Ilustracja 13), jednak da się zauważyć nieznaczną kumulację danych klasy H. *Information Systems* – pomarańczowe punkty. Taki formujący się klaster świadczy o intensywnych badaniach systemów informacyjnych w tym okresie. Lata siedemdziesiąte ubiegłego wieku – to był czas kiedy używano komputerów *mainframe*[1], zaczęto centralizować dane oraz procesy obliczeniowe. Systemy informacyjne przyjęły się w sektorze biznesowym do takich zadań jak: organizacja listy płac, księgowość czy spisy inwentarzowe[2]. Podklasy o zauważalnych rozmiarach węzłów to: H.2.1, H.2.2, H.2.4, H.3.3. Zatem w tym czasie zamieszczono wiele prac o zarządzaniu baz danych, ich projektowaniu, a także zagadnieniach wyszukiwania i przechowywania informacji (H.3). Z drugiej strony odnotować można brak lub małe natężenie koloru turkusowego, co dowodzi słabego rozwoju metodologii komputerowych.

Wzór utworzony przez klasy zdradza dużą objętość podklas klasy D. *Software* na tle innych obiektów. Największy z nich to D.3.2, odnoszący się do ówczesnych języków programowania i ich klasyfikacji. Z diagramu historii języków programowania[3] dopatrzeć się można, iż w tych latach swój początek biorą takie języki jak *Logo* (1968), *Prolog, Pascal* (1970), *sh* (1971), *Ada* (1979). Wielkości obiektów klasy G. *Matematyka obliczeń* też mówią o namnożeniu się prac na temat analizy numerycznej, prawdopodobieństwa i statystyki. Można byłoby dalej rozważać cechy lokalizacji poszczególnych kategorii na podstawie rozmiarów obiektów wizualizacyjnych reprezentujących klasy i podklasy. Jednak nie zdobędziemy to istotnych informacji o ewolucji schematu *CCS*, jeśli nie sposób dopatrzyć się podobieństwa w rozkładzie lub tendencji odwrotnych. Liczba danych jest zbyt mała na wykonywanie kolejnych faz analizy.

1988

Posługując się „tradycyjną" strategią analityczną, zaczniemy od scharakteryzowania największych węzłów na Ilustracji 14:

– H.1. *Information Systems/Models and Principles*. Kategoria ta przeznaczona dla artykułów o teorii informacji oraz już aktualnych wówczas zagadnieniach kom-

[1] *Mainframe* – superkomputery, używane głównie przez duże organizacje dla finansowych i sta-tystycznych zadań. Są to systemy o dużej wydajności przetwarzania danych. Por. *Słownik pojęć komputerowych...*, s. 194.

[2] M. Rana: *Historical Perspective on Information Systems* [on-line]. *Information Systems based on Logistics Perspective* [University of Houston] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.uh.edu/~mrana/try.htm.

[3] *The History of Programming Languages* [on-line]. O'Reilly Media [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://oreilly.com/news/graphics/prog_lang_poster.pdf.

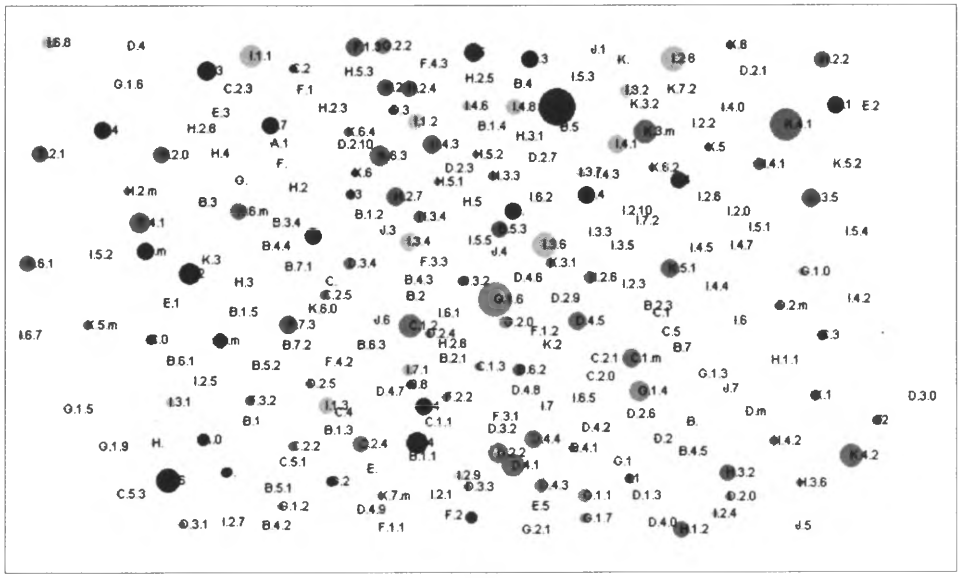
presji danych i cyfrowego kodowania nadmiarowego, używanego w celach ochrony danych przenoszonych z pomocą sygnałów cyfrowych. W pobliżu usytuowane są inne mniej liczne tematy tejże klasy głównej: bazy danych, wyszukiwanie informacji oraz interfejsy i prezentacja informacji.

– G.1.6. *Numerical Analysis/Optimization.* Termin optymalizacja w matematyce iden-tyfikowany jest z problemem znalezienia minimum zadanej funkcji. Tu omawiało się metody optymalizacji takie jak: gradientu, najmniejszych kwadratów, programowanie liniowe i nieliniowe.

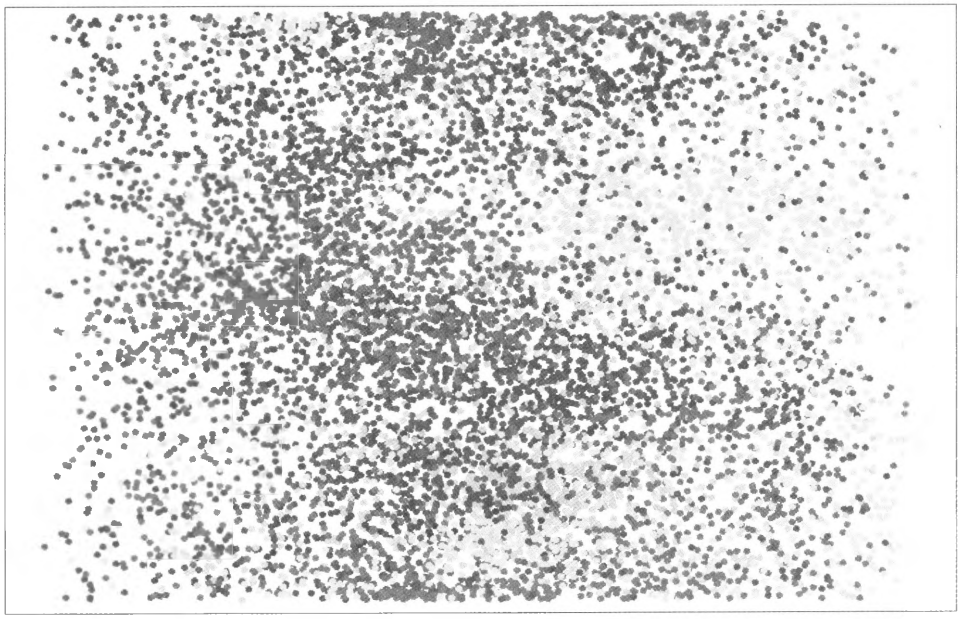
– K.4.1. *Computers and Society/Public Policy Issues.* Sekcja ta odnosi się między innymi do: etyki komputerowej, prywatności i bezpieczeństwa pracy z komputerem, regulacji prawnych tych kwestii. Nieobce w owym okresie były także problemy nadużywania i przestępczości komputerowej.

Zauważalne są mniej ważące tematy, np. projektowanie logiki (B.6), systemy operacyjne (D.4), inżynieria programowania (D.2), architektury procesorów (C.1), grafika komputerowa (I.3) itp. Dostrzegalne jest zcentralizowanie zbioru węzłów klasy D. *Software* (zielone glify), jak również rozproszenie danych klas I. *Computing Methodologies*, K. *Computing Milieux* po całym polu mapy, wskazujące na ogólne wykorzystanie tych kategorii. Rozkład dokumentów jest z wysoką dokładnością jednolity, nie wykazujący jednak cech ścisłej klasteryzacji. Przytoczone na następnej Ilustracji 15 miniatury map osobnych klas lub ich kombinacji dają możliwość zaobserwowania cech charakterystycznych. Badanie pojedynczych konfiguracji w danym przypadku daje kompletniejszy obraz formujących się działów. Bardzo charakterystyczne, że dla klas głównych o symbolach **H** i **C** skupiska węzłów tworzą szerokie ciągłe pasma – takie „drogi mleczne" (pamiętajmy, że granice mapy są ciągłe). Świadczą o tym, że artykuły o tematyce systemów komputerowych i systemów informacyjnych należą do kategorii, mających tendencję do uporządkowania swej struktury. W latach osiemdziesiątych XX w. już używano 8-bitowych komputerów osobistych (*Commodore, Atari, Spectrum*) oraz instalowano sieci lokalne w centrach informatycznych licznych instytucji. Służyły one głównie do zadań automatyzacji istniejących procesów. Najintensywniejsze przenikanie się podklas zanotowano dla par klas **H**, **I** oraz **C**, **D**. Pierwsza kombinacja mówi o powszechnym wykorzystywaniu nowych metodologii w systemach informacyjnych. Druga przy konfrontacji z ogólną mapą klas (p. Ilustracja 14) potwierdza, że systemy operacyjne (a w 1988 r. powstały między innymi *Macintosh OS, OS/400, SunOS 4.0, RISC*[4]) projektowano w oparciu o procesory o określonej architekturze – największe węzły przypadają na podrzędne klasy C1. *Processor Architectures.*

[4] *Concise Ecyclopedia of Computer Science...*, s. 572-579.

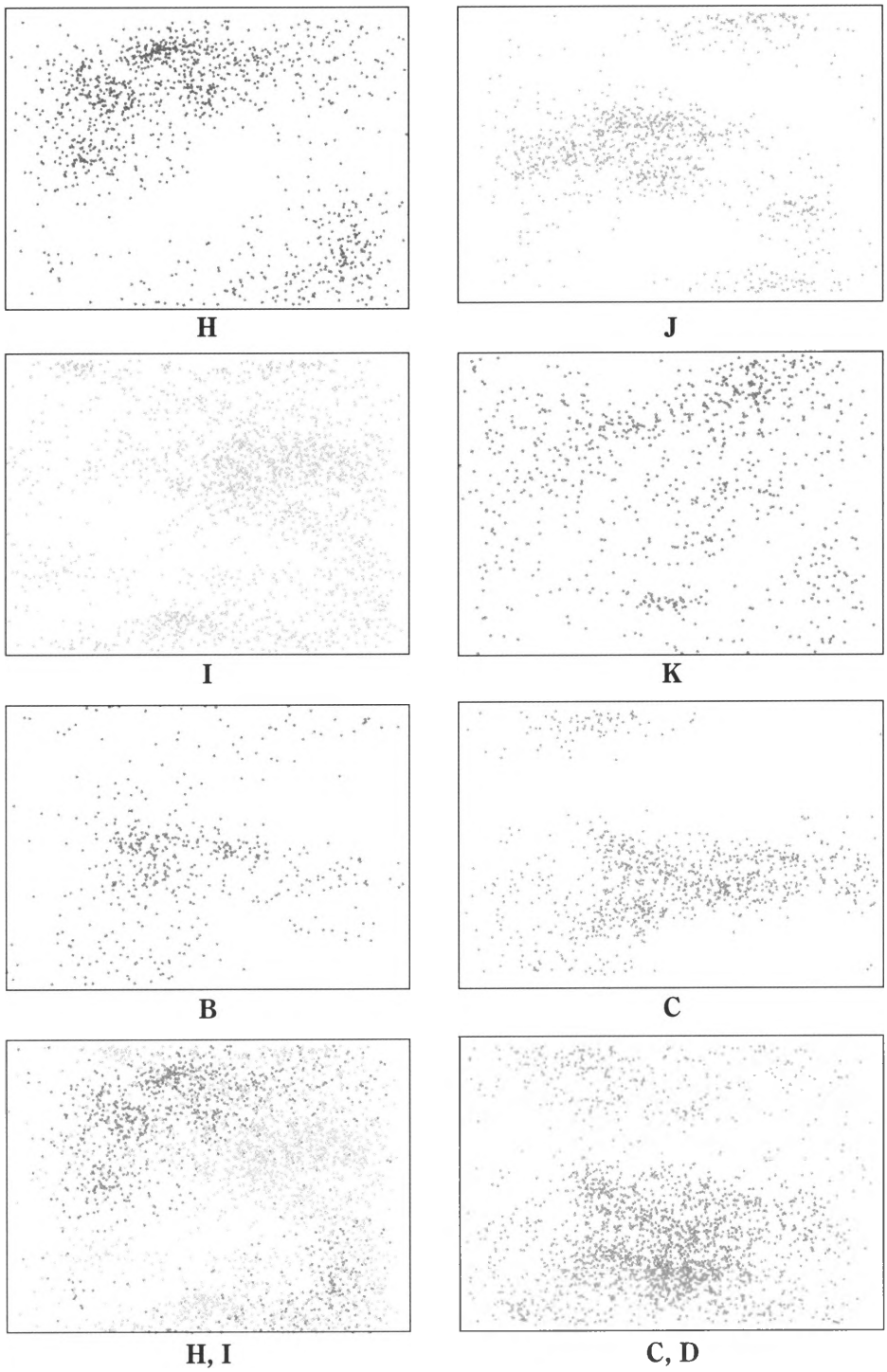


Klasy



Dokumenty

Ilustracja 14. Mapy klasyfikacji CCS z 1988 r.



Ilustracja 15. Mapy klasyfikacji CCS z 1988 r. dla wybranych kombinacji klas.

Ilustracje 16 i 17 prezentują wyniki wizualizacji dla publikacji opublikowanych w 1998 r. Najwięcej publikacji pochodzi z klasy G. *Mathematics of Computing*. Są tu dwa centra koncentracji artykułów na tematy G1. *Numerical analysis* i G.2 *Discrete mathematics*[5]. Prawdziwy rozwój analizy numerycznej jako osobnej dziedziny matematyki nastąpił z rozpowszechnieniem komputerów osobistych, gdy naukowcy tworzyć modele matematyczne, analizować i rozwiązywać problemy dotyczące rozmaitych naukowych dyscyplin. Następnie według liczności dokumentów jest węzeł J.2. Nauki fizyczne i inżynieria, czyli fizyka, chemia, astronomia, elektronika, matematyka i statystyka. Na mapie dokumentów już są zauważalne procesy klasteryzacji niektórych kategorii. Do dalszej analizy posłużymy się oddzielnymi mapami na Ilustracji 17.

Obiekty prawie wszystkich kategorii wykazują grupowanie wokół określonych centrów. Obiekty klasy G. *Mathematics of Computing* (kolor brązowy) organizują się dookoła dwóch wyżej wymienionych najliczniejszych klas. Pokrywają się one z klasą F. Teoria obliczeń, czyli dwa najbardziej teoretyczne działy w całym schemacie wykazują prawidłową lokalizację. Klasy D. *Software*, C. *Computer Systems Organization* oraz H. *Information Systems* formują ciągłe pasma danych, przy czym w przypadku klasy **C** – przypomina to odwróconą literę „S". Można się spodziewać, że ta konfiguracja poprzedza etap dyskretnej klasteryzacji, jaki odnotowaliśmy dla 2007 r. Punkty kategorii sprzęt i oprogramowanie (klasy **B**, **D**) nie nachodzą na siebie, co jest zrozumiałe. Widać, iż węzły klasy I. *Computing Methodologies* mają tendencje do grupowania się w kilka klastrów. Ostatecznie na mapie z 2007 r. naliczono 5 wyraźnych skupisk. Nowe metodologie stosowano w ówczesnych systemach informacyjnych i na ich podstawie powstawały aplikacje komputerowe szerokiego przeznaczenia – stąd pokrywanie się obszarów zajmowanych przez obiekty klas **H**, **I** oraz **J**.

Podsumowując wyniki badań w skali czasowej, zaznaczymy, iż moment uformowania się zauważalnych klastrów z węzłów dokumentów można przypisać dla roku 1998. Dokładność określenia daty niestety nie jest tu jest dość wysoka, ponieważ cały eksperyment składał z okresów wybieranych dyskretnie co 10 lat. Należy również spojrzeć na ten proces jako długotrwały. Wobec tego można uznać, iż w drugiej połowie lat dziewięćdziesiątych XX w. organizacja struktury klasyfikacji *CCS* została skutecznie zaadoptowana w praktyce. Na początku istnienia drzewa *CCS*, czyli w okresie 1978-1988 rozrastać się zaczęła literatura naukowa w zakresie systemów informacyjnych. Od 1988 r. dostrzegalna jest organizacja węzłów dokumentów w obrębie klas: **D**, **H**, **C**. Ich konfiguracje ewoluują od ciągłych pasm do osobnych ostatnich dwóch dekad. Natomiast systemy informacyjne, aplikacje oraz środowisko komputerowe tematycznie się pokrywają.

[5] **Analiza numeryczna** to zbiorcza nazwa działów matematyki (np. teoria obliczeń, analiza błędów, metody numeryczne), które zajmują się badaniem struktur ciągłych, nieprzeliczalnych, której głównym zadaniem jest badanie możliwości realizacji obliczeń przybliżonych. I odwrotnie, działy matematyki, które zajmują się badaniem struktur nieciągłych, to znaczy zawierających zbiory co najwyżej przeliczalne, (np. teoria grafów, teoria informacji, logika matematyczne, kryptografia itp.) składają się na **Matematykę dyskretną**.

4.2. Zastosowanie: wyszukiwanie dokumentów

Wydobywanie informacji (ang. *Infromation Retrieval – IR*), najpierw traktowane w informacji naukowej jako obszar badań, odwołujący się do typowych źródeł danych, aktualnie rozrosło się do rangi „tradycji badawczej"[6], obejmującej różnorodne ręczne i/lub automatyczne techniki szukania relewantnych dla użytkownika informacji. W ujęciu historycznym podejścia do wyszukiwania informacji dzieli się na: boolowskie, wektorowe i probabilistyczne. Pierwsze – traktowane jako klasyczne, opiera się na algebrze dwóch wartości logicznych {0,1} oraz operatorów Boole'a: AND, NOT i OR.

Druga metoda wykorzystuje tak zwany wektorowy model reprezentacji tekstu. Dokumenty w języku naturalnym są przedstawiane w sposób formalny przy użyciu wektorów w przestrzeni wielowymiarowej.

Procedurę tworzenia modelu przestrzeni wektorowej można podzielić na trzy etapy. Pierwszym jest indeksowanie dokumentów i „wyłuskanie" słów oddających treść dokumentu. Na drugim etapie zachodzi ważenie słów indeksowanych, czyli określenie, w jakim stopniu termin jest ważny dla dokumentu w odniesieniu do zapytania. Na koniec ustalana jest pozycja rankingowa dokumentu na liście odpowiedzi. Model wektorowy wyszukiwania dokumentów jest naturalnym rozszerzeniem modelu algebraicznego, mający tę zaletę, iż uwzględnia wagi charakteryzujące dokument. Jednym z wariantów modeli przestrzeni wektorowej jest metoda matematyczna zwana analizą ukrytych grup semantycznych (*Latent Semantic Indexing*), przybliżona w dodatku artykułu[7].

W podejściu probabilistycznym korzysta się z różnych modeli probabilistycznych, traktujących proces wydobywania informacji jako wielostanowy eksperyment losowy[8]. Zamiast cech określających podobieństwo dokumentów używa się prawdopodobieństwa ich relewancji. Na przykład w pracy[9] zdefiniowana została zasada prawdopodobieństwa rankingu wyników: *Jeśli wyszukiwane dokumenty uporządkowane są według malejącego prawdopodobieństwa relewantności danych, to system wyszukiwawczy charakteryzuje się najlepszą wydajnością*[10]. Z reguły większość modeli probabilistycznego wyszukiwania informacji obejmuje ocenę efektywności ścieżki, prowadzającej do kolekcji zawierającej potencjalne obiekty relewantne[11].

Jak zauważa B. Hjørland z *Royal School of Library and Information Science* w Kopenhadze, pomiędzy badaniami metod wyszukiwania informacji a badaniami klasyfikacji bibliotecznych występuje ujemne sprzężenie zwrotne: rozwój jednych

---

[6] B. Hjørland: *What is Knowledge Organization (KO)?* W: *Knowledge Organization* 2008, nr 35(4), s. 86-101; Sh. Koshman. *Visualization-based information retrieval on the Web*. Library and Information Science Research 2006, Vol. 28, nr 2, s. 192-207.

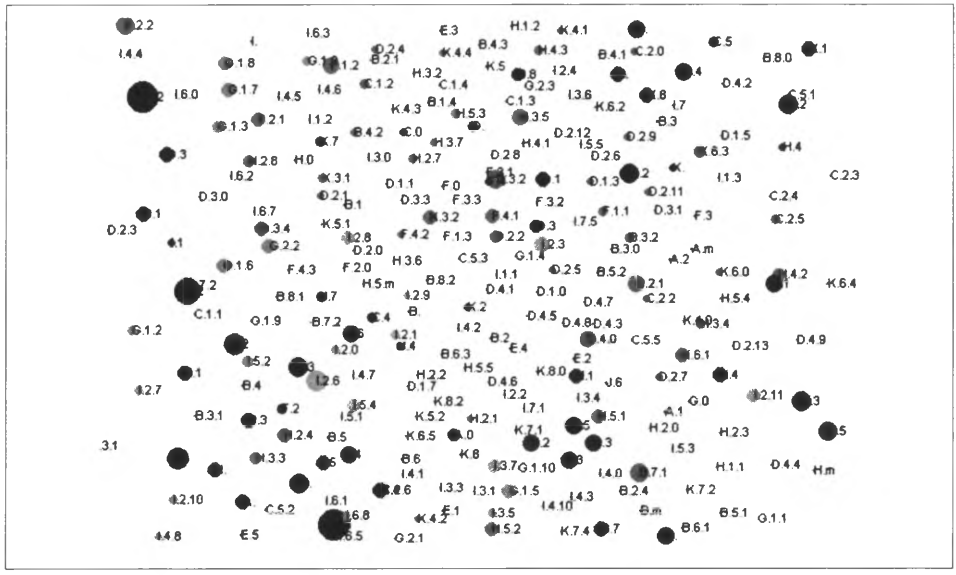
[7] V. Osińska: *Przybliżenie semantyczne w wizualizacji informacji w Internecie i bibliotekach cyfrowych*...

[8] B. Hjørland: *What is Knowledge Organization*..., s. 89-91.

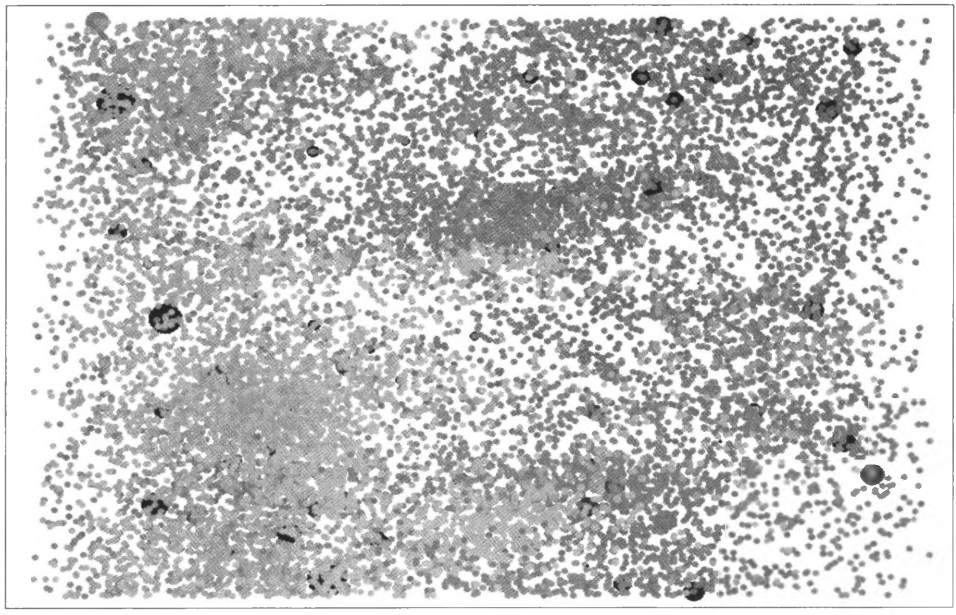
[9] S.E. Robertson: *The Probability Ranking Principle in IR*. Journal of Documentation 1977, Vol. 33, s. 294-304.

[10] Tamże.

[11] Ł. Neuman: *Transformacja relacyjnych baz danych do postaci sieci Bayesa* [on-line]. Politechnika Wrocławska. Zakład Systemów Informacyjnych [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.zsi.pwr.wroc.pl/ zsi/missi2002/pdf/s210.pdf.

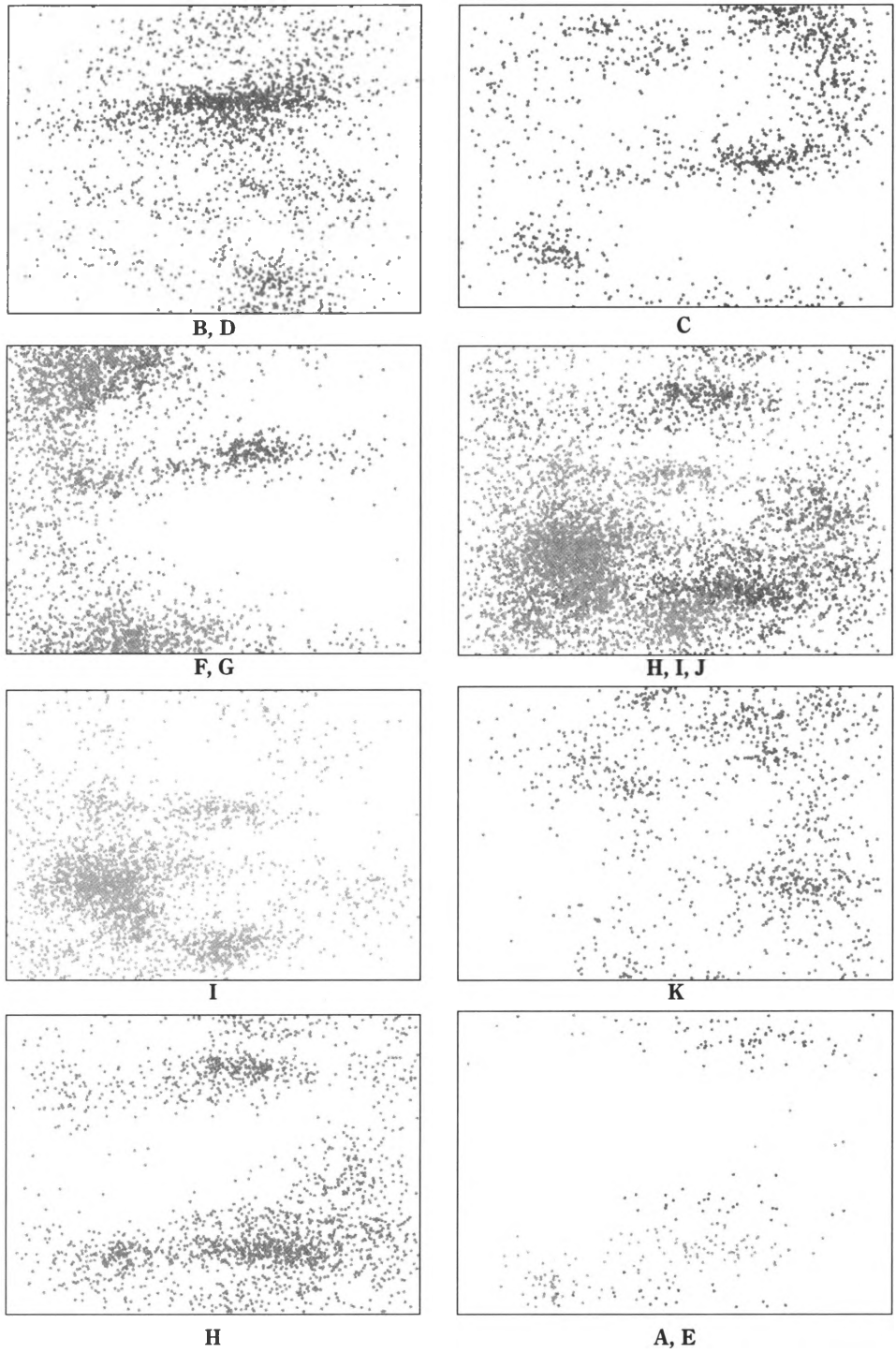


Klasy



Dokumenty

Ilustracja 16. Mapy klasyfikacji CCS z 1998 r.



Ilustracja 17. Mapy klasyfikacji CCS z 1998 r. dla wybranych kombinacji klas.

powoduje zanik drugich[12]. Eksperymenty w Cranfield, zainicjowane w latach sześćdziesiątych ubiegłego wieku, miały na celu znalezienie sposobów zwiększenia wydajności systemów wyszukiwawczych (SW) poprzez usprawnienie języków informacyjno-wyszukiwawczych (JIW)[13]. Wtedy właśnie wprowadzono aktualne do dzisiaj miary relewancji wyjściowych danych: kompletność i dokładność.

Dokładność (ang. *precision*) P jest miarą wydajności systemów wyszukiwawczych określoną jako ułamek znalezionych relewantnych dokumentów w kolekcji do pełnej liczby znalezionych dokumentów:

$$P = \frac{\left|\{dokumenty\_relewantne\} \cap \{dokumenty\_znalezione\}\right|}{\left|\{dokumenty\_przeszukane\}\right|} \tag{9}$$

Dokumenty przeszukane w danej transzy składają się z relewantnych i nie relewantnych. Te ostanie często nazywa się szumem.

Kompletność K (ang. *recall*) definiowana jest za pomocą części znalezionych relewantnych dokumentów w stosunku do wszystkich relewantnych dokumentów w zbiorze:

$$K = \frac{\left|\{dokumenty\_relewantne\} \cap \{dokumenty\_znalezione\}\right|}{\left|\{dokumenty\_relewantne\}\right|} \tag{10}$$

Ta miara wskazuje w jakim stopniu proces wyszukiwania jest wyczerpujący w zbiorze wszystkich dokumentów. Dosłownym tłumaczeniem terminu angielskiego jest „odzew"; wysoki odzew świadczy o tym, że prawie wszystkie obiekty szukane, które interesują użytkownika zostały zwrócone przez system wyszukiwawczy. Testy w Cranfield[14] oraz późniejsze prowadzone dla różnych wyszukiwarek internetowych wykazały, że systemy klasyfikacji bibliotecznych i fasetowo-analitycznych były mniej wydajne od systemów wyszukiwania tekstowego. Niekwestionowana skuteczność i popularność wyszukiwarki *Google* tylko przypieczętowały wyraźną w ostatnich latach przewagę wyszukiwania informacji, dominującą nad rozwojem systemów klasyfikacyjnych wraz z wbudowanym w nie wyszukiwaniem katalogowym.

Nowoczesna teoria wyszukiwania informacji wiąże się z konfliktem paradygmatów: fizycznego, czyli wyznaczonego systemem wyszukiwawczym oraz kognitywnego, który jest zorientowany na użytkownika[15]. Aby sprostać zadaniu stworzenia doskonałego systemu wyszukiwawczego, badacze przez lata koncentrowali się na polepszeniu algorytmów reprezentacji dokumentów i formułowania zapytań. Takie podejście, zorientowane jedynie na „potrzeby maszyny" ma tendencję ignorowania kognitywnego zachowania użytkownika. Natomiast nowoczesne systemy wizualizacji informacji, a wiele przykładów można zobaczyć w przeglądarkach nowej generacji a także projektach komercyjnych (p. Rozdział 1.3), bazują na filozofii projektowania zorientowanego na użytkownika. Potrzeby, wymagania i ogranicze-

[12] B. Hjørland: *What is Knowledge Organization...*, s. 88-90.

[13] *Cranfield Demo* [on-line]. University of Twente, Netherlands. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://dbappl.cs.utwente.nl/pftijah/Documentation/CranfieldDemo.

[14] Tamże.

[15] B. Hjørland: *What is Knowledge Organization...*, s. 88-90; P. Morville. *Ambient Findability*. Sebastopol, USA: O'Reilly Media Inc., 2005, s. 48-51.

nia końcowego użytkownika są szczegółowo badane na każdym etapie procesu projektowego. Interfejsy takich systemów są konstruowane tak, aby ułatwić zadania wyszukiwania informacji, jak również manipulowania wynikami i oszacowania ich stopnia relewantności. Dodatkowo, takie aplikacje niosą ważny pierwiastek interaktywności, kiedy to użytkownik ma możliwość zarządzania procesem wyszukiwania. Aktualne trendy społecznościowe Web 2.0 włączając: wspólne tagowanie (folksonomie), edycję treści wspólnych zasobów (wiki), P2P, blogi, kanały *RSS*, podcasty[16] i webcasty[17] – stwarzają nowe wyzwania systemom wyszukiwawczym, opartych na technikach wizualizacyjnych[18].

Wykorzystanie wizualizacji jako graficznego interfejsu do wyszukiwania informacji jest praktykowane od wczesnych lat 90-tych. Jako pierwsza przetestowana została metoda map samorganizujących się (*SOM*) do wyszukiwania dokumentów on-line na podstawie reprezentacji semantycznych relacji pomiędzy nimi[19]. Autorzy[20] porównali wyniki wyszukiwania dokumentów na mapach wizualizacji przy zastosowaniu dwóch algorytmów: *SOM* oraz *MDS*. Podstawowym celem takich prac we wczesnym okresie było skonstruowanie abstrakcyjnej przestrzeni informacyjnej, służącej jako interfejs systemu wyszukiwawczego. Te systemy wyszukiwawcze charakteryzowały się nową strategią formułowania zapytań przez użytkownika. Standardowe systemy są oparte na tekście, systemy nowszej generacji, a szczególnie ciekawe przykłady można zobaczyć w sieci, jak np. wyszukiwarki wizualizacyjne *Kartoo*, *Grokker* (p. Rozdział 1.3) w komunikacji z użytkownikiem posługują się reprezentacją dokumentów za pomocą ikon, symboli, glyfów i innych metafor.

Sh. Koshman podkreśla, iż nowoczesne sieciowe SW muszą umożliwiać użytkownikom poza manipulowaniem wynikami, także zarządzanie całym procesem wyszukiwawczym[21]. U podstaw wizualnego wyszukiwania informacji leży wiedza o funkcjonowaniu ludzkiego systemu percepcyjnego, którego założenia przedstawione są w rozdziale 1.1. Prawa Gestalt'a[22] pomagają w zrozumieniu procesów wizualnego przetwarzania informacji; definiując zasady: bliskości (ang. *proximity*), domknięcia (ang. *closure*) i ciągłości (ang. *continuity*). Odbiorcy łatwiej dostrzegają obiekty zgrupowane, skupione wokół siebie, niż odseparowane – na tym polega zasada bliskości (Rysunek 36a).

[16] Są to formy internetowej publikacji multimedialnej, najczęściej w postaci regularnych odcinków, z zastosowaniem technologii RSS.

[17] Prezentacje multimedialne (jw.), przesyłane za pomocą mediów strumieniowych.

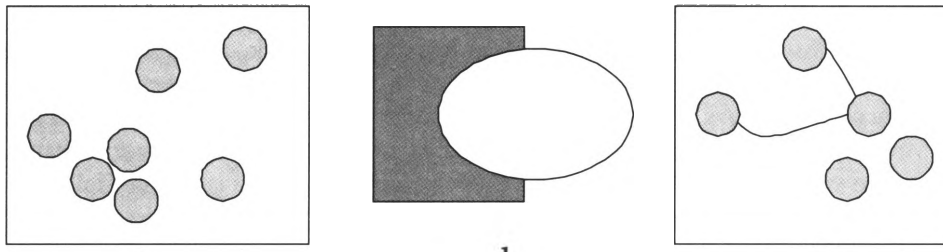
[18] Sh. Koshman, dz. cyt., s. 192-207.

[19] X. Lin, D. Soergel, G. Marchionini: *A Self-organizing semantic maps* as graphical interfaces for *information retrieval*. W: *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in Information Retrieval*. College Park, USA: University of Maryland, 1991, s. 262-269.

[20] H.D. White, X. Lin, K.W. McCain, dz. cyt.

[21] Sh. Koshman, dz. cyt.

[22] Tamże, s. 193-194.



Rysunek 36. Zasady projektowania interfejsów graficznych systemó wyszukiwawczych: bliskość; b) domknięcie; c) ciągłość.

Źródło: Na podst. Sh. Koshman. *Visualization-based information retrieval on the Web*. Library and Information Science Research 2006. Vol. 28 nr 2, s. 193-194.

Kolejny rysunek – 36b jest przykładem domknięcia i obrazuje, jak w naszym mózgu mimowolnie uzupełniana zostaje informacja o domknięciu prostokąta, mimo że jego struktura jest niekompletna. Ciągłość wskazuje na zdolność naszego mózgu do organizacji wizualnych elementów za pomocą ich właściwości wiązań. Wizualnie łączymy elementy dalej zlokalizowane, lecz powiązane ze sobą za pomocą krzywych (Rysunek 36c). Te trzy wymienione zasady leżą u podstaw projektowania interfejsów wizualnego wyszukiwania informacji.

Z punktu widzenia ewaluacji wyszukiwania informacji w systemach informacyjnych, wydajność otrzymanej w trakcie naszego eksperymentu wizualizacji można ocenić za pomocą skuteczności wyszukiwania podobnych w treści publikacji. Idea wizualizacji wykorzystuje regułę, iż macierz podobieństwa klas i podklas znalazła swoje odbicie w bliskości położenia reprezentujących ich węzłów na mapie lub powierzchni sfery. Taką samą zasadę można przyjąć także dla dokumentów w kolekcji, ponieważ ich współrzędne obliczano na podstawie współrzędnych klas i podklas, do których zostały zaklasyfikowane. Podklasy klasyfikacji *CCS* determinowały tematy badawcze prac naukowych. A zatem w przestrzeni informacyjnej artykuły o podobnej tematyce powinny znajdować się blisko siebie. Test sprawdzający tą hipotezę polega na pierwotnym wytypowaniu dokumentu wzorcowego, określenia lokalizacji jego węzła na mapie, a następnie selekcji jego najbliższych oraz dalszych sąsiadów, na końcu – scharakteryzowania wyszukanych obiektów pod względem treści. Wykorzystując formułę (9) została obliczona precyzja wyszukiwania relewantnych artykułów dla poszczególnych testów. Topologia sąsiedztwa jest określona empirycznie, co prowadzi do załamania symetrii kierunkowej w przypadku, kiedy przestrzeń trafnych danych będzie rozciągnięta w kierunkach dominujących podklas. W takim przypadku może się okazać, iż potrzebna będzie modyfikacja powierzchni sfery, aby wyróżnić topologicznie odwzorowane obszary zwiększonej relewantności.

Jak wykazały poprzednie badania słów kluczowych, najbardziej zachęcające wyniki otrzymano w obszarach o dużym zagęszczeniu dokumentów. Klastry te wskazywały na klasy główne *CCS* w sposób jednoznaczny, co wywnioskowano z jednolitego ich zabarwienia. Testy wyszukiwania podobnych artykułów przeprowadzono także w miejscach nakładania się pól, aby sprawdzić spójność tematów. Wybrano charakterystyczny wspólny obszar dla węzłów klas H. *Information Systems*, I. *Computing*

Methodologies i K. *Milleux*. Do identyfikacji węzłów na mapie posłużono się numerami rekordów w bazie. Z powiększonego fragmentu mapy wytypowano węzeł wzorcowy (dokument nr 1622) i wybrano jego najbliższych sąsiadów – zaznaczonych na ilustracji pod Tabelą B1. Na sąsiadów wybierano nie tylko obiekty tej samej klasy głównej (w tym samym kolorze), lecz również najbliżej położone węzły z innych klas. W ten sposób można było wykryć ewentualne związki semantyczne pomiędzy dokumentami przypisanych do różnych podklas.

Artykuł nr 1622 nosi tytuł: *Combining bibliometrics, information retrieval, and relevance theory. Some implications for information science*. Wyłącznie na jego podstawie nie da się kompletnie scharakteryzować obszernej tematyki dokumentu, która jeszcze poza głównym tematem omawia zagadnienia infometrii, psychometrii i modeli kognitywnych. Analogicznie niewygodna sytuacja powstaje kiedy trzeba określić podobieństwo wybranych na mapie artykułów. W konsekwencji analizę artykułów rozszerzono o przegląd słów kluczowych, deskryptorów tematycznych oraz abstraktów. W Tabeli B1 wymienione są merytoryczne właściwości zbadanych dokumentów: tytuł, autor(zy), symbole przyporządkowanych klas, słowa kluczowe i terminy główne. Zbliżone trzy- i czterocyfrowe numery dokumentów wskazują na bliską lokalizację tych rekordów w oryginalnej bazie danych *ACM* w procesach indeksowania i sortowania. Pytanie czy niska rozbieżność identyfikatorów rekordów ma jakieś odwzorowanie w podobieństwie rzeczowym dokumentów jest drugorzędnym i nadaje się do osobnego zbadania problemu. Dokładnie przeglądając poszczególne pozycje, widać jak mylący lub nieznaczący może być tytuł, w którym często używa się nazw własnych technologii, programów oraz nazw skrótowych narzędzi. Większa wartość informacji użytecznej jest zawarta w słowach kluczowych i symbolach klasy głównej. Zapoznanie się ze streszczeniem publikacji z numerem 1622 (wzorzec) pozwoliło na konkluzję, iż głównymi tematami pracy są: wyszukiwanie informacji przy równoległym rozwiązywaniu problemów formułowania zapytań, jak również efekty kognitywne, wywoływane w procesie interpretacji relewantnych odpowiedzi przez użytkownika. Dane bibliometryczne natomiast służą jedynie jako środek badawczy i nie należy bazować na wyrazie *bibliometric* w oszacowaniu wyników. Z prezentowanej listy wykluczono również pozycje *Book review*, jako nie prezentującej konkretnej jednostki tematycznej. Otrzymano precyzję wyszukiwania 61%, która wzrosła po zastosowaniu korekty danych wejściowych. Zdecydowano się wyeliminować ze przeszukiwanego zbioru dokumenty o numerach 2310 i 2317 (w kolorze turkusowym), jako zlokalizowanych w dość dalekiej odległości od „wysepki" testowanej próbki danych (p. ilustrację pod Tabelą B1) i wówczs precyzja wyniosła 68,7%. Zauważono, iż największym podobieństwem charakteryzują się artykuły z pierwszej połowy listy – należą one do tej samej podklasy – H.3.3 w klasyfikacji podstawowej. Wywnioskowano zatem, jeśli mapując obiekty, zwiększymy wagę klasyfikacji podstawowej, to tym samym przyciągniemy do siebie węzły dokumentów o maksymalnie pokrewnej tematyce badań.

Następne trzy testy przeprowadzono dla wag 0.7:0.3 klasyfikacji głównej i klasyfikacji dodatkowych odpowiednio. W pierwszym powtórzono wyszukiwanie tematycznie bliskich dokumentu z numerem rekordu w bazie 1622. Wyniki oraz fragment mapy wizualizacji są prezentowane w Tabelach B2-B4. Przygotowując zbiór

danych do końcowych obliczeń, pozbyto się w nim pozycji 1895 (*Recomended Systems – Proceedings of the 40th Annual Hawaii International Conference on System Sciences*), jako elementu o niesprecyzowanej tematyce, oraz pozycji 408 ze względu na odległość. Tak jak oczekiwano trafność wyszukiwania się zwiększyła o ponad 10% (Tabela 7). Tematyka badawcza przeszukiwanych artykułów rozszerzyła się o metodo-logie wyszukiwania semantycznego (nr 3151 w Tabeli B2).

Tabela 7.

Uzyskane charakterystyki wyszukiwania dokumentów dla różnych map klasyfikacji podstawowej i dodatkowych

Parametr	Wagi			
	0.6:0.4	0.7:0.3	0.7:0.3	0.7:0.3
Nr dokumentu wzorcowego	1622	1622	882	719
Klas. Podstawowa (KP)	H.3.3	H.3.3	C.2.1	D.2.11
Kategoria Główna KP	Systemy informacyjne	Systemy informacyjne	Systemy komputerowe	Software
Dokładność *P*(%)	**68.7**	**82.6**	**71.7**	**86.8**
Il. dok. w próbce	21	15	13	18
Il. dok. wyeliminowanych	2	2	3	0
Pole obszaru na mapie (j.u.)	0.211	0.272	0.160	0.420

Źródło: Opracowanie własne.

W następnym teście zbadano artykuły sąsiednie węzła nr 882, który prezentował dokument pod tytułem *Communication over Hypercomplex Kähler Manifolds: Capacity of Multidimensional-MIMO Channels*. Pomijając nazwę technologii *MIMO*[23], zwiększającej przepustowość sieci bezprzewodowej, można skupić na identyfikacji artykułów nawiązujących do komunikacji bezprzewodowej. Niższy wynik precyzji, niż w poprzednim przypadku (p. Tabela 7) można wytłumaczyć obecnością czasopism na liście zamiast pojedynczych prac o konkretnej tematyce.

Ostatni test był najważniejszy z perspektywy badań krzyżowania się gałęzi drzewa klasyfikacyjnego. Wybrano taki obszar na mapie wizualizacji, gdzie występowało duże mieszanie się kolorów węzłów dokumentów przypisanych do różnych klas w klasyfikacji podstawowej. Autorzy pracy pod tytułem: *CCLRC Portal to support research facilities: Research Articles* skupili się nad badaniem specyfikacji i architektury portalu *CCLRC*[24]. Z cech funkcjonalności wymieniono między innymi zastosowania gridowe, podając ten wyraz również jako słowo kluczowe. W wyniku wybrania sąsiednich dokumentów na mapie otrzymano dużą część publikacji z odwołaniem do tematu tematu zastosowania gridowe. Precyzja wyniosła ponad 86%. Żadnej pozycji nie odrzucono: wybrane dokumenty na mapie zlokalizowane były w zwartym obszarze. Dla pełniejszego opisu niektórych prac, w tabeli zamieszczono streszczenie (kolor zielony w Tabeli B4). Nieco niejasna jest rola artykułu 1975: *Towards the theoretical foundation of choreography*. Z abstraktu wynika, iż autorzy zbadali możli-

[23] MIMO (ang. *Multiple Input, Multiple Output*) – nowoczesna technologia, stosowana w systemach radiowych.

[24] *CCLRC* (ang. *Council for the Central Laboratory of the Research Councils*) – jednostka rządowa Wielkiej Brytanii do spraw badań naukowych.

wości implementacji „prostego" języka choreografii włączając semantykę w teorii formalnej, która rozciąga się na kategorie klas głównych B. *Hardware*, D. *Software* oraz F. *Theory of Computation*. Przez to zaklasyfikowanie tej teoretycznej pracy w pierwotnym podejściu do klasy H. *Information Systems* potraktujemy jako pomyłkę autorów. Takie błędy niestety zawsze występują jako efekt działania czynnika ludzkiego. Z pobieżnej inspekcji listy tematów wynika, iż stanowią one niewielki procent w bazie danych i nieznacznie zaniżają precyzję wyszukiwania podobnych dokumentów. Do testów w trybie półautomatycznym wybierano niewielkie próbki dokumentów – rzędu kilkunastu, co przekładało się na małe obszary na mapie wizualizacji. Rzeczą oczywista jest, że należy powtórzyć eksperyment dla większych fragmentów mapy, na przykład kilku sąsiadujących klastrów. Można się spodziewać obniżenia precyzji przy znacznym wzroście liczby badanych dokumentów, lecz przy zastosowaniu mechanizmów różnicowania poziomu podobieństwa ogólna wydajność wyszukiwania może zostać zachowana. Pewnikiem jest, że wysoka precyzja systemów wyszukiwania świadczy o relewantności dla użytkownika dużej części otrzymanych artykułów. Natomiast wysoki odzew (kompletność) wskazuje na to, że prawie wszystkie relewantne dokumenty zostały odnalezione. Ponieważ zmapowane klastry zrzeszają artykuły z tej samej klasy głównej, to wewnątrz lub w pobliżu tych obszarów kompletność wyszukania tematycznie podobnych prac będzie największa. Pole i lokalizacja badanego fragmentu mapy determinuje wartość tego parametru.

Bardziej treściwe, niż tytuły i/lub słowa kluczowe są abstrakty, gdzie można zastosować na przykład algorytmy zliczania częstości występowania form wyrazowych, wykrycia konstrukcji składniowych i inne techniki lingwistyki komputerowej. Otrzymane wyniki w takich eksperymentach, jak na przykład w pracy[25] o wyszukiwaniu podobnych dokumentów korzystając z korpusów słownikowych ich abstraktów dawały precyzję nie przekraczającą 70%. Jak przedstawiono powyżej zaproponowane metody wyszukiwania dokumentów, oparte o metody wizualizacyjne pozwoliły na zwiększenie precyzji.

Zaprojektowana na potrzeby wizualizacji baza danych nie pozwala na korzystanie z dodatkowych atrybutów bibliometrycznych, takich jak abstrakt, tekst lub bibliografię, które znacząco mogłyby podnieść precyzję wyszukiwania. Dlatego jako uzupełnienie badań nad implementacją rozkładu wizualizacji w wyszukiwaniu artykułów wybrano metodę odwrotną. Jeśli w pierwszym etapie sprawdzono stopień podobieństwa tematów wybranych na mapie dokumentów, to następnym krokiem było zlokalizowanie na mapie wyników wyszukiwania specjalistycznego systemu wyszukiwawczego.

Kilka miesięcy po skompletowaniu danych publikacji z 2007 r., autorzy serwisu biblioteki cyfrowej *ACM* zaczęli go intensywnie rozwijać. Dużym udogodnieniem dla użytkowników stała się możliwość sortowania wyników według daty publikacji, tytułu, liczby cytowań, stopnia relewantności itp. Funkcje wyszukiwania zostały udoskonalone w wyniku zaimplementowania nowych filtrów na rozbudowanych

[25] B. Zyglarski, T. Schreiber, P. Bała: *Web Services Based Scientific Article Manager*. W: *Information Systems Architecture and Technology. Web Information Systems: Models, Concepts and Challenges*. Wrocław: Wydawnictwo Politechniki Wrocławskiej, 2008, s. 205-216.

danych bibliometrycznych, na przykład wybór dokumentów związanych z poszczególnymi nazwiskami wymienionymi w publikacji, kwerendy współpracowników autorów. Dostępne jest również wyszukiwanie zaawansowane, z którego skorzystano w dalszych testach. Głównymi motywacjami były wygoda użytkowania, multiplikatywnie komplementarnej bazy danych oraz dostępność profesjonalnych algorytmów (niejawnych dla użytkowników) wyszukiwania w serwisie *ACM*. Wyniki testów będą pozytywne jeśli odfiltrowane dokumenty o podobnej tematyce, lecz przypisane do różnych gałęzi drzewa klasyfikacji, ulokują się na mapie wizualizacji w bliskiej odległości od siebie.

W testach posłużono się metoda ręczną, ponieważ automatyczna nie nadawała się ze względu na upływ czasu (minęło 2 lata) i zmianę identyfikatorów dokumentów w bazie *ACM* w porównaniu do skompletowanej pierwotnie. Dla ułatwienia zadania w rozkładzie nieposortowanych wyników wybrano co drugi lub trzeci rekord. Zapytanie powinno było odwoływać się do specyficznego tematu badań, aby zawęzić zakres wyszukiwawczy. Z takich dostępnych pól przeszukiwania, jak tytuł, abstrakt, słowa kluczowe oraz przegląd/recenzja wybrano drugie, ponieważ z dużym prawdopodobieństwem w streszczeniu znajdą się i słowa kluczowe, i deskryptory tematyczne lub wyrazy semantycznie im bliskie. Tytuł w takiego typu publikacjach często jest zlepkiem nazw własnych albo nic nie znaczącym np. *Book Review*. (p. Tabela B1).

Pierwszym wyrażeniem testowym było *distance learning*, pojęcie występujące w drzewie klasyfikacyjnym na poziomie deskryptora tematycznego. Jest ono szersze, niż popularne w kręgach naukowych i oświatowych *e-learning*. Na podstawie zwróconych wyników dało się zaobserwować, że system wyszukiwawczy uwzględnił na równi z wytypowanym termin drugi (*e-learning*), jak również *on-line learning*. Uzyskano w odpowiedzi 65 rekordów, wizualizacji poddano 30 – mapa z zaznaczonymi artykułami jest przedstawiona na Ilustracji 18. Zanotowano wyraźne w skali całej mapy skupienie obiektów wokół siebie. Analiza dokumentów wskazała, iż najczęściej powtarzającą się w wynikowym zbiorze klasą główną była **K** – punkty w kolorze liliowym. Odnosi się ona do tematyki środowiska komputerowego, czyli nie jest ściśle określona. Tym bardziej zaobserwowana koncentracja obiektów świadczy na korzyść proponowanej metody wyszukiwania podobnych artykułów.

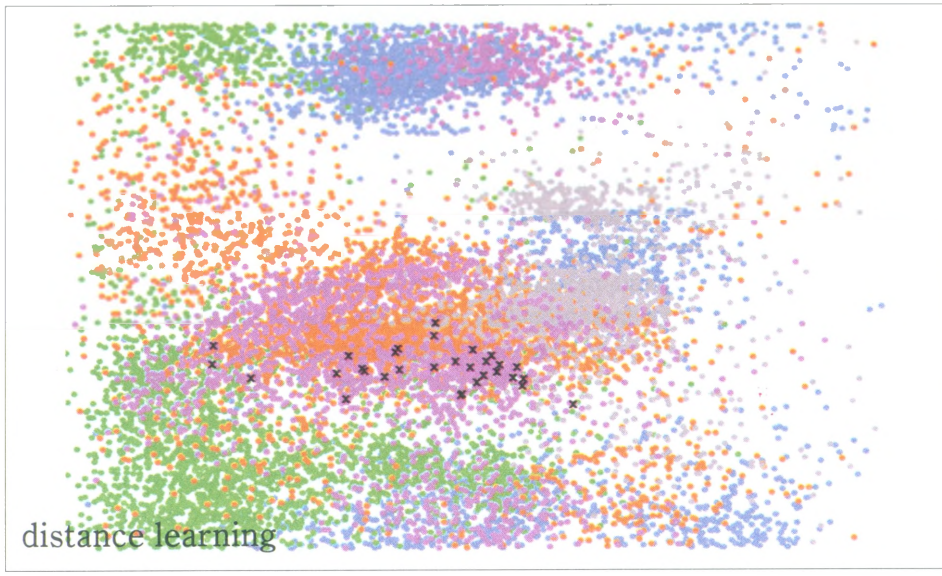
Następnie wprowadzono wyrażenie *grid computing*[26] jako frazę, aby uniknąć prac, gdzie siatka (wyraz *grid*) występuje nie w roli technologii komputerowej, lecz narzędzia mierniczego np. w kartografii. Otrzymano w odpowiedzi 105 pozycji. W wy-niku wizualizacji (Ilustracja 18) zauważamy, iż wykształciły się trzy skupiska położone blisko siebie (ciągłość powierzchni sfery wyznacza, iż górny fragment trzeba rozpatrywać jako kontynuacja dolnego). Wyłowione artykuły z reguły odnosiły się do klas **C**, **D** i **H** w klasyfikacji głównej. Ponieważ klasa **C** – zajmuje się generalnie problematyką sieci komputerowych, a węzły artykułów do niej należące mieszczą się na biegunach sfery, to tłumaczy dlaczego jeden z klastrów jest odosobniony w tym obszarze mapy i rozciągnięty. Drugie ognisko zlokalizowane jest

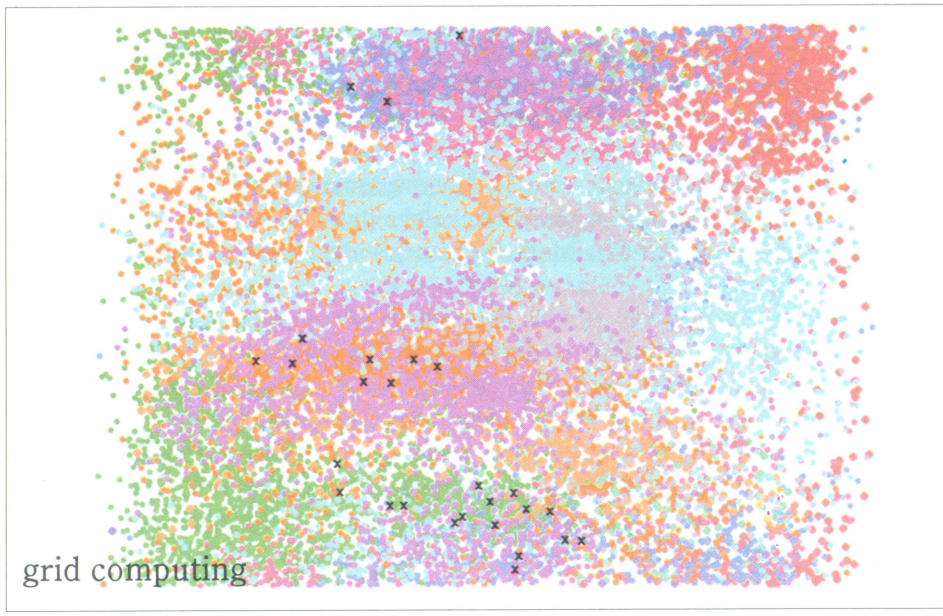
[26] Można to określić jako zastosowanie kilku komputerów do pojedynczego problemu w tym samym czasie. Zwykle to są naukowe lub techniczne zagadnienia, które wymagają dużej liczby cykli procesora aby przetworzyć ogromne ilości danych.

w miejscu, gdzie klastry brązowy (*Information Systems*) i liliowy (*Milleux*) zachodzą na siebie. Wynika stąd, iż publikacje o tematyce obliczeń gridowych poruszają zarówno stronę programowo-techniczną takich systemów, tak i organizacyjno-informacyjną przy równoległym traktowaniu ich jako usługi użyteczności publicznej.

W pierwszej fazie testów przeanalizowano właściwości wytypowanych na mapie węzłów dokumentów sąsiadujących z wzorcowym. Druga faza opierała się o wskazane przez system wyszukiwawczy *ACM* relewantne dokumenty, które następnie zlokalizowano na mapie. Wysoka precyzja w pierwszym przypadku oraz bliska lokalizacja – w drugim dowodzą pewności zastosowania proponowanej metody wizualizacji w zadaniu wyszukiwania informacji. Równoległe występowanie kategorii tematycznych z różnych klas głównych w wynikach pierwszego etapu, oraz skupiska obiektów w miejscach mieszania się kolorów na mapie – drugiego etapu świadczą o tym, iż strategia wyszukiwawcza została oddzielona od schematu klasyfikacji pierwotnej. W ten sposób możliwe jest wynajdowanie publikacji naukowych, należących do różnych gałęzi drzewa klasyfikacyjnego, lecz o zbliżonej tematyce. Co prawda za pomocą dobrze sformułowanych słów kluczowych można również efektywnie wyszukiwać dokumenty w kolekcji. Jednak nie każdy artykuł jest wyposażony w ten atrybut. Poza tym żadna sekwencja wyrazów nie dokona kompletnego semantycznego opisu tematu badań pracy. Wybór słowa z kontekstu reszty sekwencji pociąga za sobą obniżenie precyzji wyszukiwania tematycznie podobnych artykułów.

Obiecujących wyników natomiast można spodziewać się w przypadku posiłkowania się uprzednio skonstruowaną mapą słów kluczowych (p. Rozdział 3.3.b), gdzie statystycznie zachowane zostały zestawy słów kluczowych. Zaznaczone obszary trafności (lub relewancji) wyrazów kluczowych na mapie (Ilustracja 10) mogą posłużyć jako pole danych do uruchomienia silnika wyszukiwawczego. Takie czynności zwiększają nie tylko precyzję wyszukiwania, lecz przede wszystkim kompletność, ponieważ badane miejsce ma największy potencjał relewantności.





Ilustracja 18. Identyfikacja wyszukanych obiektów na mapie wizualizacji.

Podsumowanie i wnioski

Konieczność badań nad reprezentacją graficzną zasobów cyfrowych jest bardzo ważna, co potwierdza powstanie w ostatnich latach wielu nowatorskich technik wizualizacji informacji.

Nowa metoda wizualizacji informacji

W niniejszej pracy przedstawiona została nowa metoda wizualizacji danych, pochodzących z biblioteki cyfrowej *ACM* i zorganizowanych według klasyfikacji nauk komputerowych – *Computing Classification System*. Dane stanowiła kolekcja artykułów o tematyce informatycznej, opublikowanych w 2007 roku. Zgodnie z uzasadnioną i opisaną koncepcją stworzono model sferyczny przestrzeni informacyjnej drzewa klasyfikacyjnego. Zamiast liniowych dendrogramów z ograniczonymi możliwościami przedstawienia relacji i podobieństw klas, skonstruowano zakrzywioną powierzchnię, która zapewniła nie tylko pełniejszą analizę danych i ich właściwości, lecz również odzwierciedlenie struktury hierarchicznej za pomocą nowoczesnych technik wizualizacyjnych. Należy tu nadmienić, iż dotychczas nikt jeszcze nie podjął się przedstawienia schematu klasyfikacji oraz jej uniwersum w podobny sposób. Zmapowanie dokumentów na sferę, jak się spodziewano, stało się przysłowiową „kopalnią wiedzy" o schemacie badanej klasyfikacji, jej rozwoju, a także samej dziedzinie nauk komputerowych.

Badania zostały przeprowadzone w bardzo szerokim zasięgu, włączając w kolejnych etapach zestawy przetworzonych danych. W związku z tym należy wspomnieć o organizacji całego eksperymentu, który stał się immanentną częścią stosowej metody. Składał się on z kilku etapów:

- kolekcjonowanie metadanych;
- ekstrakcji danych;
- wizualizacji danych;
- analizy danych za pomocą:
  - rzutów kartograficznych;
  - modyfikacji zestawu danych;
  - obliczeń wymiaru fraktalnego dla reprezentacji graficznych;
- przetestowaniu możliwości zastosowań, a mianowicie:
  - w mapowaniu semantycznym słów kluczowych;
  - w modernizacji i ewaluacji istniejącego schematu klasyfikacji;

- badaniu ewolucji schematu klasyfikacji;
- badaniu rozwoju dziedziny nauk komputerowych;
- wyszukiwaniu dokumentów o podobnej tematyce.

Jedno już spojrzenie na otrzymane mapy graficzne wystarczyło, aby zauważyć słuszność postulowanych założeń nowej metody. Uzyskane metody wizualizacji danych pozwoliły na łatwiejszą i bardziej intuicyjną interpretacje. Dodatkowo uzyskano szereg nowych możliwości wynikających bezpośrednio ze sposobu prezentacji danych. W pierwotnym założeniu eksperymentu zamierzano dokonać tylko wizualizacji klas i dokumentów oraz powtórzenia sekwencji badań cyklach 10-letnich.

Wizualizacja jako interfejs Information Retrieval

Eksploracja i wyszukiwanie relewantnych dokumentów na powierzchni sfery było głównym zadaniem realizowanym w niniejszej pracy. W trakcie obliczeń dokonano kilku istotnych czynności metodologicznych m.in.: sprawdzono właściwości dokumentów, które znajdowały się w bliskim sąsiedztwie topologicznym z wzorcowym obiektem oraz zlokalizowano na mapie wytypowane przez wyszukiwarkę *ACM* relewantne dokumenty. Otrzymano wysoką precyzję w pierwszym przypadku (80%-90%) oraz bliską lokalizację – w drugim. Dowodzi to poprawności proponowanej metody wizualizacji i jej skuteczności w wyszukiwaniu informacji. W ten sposób możliwe jest znajdowanie publikacji naukowych o zbliżonej tematyce, lecz należących do różnych gałęzi drzewa klasyfikacyjnego i w związku z tym trudnych do znalezienia metodami tradycyjnymi.

Idealna wyszukiwarka charakteryzuje się zarówno wysoką dokładnością, jak i tzw. odzewem (kompletnością). Współczesne wyszukiwarki internetowe mają wysoki odzew kosztem niskiej precyzji, albo odwrotnie. Kompletność prawniczych systemów wyszukiwawczych (np. LEX[1]) wynosi jedynie 20%. W profesjonalnych systemach wyszukujących np. *STAIRS* (*Storage and Information Retrieval System*), zaprojektowanym przez *IBM* osiągnięto wartości 75-80%[2]. Przedmiotowe wyszukiwarki biblioteczne stosowane w katalogach *OPAC* mają wysoką precyzję dzięki zastosowaniu słownictwa kontrolowanego[3]. Jednak tym samym mają niski „odzew", gdyż nie wykorzystują relacji hierarchicznych[4]. Zaproponowana metoda wizualizacji pozwala na uniknięcie wymienionych niedogodności.

[1] LEX – jeden z największych i najbardziej znanych polskich systemów informacji prawnej, istniejący od 1987 roku. Aktualnie wydawany jest przez koncern *Wolters Kluwer Polska*.

[2] P. Morville, dz. cyt., s.50.

[3] D. Dačko: *Zastosowanie ontologii do odkrywania wiedzy* [on-line]. *Consensual Knowledge* [Strona domowa D. Dačko] [Dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.consensualknowledge.net/files/semlib_thesis.pdf.

[4] LEP (ang. *Light Emitting Polymers*) – polimery emitujące światło pod wpływem przyłożonego do nich napięcia elektrycznego.

Dynamika zmian w klasyfikacji Computing Classification System

Aby wykryć istotne zmiany w światowej literaturze naukowej w zakresie informatyki badania nad wizualizacją dokumentów powtórzone zostały w skali czasu z interwałem co 10 lat. Na podstawie wyników oraz informacji na temat historii rozwoju nauk komputerowych przeprowadzono wnioskowanie o ewolucji ich przedmiotowej klasyfikacji, kierunkach integracji z innymi dyscyplinami oraz o trendach w przyszłości. Systemy informacyjne – to kategoria, która zapoczątkowała rozrost drzewa klasyfikcji *CCS*. Od 1988 r. systematyka takich kategorii jak *Software* (oprogramowanie), *Hardware* (sprzęt), *Computer Systems Organizations* (systemy Komputerowe) i Information Systems (systemy informacyjne) jest wyraźnie dostrzegalna. Kategorie *Software* oraz *Hardware* semantycznie wykazują dużą różnicę – potwierdzają to mapy wizualizacji z ostatnich dwóch dekad. Najważniejszym wnioskiem z prezentowanych zmian map klasyfikacji jest to, iż moment klasteryzacji przypada na lata dziewięćdziesiąte XX w. Mapy z dwóch ostatnich dekad świadczą o tym, iż proces zmian klasyfikacji rozwija się w jednym kierunku – coraz mocniejszej adaptacji struktury *CCS* w zbiorach biblioteki cyfrowej *ACM*. Z jednej strony – redaktorzy serwisu nanoszą ciągle poprawki w drzewie klasyfikacyjnym, tym samym dopasowując schemat do dynamicznych zmian w technologiach komputerowych. Z drugiej, autorzy prac naukowych coraz precyzyjniej potrafią je klasyfikować oraz szerzej stosować w opisach dane bibliometryczne. Ostateczna decyzja o kategoryzacji artykułów w drzewie klasyfikacji podstawowej i dodatkowych należy do redaktorów portalu. Słowa kluczowe natomiast wprowadzają wyłącznie autorzy publikacji. Mapa wizualizacji powstała dzięki takim krzyżującym się kategoriom, natomiast mapę semantyczną skonstruowano przy pomocy słów kluczowych. W wyniku uzyskano transformację wizualizacji tematycznej na mapę semantyczną. Na tych dwóch mapach skonfrontowane zostały dwie koncepcyjno-skojarzeniowe ścieżki, pochodzące z niezależnych źródeł.

Zmodernizowany schemat klasyfikacji CCS

Przeznaczenie takiej mapy semantycznej dostrzeżono w budowaniu nowego schematu klasyfikacyjnego, który jednak mieściłby się w granicach stałych struktur klas głównych klasyfikacji *CCS*. W związku z tym w ramach niniejszej pracy wyznaczono kolejny cel: na ile kategorie tematyczno-semantyczne pokrywają się w obu drzewach: oryginalnym *CCS* i wyznaczonym za pomocą zaproponowanych metod? Tabela A prezentuje wyniki szczegółowego zestawienia do poziomu drugiego. Na jej podstawie przeprowadzono ewaluację organizacji kategorii tematycznych oryginalnej klasyfikacji. Uzyskane wyniki pozwoliły na wywnioskowanie, iż system klasyfikacji *CCS* nie jest optymalny pod względem podziału logicznego na wyższych poziomach szczegółowości, jak również pod względem ilości stopni zagnieżdżenia. W oparciu o poklasteryzowane słowa kluczowe skonstruowano nowe drzewo klasyfikacji utrzymując podział klas głównych. Zmodernizowana struktura powstała poprzez zastosowanie tematyczno-semantycznego kryterium grupowania doku-

mentów. Zredukowanie poziomów hierarchii było podstawowym rezultatem transformacji pierwotnej przestrzeni klasyfikacyjnej na semantyczną.

Otrzymane wyniki pozwalają na stwierdzenie, że proponowana metoda wieloaspektowej organizacji zasobów może być wykorzystana zarówno do generowania, jak i automatycznej ewaluacji schematów klasyfikacyjnych.

Propozycje dalszych badań

Otrzymane wyniki wizualizacji są mocną zachętą do dalszych badań w kierunku wykrywania charakterystyk dynamiki klasyfikacji. W obecnym dziesięcioleciu doświadczamy, jak szybko następuje transformacja Internetu od postaci społecznej (Web 2.0) do nowej generacji – sieci semantycznej zaopatrzonej w technologie ontologiczne. Web 3.0 proponuje bardziej płynną organizację informacji, jak również nowe formy analizy i eksploracji danych, bazujące na inżynierii ontologicznej i/lub uczeniu maszynowym. Zmiany te powinny być także zauważalne w literaturze naukowej. A zatem w celu ich wykrycia, badanie kolejnych roczników publikacji *ACM* powinno być przeprowadzone dla krótszych cyklów. Duży potencjał badawczy przedstawionej metody wizualizacji z jednej strony oraz rozszerzalna struktura bazy danych o nowe metadane to powody, dla których planuje się dołączyć do analizy inne dane bibliometryczne (np. cytowania, autorów). Planuje się również rozszerzyć metodę od strony technicznej: do redukcji wymiaru danych zamierza się zastosować algorytmy *Self Orginizing Map*.

Przeprowadzone badania mogą być punktem wyjścia do projektowania systemów wyszukiwawczych opartych na wizualizacji danych. Zaprojektowany system powinień dostarczyć użytkownikowi alternatywny sposób wyszukiwania informacji przy wykorzystaniu topologii mapy. Precyzja trafności wyszukiwania będzie uwarunkowana rozmiarami wykształconych z odfiltrowanych obiektów pól. W różnorodności istniejących metod wizualizacja generalnie służy do prezentacji wyników wyszukiwania o interaktywnie zmienianych parametrach. Dla proponowanego systemu łączy ona funkcje bezpośredniej bazy wyszukiwawczej i filtra oraz interpretacji zapytania.

Najbardziej obiecujące, z metodologicznej perspektywy, byłoby zastosowanie zaproponowanego modelu wizualizacji do badania zasobów należących do różnych domen naukowych. Interpretacja organizacji wiedzy w nim byłaby uwarunkowana zarówno poziomem erudycji, rozległym horyzontem poznawczym, jak również intuicją naukowca. Dlatego w celu przeprowadzenia takich badań należałoby stworzyć interdyscyplinarny zespół naukowców a także uzyskać szerszy dostęp do bibliotek cyfrowych.

W przedstawionej metodzie kryją się ogromne możliwości naukometryczno-taksonomiczne w epoce, kiedy nauka zmierza w kierunku coraz wyraźniejszej interdyscyplinarności (nauk ścisłych i humanistycznych, które konsekwentnie zmierzają w kierunku integracji).

Rozwijając wątek przyszłych zastosowań, należy zwrócić szczególną uwagę na wybraną w wizualizacji – topologię sfery. Temat wizualizacji bezpośrednio dotyczy

warunków prezentacji obrazu. Monitory *LCD* dzisiaj nie są już nowatorską technologią. W chwili obecnej wchodzą na rynek ekrany polimerowe (*LEP*[5]) które w porównaniu z poprzednią generacją są cieńsze, lżejsze, bardziej energooszczędne, wytrzymałe mechanicznie oraz tańsze w produkcji. Mają one jednakże tę najważniejszą zaletę, że są elastyczne. Zupełnie naturalnie podchodzimy do informacji (między innymi inspirację czerpiąc z literatury i filmów *SF*), iż dominującą technologią wyświetlaczy 21-go tysiąclecia będzie sprzętowa reprezentacja 3D (np. holografia). Interaktywne ekrany holograficzne nie będą płaskie, nie będą miały ograniczonych kierunków obserwacji. Można w takim razie wywnioskować, że kula, która jest symetrycznym i naturalnym obiektem percepcyjnym dla ludzkiego układu nerwowego, zostanie preferowaną formą interfejsów aplikacji komputerowych w przyszłości. Trendy te już znalazły implementację w takich zastosowaniach, jak kuliste przeglądarki graficzne. Obecnie prym w tych badaniach na razie wiedzie Microsoft, który zademonstrował ostatnio prototyp interaktywnego urządzenia do wyświetlania zdjęć: *Multi-Touch Interactive Spherical Display*[6]. Innym przykładem nowoczesnego interfejsu do przeglądania obrazów nawiązujących do relewantnych dokumentów jest aplikacja *taggalaxy*.

Przytoczone przykłady pokazują, że prowadzone prace z zakresu wizualizacji danych są bardzo istotne i mogą zostać wykorzystane w praktycznych zastosowaniach nie tylko do prezentacji danych bibliometrycznych. Pozostaje mieć nadzieje że kolejne implementacje sprzętowe będą już uwzględniały najnowsze wyniki badań z zakresu wizualizacji informacji np. takich jak przedstawione w niniejszej pracy.

[5] Sphere: *A Multi-Touch Interactive Spherical Display* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://research.microsoft.com/en-us/um/people/benko/projects/sphere/.

[6] *Tag Galaxy* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.taggalaxy.de.

Bibliografia

1. *1100+ examples of information visualization* [on-line]. Infovis.info [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.infovis.info/index.php?cmd=search&words=science&mode=normal.
2. *ACM Computing Classification System* [on-line]. Association for Computing Machinery [dostęp 19 maja 2009]. Dostępny w World Wide Web:http://www.acm.org/about/class/.
3. Adamski M. *Informatyka – nauka, sztuka, czy rzemiosło?* [on-line]. Uniwersytet Zielonogórski – Miesięcznik Społeczności Akademickiej, 2002 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.uz.zgora.pl/wydawnictwo/ miesiecznik11-2002/17.pdf.
4. *American Society for Cybernetics* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.asc-cybernetics.org/.
5. *Association for Computing Machinery* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/.
6. Barát Á.H. *The Structures of Concept And its Connection to Sciences*. W: *Proceedings of IX ISKO Congres Spain Group, New Perspectives for the organization and dissemination of knowledge*. Valencia: UPV, 2009.
7. Bollen J. i in. *Clickstream Data Yields High-Resolution Maps of Science*. PLoS ONE [on-line] 2009, Vol. 4, no. 3 [dostęp 19 maja 2009] Dostępny w World Wide Web: http://www.plosone.org/article/info:doi/10.1371/journal.pone. 0004803.
8. Börner K., Chen Ch., Boyack K.W. *Visualizing Knowledge Domains*. W: Blaise Cronin (red.). *Annual Review of Information Science & Technology*. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology 2003, Vol. 37, s. 179-255.
9. Boyack Kevin W. i in. *Domain visualization using VxInsight for science and technology management*. Journal of the American Society for Information Science and Technology 2002, nr 53(9), s. 764-774.
10. Boyack K. W. i in. *Mapping the backbone of Science*. Scientometrics 2005, Vol. 64, nr 3, s. 351-374.
11. *Browse map. Overview* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.scimaps.org/browse/.
12. Budrewicz J. *Fraktale*. Warszawa: Wydawnictwo Naukowo-Techniczne, 1996.
13. Burns M., Bitner T. *Sztuka informowania*. Digit online [on-line] 2003, nr 6 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.digit.pl/artykuly/34291_6/sztuka.informowania.html.
14. *CAIDA: The Cooperative Association for Internet Data Analysis* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.caida.org/home/.
15. Case D. O. *Looking for Information: a survey of research on information seeking, needs and behavior*. 2 ed. 7London, UK: Elsevier 2002.
16. *Chaomei Chen's Homepage* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web http://www.pages.drexel.edu/~cc345/.

17. Chen Ch. *Information Visualization. Beyond the Horizon*. 2nd ed. London: Springer, 2006.
18. Chen Ch. *Mapping Scientific Frontiers. The Quest for Knowledge Visualization*. London: Springer, 2003.
19. Clayton K. *Fractals and the Fractal Dimension* [on-line]. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.vanderbilt.edu/AnS/ psychology/cogsci/chaos/workshop/Fractals.htm
20. Computing Curricula 2004. Overview Report including A Guide to Undergraduate Degree Programs in Computing. A cooperative project of ACM, AIS, IEEE-CS [on-line]. Shanghai Jiao Tong University. School Of Software [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://se.sjtu.edu.cn/ sites/se/gb/CCSE/CCCS-040601-Overview-Strawman-Rev4.pdf.
21. *Concise Encyclopedia of Computer Science*. Ed. by E. D. Reilly. Chichester, UK: Wiley, 2004.
22. Coulter N. i in. *Computing Classification System 1998: Current Status and Future Maintenance Report of the CCS Update Committee* [on-line]. New York: ACM, 1998 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/about/class/ccsup.pdf.
23. *Cranfield Demo* [on-line]. University of Twente, Netherlands. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://dbappl.cs.utwente.nl/pftijah/ Documentation/CranfieldDemo.
24. Crowley Ch. *Overview of Complexity* [on-line]. Data ostatatniej aktualizacji 10.05.2002. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://wynchar.com/charlie/Complexity/overviewOfComplexity.html
25. Daćko D. *Zastosowanie ontologii do odkrywania wiedzy* [on-line]. *Consensual knowledge* [Strona domowa D. Daćko] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.consensualknowledge.net/files/semlib_thesis.pdf.
26. Daconta M.C., Obrst Leo J., Smith Kevin T. *The Semantic Web*. Indiana, USA: Wiley 2002.
27. Denning P. J. *Is Computer Science Science?* Communications of the ACM 2005, Vol. 48, nr 4, s. 27-31.
28. Denning P. J. i in. *Computing as a Discipline*. Communication of the ACM [on-line] 1990, Vol. 32, no. 1 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://cs.gmu.edu/cne/pjd/GP/ CompDisc.pdf.
29. *DESIRE Information Gateways Handbook* [on-line]. DESIRE (Development of a European Service for Information on Research and Education), 1998-2000 [dostęp 19 maja 2009]. Dostępny World Wide Web: http://www.desire.org/ handbook/.
30. Dodge M., Kitchin R. *An Atlas of Cyberspace* [on-line]. Manchester, 2007 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/ atlas/atlas.html.
31. Duch W. *Fascynujący Świat Komputerów*. Poznań: Wydawnictwo NAKOM, 1997.
32. Duch W. *Pamięć* [on-line]. Wstęp do kognitywistyki 2007, nr 187 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.fizyka.umk.pl/~duch/ Wyklady/kog-m/04-pam.htm.

33. Dürsteler J. C. *Diagrams for Visualisation.* The digital magazine of InfoVis.net [online] 2007, nr 186 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.infovis.net/printMag.php?num= 186&lang=2.
34. Dürsteler J. C. *InfoVis Diagram.* The digital magazine of InfoVis.net [on-line] 2007, nr 187 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.infovis.net/printMag.php?num=187&lang=2.
35. *Elektroniczny Podręcznik Statystyki PL* [on-line]. Kraków: StatSoft, 2006 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.statsoft.pl/textbook/stathome.html.
36. *Eugene Garfield, Ph. D. Home Page* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://garfield.library.upenn.edu/.
37. *Exhibit Purpose and Goals* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://scimaps.org/.
38. Frame M., Mandelbrot B., Neger N. *Fractal Geometry* [on-line]. Yale University [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://classes.yale.edu/fractals/.
39. Garfield E. *Essays/Papers on „Mapping the World of Science"* [on-line]. *Eugene Garfield, Ph. D. Home Page* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://garfield.library.upenn.edu/ mapping/mapping.html .
40. Garfield E. *Scientography: Mapping the tracks of science.* Current Contents: Social & Behavioural Sciences 1994, nr 7(45), s. 5-10.
41. Gelernter J. *Visual Classification with Information Visualization (Infoviz) for Digital Library Collections.* Knowledge Organization 2007, nr 34, s. 128-143.
42. Gershon N. D., Eick S. G. *Guest Editors' Introduction: Information Visualization. The Next Frontier.* Journal of Intelligent Information Systems. 1998, Vol. 11 (3), s. 199-201.
43. Golub K. *Automated subject classification of textual Web pages based on a controlled vocabulary.* New Review of Hypermedia and Multimedia 2006, Vol. 12, no. 1, s. 11–27.
44. *Grokker – Enterprise Search Management and Content Integration* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.grokker.com/.
45. Hjørland B. *Domain analysis in information science: eleven approaches – traditional as innovative.* Journal of documentation 2002, nr 58, s. 422-462.
46. Hjørland B. *What is Knowledge Organization (KO)?* Knowledge Organization 2008, nr 35(4), s. 86-101.
47. Hjørland B., Albrechtsen H. *Toward a new horizon in information science: domain analysis.* Journal of the American Society for Information Science (JASIS) 1995, nr 46, s. 400-425.
48. Holloway T., Bozicevic M., Börner K. *Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors.* Wyd. specjalne: *Understanding Complex Systems.* Complexity 2007, Vol. 12, nr 3, s.30-40.
49. Hurvich M. L., Jameson D. *An opponent-process theory of color vision.* Psychological Review 1957, Vol. 64 (6), s. 384-404.
50. *IEEE Symposium on Information* Visualization (*InfoVis 2003*) [on-line]. Seattle, Washington, 2003 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.pages.drexel.edu/~cc345/papers/ infovis03.pdf.

51. *InfoVis 2007 Welcome* [on-line] [dostęp 19 maja 2009]. Dostepny w World Wide Web: http://conferences.computer.org/infovis/infovis2007/.
52. *Inxight [on-line] [dostęp 31 sierpnia 2008]. Dostępny w World Wide Web: http://* www.inxight.com (adres aktualny do grudnia 2008 r.).
53. Kaplan I. G. *Handbook of Molecular Physics and Quantum Chemistry*. Ed. by S. Wilson. New York: Wiley, 2003.
54. Kaliczyńska M. *Badanie struktury akademickiego społeczeństwa Informacyjnego z wykorzystaniem metody mds*. Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej 2005, nr 21, s. 137-144.
55. Kąkolewicz M. *Technologie informacyjne i konstruowanie wiedzy a qualia*. W: *Komputer w edukacji*. Morbitzer J. (red.). Kraków: Pracownia Technologii Nauczania, Akademia Pedagogiczna, 2008, s. 116-122.
56. Kingston J. *Ontologies, Multi-Perspective Modelling and Knowledge Auditing* [on-line]. *CEUR Workshop Proceedings* [RWTH Aachen University. Informatik 5] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://ftp.informatik. rwth-aachen.de/ Publications/CEUR-WS/Vol-48/kingston.pdf.
57. Kingston J. *Ontology, Knowledge Management, Knowledge Engineering and the ACM Classification Scheme* [on-line]. University of Edinburgh. School of Informatics [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.inf.ed.ac.uk/ publications/online/0169.pdf.
58. Klavans R., Boyack K. *Maps of Science: Forecasting Large Trends in Science* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.scimaps.org/dev/big_thumb.php?map_id=164.
59. Klavans R., Boyack K.W. *Identyfying a better measure of relatedness for mapping science*. Journal of the American Society for the Information Science and Technology 2005, Vol. 57, nr 2, s. 251 – 263.
60. Koch T., Neuroth H., Day M. *Renardus: Cross-browsing European subject gateways via a common classification system (DDC)*. W: *Subject Retrieval in a Network Environment: Papers Presented at an IFLA Satellite Meeting Sponsored by the IFLA Section on Classification and Indexing and IFLA Section of Information Technology, Dublin, Ohio, USA, 14-16 August 2001* [on-line]. Dublin, Ohio: OCLC, 2001 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.ukoln.ac.uk/metadata/renardus/ papers/ifla-satellite/ifla-satellite.pdf.
61. *Komputer w edukacji*. Pod red. J. Morbitzera. Kraków: Pracownia Technologii Nauczania, Akademia Pedagogiczna, 2008.
62. *Konwersja danych marc bn ===> MARC21* [on-line]. Warszawa: Biblioteka Narodowa, 2009 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://mak.bn.org. pl/wykaz5.htm.
63. Koshman Sh. *Visualization-based information retrieval on the Web*. Library and Information Science Research 2006, Vol. 28, nr 2, s. 192-207.
64. Larsen I. *Vector Space Modeling* [on-line]. THOR Center for Neuroinformatics [dostęp 19 maja 2009]. Dostepny w World Wide Web: http://eivind.imm.dtu.dk/thor/ projects/multimedia/textmining/ node5.html.
65. *Library of Congress Classification* [on-line]. The Library of Congress [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.loc.gov/catdir/cpso/lcc.html.

66. *List of Implicit Subject Descriptors in ACM CCS* [on-line]. *The ACM Portal* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://portal.acm.org/lookup/ccsno-un.cfm.
67. Lin X., Soergel D., Marchionini G. *A Self-organizing semantic maps* as graphical interfaces for *information retrieval.* W: *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in Information Retrieval.* College Park, USA: University of Maryland, 1991, s. 262-269.
68. Luther J., Kelly M., Beagle D. *Visualize This.* Library Journal [on-line] 2005, nr 3 (1) [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.libraryjournal.com/article/ CA504640.html.
69. Ma K. L. *Introduction to Visualization* [on-line]. DOE Office of Science Homepage [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.er.doe.gov/ascr/Research/scidac/intro_datavis.pdf;
70. Malina W., Smiatacz M. *Metody cyfrowego przetwarzania obrazów.* Warszawa: EXIT, 2005.
71. Małyszko D. *Systemy informacji przestrzennej* [on-line]. Politechnika Białostocka. Katedra Systemów Informacyjnych i Sieci Komputerowych [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://aragorn.pb.bialystok.pl/~dmalyszko/GIS_Materialy/SIP_Zajecia/SIP_Odwzorowania.htm.
72. Mandelbrot B. *Fractal Geometry of Nature.* Gordonsville, San Francisko, USA: W. H. Freeman & Co, 1982.
73. *Many Eyes* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://manyeyes.alpha works.ibm.com/manyeyes/.
74. Marshakova I. V. *A system of document connection based on references.* Scientific and Technical Information Serial of VINITI 1973, Vol. 6 (2), s. 3-8.
75. Mirkin B., Nascimento S., Pereira L. M. *Representing a Computer Science Research Organization on the ACM Computing Classification System* [on-line]. *CEUR Workshop Proceedings* [RWTH Aachen University. Informatik 5] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-354/p19.pdf.
76. Morville P. *Ambient Findability.* Sebastopol, USA: O'Reilly Media Inc., 2005.
77. Moya-Anegón F. i in. *A new technique for building maps of large scientific domains based on the cocitation of classes and categories.* Scientometrics 2004, Vol. 61, nr 1, s. 129-145.
78. *Multimedialne i sieciowe systemy informacyjne. Materiały konferencyjne.* Pod red. Cz. Daniłowicza. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2002.
79. Narojczyk K. *Komputerowa wizualizacja danych historycznych.* W: *Megabajty dziejów. Informatyka w badaniach, popularyzacji i dydaktyce historii,* (red.). R. T. Prinke, Poznań: Instytut Historii UAM, 2007, s. 79-95.
80. Neuman Ł. *Transformacja relacyjnych baz danych do postaci sieci Bayesa* [on-line]. Politechnika Wrocławska. Zakład Systemów Informacyjnych [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.zsi.pwr.wroc.pl/ zsi/missi2002/pdf/s210.pdf.
81. Niskanen I. *An interactive ontology visualization approach for the domain of networked home environments* [on-line]. Oulu: Julkaisija-Utgivare, 2007 [dostęp 19 maja

2009]. Dostępny w World Wide Web: http://www.vtt.fi/inf/pdf/ publications/2007/ P649.pdf.
82. Noyons E.C.M., Moed H.F. *Combining Mapping and Citation Analysis for Evaluative Bibliometric Purposes: A Bibliometric Study*. Journal of the American Society for Information Science 1999, nr 50(2), s. 115-131.
83. *OCLC DeweyBrowser Beta v.1.0*. OCLC Online Computer Library Center [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://deweybrowser.oclc.org/ddcbrowser/wcat.
84. *Online Computer Library Center* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.oclc.org.
85. *Orbiter model* [on-line]. NASA [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.nasa.gov/multimedia/imagegallery/image_feature_431.html.
86. Osińska V. *Przybliżenie semantyczne w wizualizacji informacji w Internecie i bibliotekach cyfrowych*. Biuletyn EBIB [on-line] 2006, nr 7 (77) [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.ebib.info/2006/77/osinska.php.
87. Osińska V., Bała P. *Classification Visualization across Mapping on a Sphere*. W: New trends of multimedia and Network Information Systems. Amsterdam: IOS press, 2008.
88. Osińska V., Bała P. *Nonlinear approach in classification visualization and evaluation*. W: *Proceedings of IX ISKO Congres Spain Group, New Perspectives for the organization and dissemination of knowledge*. Valencia, Hiszpania: UPV 2009. s. 222-231.
89. Osińska W. *Dynamika historycznego rozwoju stron WWW*. Biuletyn EBIB [on-line] 2007, nr 7 (88) [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.ebib.info/2007/88/.
90. Owen S.G. *Definitions, History, and Goals of Visualization: Overview* [on-line]. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.siggraph.org/ education/materials/HyperVis/ visgoals/definiti.htm.
91. Pajares F. *The Structure of Scientific Revolutions by Thomas S. Kuhn. Outline and Study Guide* [on-line]. Emory University. Division of Education Studies [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.des.emory.edu/mfp/Kuhn.html.
92. Pamuła N. *Typologia zasobów Ukrytego Internetu*. Przegląd biblioteczny, 2006, nr 2, s. 153-164.
93. Pitas I., Venetsanopoulos A.N. *Nonlinear digital filters: principles and applications*. Boston, USA: Springer, 1990.
94. Plotnick R. E., Gardner R. H. *Lacunarity indices as measures of landscape texture*. Landscape Ecology 1993, Vol. 8, nr 3, s. 201-211.
95. *Proceedings of I International Conference on Multidisciplinary Information Sciences and Technologies, Mérida (Spain), 25th-28th October, 2006* [on-line]. *E-LIS. E-prints in Library and Information Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://eprints.rclis.org/8170/1/Domain_analysis _by_means_of_the_visualization_of_maps_of_vast_scientific_domains.pdf.
96. *Proceedings of IX ISKO Congres Spain Group, New Perspectives for the organization and dissemination of knowledge*. Valencia: UPV, 2009.

97. *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in Information Retrieval*. College Park, USA: University of Maryland, 1991.
98. Rana M. *Historical Perspective on Information Systems* [on-line]. *Information Systems based on Logistics Perspective* [University of Houston] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.uh.edu/~mrana/try.htm.
99. Reichenstein O. *Trend Map 2008. What's new?* [on-line]. Information Architects Japan [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://informationarchitects.jp/trendmap3-countdown-sneak-peak/.
100. Rieger B., Kleber A., (von) Maur E. *Metadata-Based Integration of Qualitative and Quantitative Information Resources Approaching Knowledge Management* [online]. ECIS (European Conference on Information Systems) resources [London School of Economics and Political Science. Department of Information Systems] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://is2.lse.ac.uk/asp/ aspecis/20000115.pdf.
101. Robertson S. E. *The Probability Ranking Principle in IR*. Journal of Documentation 1977, Vol. 33, s. 294-304.
102. Rutkowski L. *Metody i techniki sztucznej inteligencji*. Warszawa: Wydawnictwo Naukowe PWN, 2005.
103. Sadowska J., Turowska T. *Języki informacyjno-wyszukiwawcze. Katalogi rzeczowe*. Warszawa: CUKB, 1990.
104. Samoylenko I., Chao T.-C., Liu W.-C., Chen C. M. *Visualizing the scientific world and its evolution*. Journal of the American Society for Information Science and Technology 2006, Vol. 57 (11), s. 1461-1469.
105. *Science – Thompson Reuters* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://scientific.thomson.com/.
106. *Science Frontiers 1976–* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.science-frontiers.com/.
107. Shneiderman B. *Treemaps for space-constrained visualization of hierarchies* [on-line]. University of Maryland. Department of Computer Science [dostęp 15 marca 2009]. Dostępny w World Wide Web: http://www.cs.umd.edu/hcil/treemap-history/.
108. Shoichiro N. *Numerical Analysis and Graphic Visualization with MATLAB. Upper Saddle River*, USA: Prentice Hall, 2002.
109. *Słownik pojęć komputerowych*. Pod red. V. Illingworth i J.Daintitha. Warszawa: Świat Książki, 2004.
110. Small H. *Co-citation in the scientific literature: A new measurement of the relationship between two documents*. Journal of the American Society of Information Science 1973, Vol. 24 (4), s. 265-269.
111. Small H., Griffith B. C. *The structure of scientific literatures I: Identifying and graphing specialities*. Science Studies 1974, nr 4, s. 17-40.
112. *Software Development Lohninger. SDL Component Suite – Kohonen* [on-line]. Software Development Lohninger [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.lohninger.com/ kohonen.html.
113. Sosińska-Kalata B. *Klasyfikacja. Struktury organizacji wiedzy, piśmiennictwa i zasobów informacyjnych*. Warszawa: SBP, 2002.

114. Sosińska-Kalata B. *Struktury klasyfikacyjne w organizacji zasobów informacyjnych Internetu* [on-line]. W: *Multimedialne i sieciowe systemy informacyjne. Materiały konferencyjne.* Pod red. Cz. Daniłowicza. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2002 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.zsi.pwr.wroc.pl/zsi/missi2002/pdf/ s403.pdf.
115. Soukup T., Davidson I. *Visual Data Mining: Techniques and Tools for Data Visualization and Mining.* New York, USA:John Wiley & Sons 2002.
116. Sperber D. *Why Rethink Interdisciplinarity?* [on-line] *Interdisciplines* 2009. [dostęp 16 maja 2009]. Dostępny w World Wide Web: http://www.interdisciplines.org/ interdisciplinarity/papers/1.
117. *Sphere: A Multi-Touch Interactive Spherical Display* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://research.microsoft.com/en-us/um/people/ benko/projects/sphere/.
118. Stasko J. *HCC Education Digital Library: Information Vizualization.* [on-line]. College of Computing, Georgia [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://hcc.cc.gatech.edu/taxonomy/ cat.php?cat=86.
119. Stasko J. *SunBurst* [on-line]. College of Computing, Georgia [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.cc.gatech.edu/gvu/ ii/sunburst/.
120. *Subject Retrieval in a Network Environment: Papers Presented at an IFLA Satellite Meeting Sponsored by the IFLA Section on Classification and Indexing and IFLA Section of Information Technology, Dublin, Ohio, USA, 14-16 August 2001* [on-line]. Dublin, Ohio: OCLC, 2001 [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.ukoln.ac.uk/metadata/renardus/ papers/ifla-satellite/ifla-satellite.pdf.
121. Tadeusiewicz R. *Sieci neuronowe.* Wyd. 2. Warszawa: Akademicka Oficyna Wydaw. RM, 1993.
122. *Tag Galaxy* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.taggalaxy.de.
123. *Taxonomy of Computer Science and Engineering.* Ed. by A. Ralston. Arlington,USA: AFIPS (American Federation of Information Processing Societies) Press, 1982.
124. *Teoria sieci neuronowych* [on-line]. *Neuralnets.eu – Polski Wortal Sztucznej Inteligencji* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://neuralnets.eu/ index.php?page=teoria&art=1.
125. *The 1964 Computing Reviews Classification System* [on-line]. Association for Computing Machinery [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/about/class/.
126. *The 1998 ACM Computing Classification System (1998)* [on-line]. Association for Computing Machinery [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/about/class/ ccs98-html.
127. *The ACM Computing Classification System [1998 Version]. Valid through 2009* [on-line]. Association for Computing Machinery [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.acm.org/class/1998/.
128. *The ACM Digital Library* [on-line]. *The ACM Portal* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://portal.acm.org/dl.cfm.

129. *The History of Programming Languages* [on-line]. O'Reilly Media [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://oreilly.com/news/graphics/prog_lang_poster.pdf.
130. *The History of Visual Communication* [on-line]. Sabanci University, Istanbul [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.citrinitas.com/history_of_viscom/.
131. *The psychology of learning and motivation: Advances in research and theory*. Ed. by D. L. Medin. San Diego in: Academic Press, 1989, Vol. 24.
132. *Thinkmap Visual Thesaurus* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.visualthesaurus.com/.
133. Tufte E.R. *Envisioning Information*. Connecticut, USA: Graphics Press 1990.
134. *Udostępniaj swoją twórczość na jasnych, przyjaznych zasadach – na licencjach CreativeCommon* [on-line]. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://creativecommons.pl/.
135. van Rijsbergen C. J. *Information Retrieval* [on-line]. 2nd ed. London: Butterworths, 1979. [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.dcs.gla.ac.uk/Keith/Preface.html#PREFACE.
136. *Viz4All – visualization survey* [on-line]. University of Maryland. College Park [dostep 19 maja 2009]. Dostępny w World Wide Web: http://www.cs.umd.edu/class/spring2005/cmsc838s/viz4all/viz4all _a.html.
137. *Walrus – Gallery: Visualization & Navigation* [on-line] CAIDA, the Cooperative Association for Internet Data Analysis [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.caida.org/tools/visualization/walrus/gallery1/.
138. *Walrus – Graph Visualization Tool* [on-line] CAIDA, the Cooperative Association for Internet Data Analysis [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.caida.org/tools/visualization/walrus/.
139. *Wapedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://wapedia.mobi/pl/.
140. Ware C. *Information Visualization: Perception for Design*. Wyd. 2. San Francisco: Morgan Kaufmann, 2004.
141. *WEBSOM – A novel SOM-based approach to free-text mining* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://websom.hut.fi/websom/.
142. *WEBSOM map* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://websom.hut.fi/websom/comp.ai.neural-nets-new/html/root.html.
143. *What is Visualization?* [on-line]. Infovis. Information Visualization Resources [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://infovis.org/.
144. White H. D., McCain K. W. *Visualization of literatures*. Annual Review of Information Science and Technology 1997, Vol. 32, s. 99-168.
145. *Wikipedia. The Free Encyclopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/Main_Page.
146. *Wikipedia. Wolna Encyklopedia* [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Strona_główna.
147. *WP12: Cross concordances of classifications and thesauri* [on-line]. *CARMEN: Content Analysis, Retrieval and MetaData: Effective Networking* [Universitätsbibliothek

Regensburg] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.mathematik.uni-osnabrueck.de/projects/carmen/ index.en.shtml.
148. Yurcik W.J. *Scientific Visualization* [on-line]. BookRags [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.bookrags.com/research/scientific-visualization-csci-03/scientific-visualization-csci-03.html.
149. Young F. W. *Multidimensional Scaling* [on-line]. University of North Karolina. Department of Psychology [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://forrest.psych.unc.edu/teaching/ p208a/mds/mds.html.
150. Zeller D. *Hypothetical Model of the Evolution of Science* [on-line]. *Places@Spaces: Mapping Science* [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://www.scimaps.org/dev/map_detail.php? map_id=163.
151. Zyglarski B., Schroeiber T., Bała P. *Web Services Based Scientific Article Manager*. W: *Information Systems Architecture and Technology. Web Information Systems: Models, Concepts and Challenges*. Wrocław: Wydawnictwo Politechniki Wrocławskiej, 2008, s. 205-216.
152. Википедия – Свободная энциклопедия [on-line] [dostęp 19 maja 2009]. Dostępny w World Wide Web: http://ru.wikipedia.org/wiki/.

Spis tabel z wynikami badań

Tabela A. Zestawienie oryginalnego schematu klasyfikacji CCS i zmodernizowanego ... 152

Tabela B1. Charakterystyki artykułu nr 1622 i sąsiadujących z nim na mapie wizualizacji przy relacji klasyfikacji podstawowej i dodatkowych jako 0.6:0.4 ... 172

Tabela B2. Charakterystyki artykułu nr 1622 i sąsiadujących z nim na mapie wizualizacji przy relacji klasyfikacji podstawowej i dodatkowych jako 0.7:0.3 ... 175

Tabela B3. Charakterystyki artykułu nr 882 i sąsiadujących z nim na mapie wizualizacji przy relacji klasyfikacji podstawowej i dodatkowych jako 0.7:0.3 ... 178

Tabela B4. Charakterystyki artykułu nr 719 i sąsiadujących z nim na mapie wizualizacji przy relacji klasyfikacji podstawowej i dodatkowych jako 0.7:0.3 ... 180

Tabela A. Zestawienie oryginalnego schematu klasyfikacji CCS i zmodernizowanego

Klasa	Klaster Il. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne	Słowa kluczowe (wg listy rankingowej)
A General Literature	1 82 / 917	A.0	General		Software engineering
					IEEE Computer Society
					privacy
					security
					Turing award
					computational science
					pervasive computing
					engineering
	2	A.0	General		interaction design
		A.2	Reference		user experience
A		A.m	Miscellaneous		Web 2.0
B Hardware	1 góra 676/ 1166	B.1.1	Control Design Styles		FPGA chips
		B.1.2	Control Structure Performance Analysis		low power
		B.2	Arithmetic and logic structures		CMOS
		B.2.1	Design Styles		performance analysis
		B.2.4	High-Speed Arithmetic		simulation
		B.3.1	Semiconductor Memories		cache
		B.3.2	Design Styles		VLSI
		B.3.3	Performance Analysis and Design Aids		Emerging technologies
		B.4	Input/output and data communications		fault-tolerance
		B.4.1	Data Communications Devices		reliability
		B.4.2	Input/Output Devices		Leakage current
		B.4.3	Interconnections (Subsystems)		statistical analysis
B		B.5.1	Design		Modelling

Klasa	Klaster II. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne	Słowa kluczowe (wg listy rankingowej)
A	1 góra 676/ 1166	B.5.2	Design Aids		Nanotechnology
		B.6	Logic design		Genetic algorithm
		B.6.1	Design Styles		High-k dielectric
		B.6.3	Design Aids		high-level synthesis
		B.7	Integrated circuits		Networks-on-chip
		B.7.1	Types and Design Styles		
		B.7.2	Design Aids		
		B.7.3	Reliability and Testing		
		B.7.m	Miscellaneous		
		B.8	Performance and reliability		
		B.8.1	Reliability, Testing, and Fault-Tolerance		
		B.8.2	Performance Analysis and Design Aids		
Hardware	1 dół 99 / 156	B.2.4	High-Speed Arithmetic		FPGA chips
		B.3.1	Semiconductor Memories		low power
		B.3.2	Design Styles		evolvable hardware
		B.5.1	Design		Genetic algorithm
		B.5.2	Design Aids		high-level synthesis
		B.6.1	Design Styles		
		B.6.3	Design Aids		
		B.7	Integrated circuits		
		B.7.1	Types and Design Styles		
		B.7.2	Design Aids		
		B.8	Performance and reliability		
A		B.8.1	Reliability, Testing, and Fault-Tolerance		

C Computer Systems Organizations	1 góra 1128 / 1717	C	Computer Systems Organization	Hardware/software interfaces Instruction set design Modelling of computer architecture System architectures Systems specification methodology Adaptable architectures Cellular architecture (e.g., mobile) Data-flow architectures Frame relay networks Network topology Wireless communication Circuit-switching networks ISDN (Integrated Services Digital Network)	RISC/CISC, VLIW architectures Array and vector processors Associative processors Connection machines Interconnection architectures (e.g., common bus, multiport memory, crossbar switch) Heterogeneous (hybrid) systems Data communications Open Systems Interconnection reference model (OSI) Security and protection (e.g., firewalls) Packet-switching networks Store and forward networks	Wireless LAN
		C.1	Processor architectures			sensor networks
		C.1.1	Single Data Stream Architectures			Quality of service
		C.1.2	Computer Systems Organization			802.11
		C.1.3	Processor architectures			performance
		C.2	Computer-communication networks			ad hoc networks
		C.2.1	Network Architecture and Design			security
						mobile ad hoc networks (MANETs)
						Internet
						energy efficiency
						mobility
						routing
						multicast
						Fault tolerance
						Intrusion Detection
						broadcast
						scheduling
	1 dół 371 / 567	C	Computer Systems Organization	Network management Network monitoring Public networks Distributed applications Distributed databases Network operating systems	Process control systems Real-time and embedded systems Smartcards Design studies Fault tolerance Modelling techniques Performance attributes Reliability, availability, and serviceability	peer-to-peer
		C.1	Processor architectures			distributed computing
		C.1.1	Single Data Stream Architectures			grid computing
		C.1.2	Computer Systems Organization			quality of service
		C.1.3	Processor architectures			Scheduling
		C.2	Computer-communication networks			fault-tolerance
		C.2.1	Network Architecture and Design			
C		C.2.3	Network Operations			

Klasa	Klaster Il. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne		Słowa kluczowe (wg listy rankingowej)
C	1 dół 371 / 567	C.2.4	Distributed Systems			
		C.2.m	Miscellaneous			
		C.3	Special-purpose and app-based system			
		C.4	Performance of systems			
		C.5.0	Computer system implementation			
		C.m	Miscellaneous			
Computer Systems Organizations	2 318 / 452	C.2.2	Network Protocols	Protocol architecture Protocol verification Routing protocols	Applications (SMTP, FTP, etc.) Open Systems Interconnection reference model (OSI)	routing
		C.5	Computer system implementation			wireless networks
						Ad hoc networks
						Quality of Service
						multicast
						Performance evaluation
						Sensor networks
						TCP
						mobile ad hoc networks
						Internet
						Security
						protocols
	3 214 / 345	C.1.2	Computer Systems Organization	Client/server Process control systems Real-time and embedded systems Signal processing systems Smartcards	Open Systems Interconnection reference model (OSI) Microprocessor/microcomputer applications Array and vector processors	embedded systems
		C.1.m	Miscellaneous			FPGA
		C.2.0	Computer-communication networks			performance analysis
		C.2.4	Distributed Systems			sensor networks
		C.2.6	Internetworking			smart card
C		C.3	Special-purpose and app-based system			

D		D.1	Programming techniques		Java
		D.1.1	Applicative (Functional) Programming		software testing
		D.1.2	Automatic Programming		model checking
		D.1.3	Concurrent Programming		aspect-oriented programming
		D.1.5	Object-oriented Programming		Software Architecture
		D.1.7	Visual Programming		parallel programming
		D.1.m	Miscellaneous		object-oriented programming
		D.2	Software engineering		compiler
		D.2.1	Requirements/Specifications		Debugging
		D.2.11	Software Architectures		performance
		D.2.13	Reusable Software		static analysis
		D.2.4	Software/Program Verification		verification
		D.2.5	Testing and Debugging		component-based systems
Software	1 831/ 273	D.2.7	Distribution, Maintenance, Enhancem.		program analysis
		D.2.8	Metrics		concurrency
		D.2.9	Management		functional programming
		D.3	Programming languages		security
		D.3.2	Language Classifications		regression testing
		D.3.3	Language Constructs and Features		virtual machines
		D.3.4	Processors		automatic test generation
		D.4.0	Operating systems		access control
		D.4.2	Storage Management		UML
		D.4.5	Reliability		multithreading
		D.4.6	Security and Protection		XML
		D.4.7	Organization and Design		
D		D.1.1	Applicative (Functional) Programming		component-based systems
		D.1.2	Automatic Programming		Software Architecture

Klasa	Klaster II. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne	Słowa kluczowe (wg listy rankingowej)
D		D.1.3	Concurrent Programming		software engineering
		D.1.5	Object-oriented Programming		web services
		D.1.7	Visual Programming		Design Patterns
		D.2	Software engineering		QoS
		D.2.1	Requirements/Specifications		Requirements engineering
		D.2.11	Software Architectures		adaptation
	2	D.2.12	Interoperability		Process model
	359 / 691	D.2.13	Reusable Software		UML
		D.2.5	Testing and Debugging		programming
		D.2.6	Programming Environments		semantic web
		D.2.7	Distribution, Maintenance, Enhancem.		SOA
		D.2.8	Metrics		Grid computing
Software		D.2.9	Management		java
		D.4.6	Security and Protection		
		D.1	Programming techniques		model checking
		D.1.1	Applicative (Functional) Programming		formal verification
		D.1.6	Logic Programming		Security
		D.1.7	Visual Programming		access control
		D.2.0	Software engineering		Verification
	3	D.2.4	Software/Program Verification		static analysis
	270 / 467	D.3.1	Formal Definitions and Theory		Trusted computing
		D.3.2	Language Classifications		program analysis
		D.3.3	Language Constructs and Features		Abstract interpretation
		D.3.4	Processors		distributed systems
		D.4.6	Security and Protection		information flow
D		D.4.9	Systems Programs and Utilities		complexity

D Software D	4 281 / 467	D.1.3	Concurrent Programming		grid computing
		D.2.5	Testing and Debugging		Real-time systems
		D.3.4	Processors		Distributed computing
		D.4	Operating systems		Performance analysis
		D.4.1	Process Management		scheduling
		D.4.2	Storage Management		parallel computing
		D.4.3	File Systems Management		fault-tolerance
		D.4.4	Communications Management		Resource management
		D.4.5	Reliability		
		D.4.7	Organization and Design		
		D.4.8	Performance		
	5 105 / 195	D.1.3	Concurrent Programming		software engineering
		D.2.2	Design Tools and Techniques		Object oriented languages
		D.2.3	Coding Tools and Techniques		aspect-oriented programming
					conceptual modeling
					UML
E Data E	1 273 /477	E.1	Data structures		Cryptography
		E.3	Data encryption		security
		E.4	Coding and information theory		Quantum computing/cryptography
					cdma
					digital signatures
					bilinear maps

Klasa	Klaster II. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne	Słowa kluczowe (wg listy rankingowej)
F		F	Theory of Computation		complexity
		F.1	Computation by abstract devices		automata
		F.1.2	Modes of Computation		Scheduling
		F.1.3	Complexity Measures and Classes		approximation algorithms
		F.2	Analysis of alg. and problem complexity		Quantum computing
		F.2.1	Numerical Algorithms and Problems		Combinatorics
	1	F.2.2	Nonnumerical Algorithms and Problems		formal languages
	667 / 905	F.3	Logics and meanings of programs		Distributed systems
		F.4	Mathematical logic and formal languag.		regular expressions
		F.4.2	Grammars and Other Rewriting Syst.		Analysis of algorithms
		F.4.3	Formal Languages		Heuristics
					Robustness
					Rewriting logic
					abstract interpretation
Theory of Computation		F.1.1	Models of Computation		Constraint networks
		F.2	Analysis of algor. and problem complex.		formal methods
		F.4.1	Mathematical Logic		Hybrid Logic
	2	F.m	Miscellaneous		linear logic
	139 / 208				modal logic
					rough sets
					Temporal logic
		F.1.1	Models of Computation		Program analysis
		F.3.1	Specifying, Verifying and Reasoning		Semantics
		F.3.2	Semantics of Programming Languages		bugs
	3	F.3.3	Studies of Program Constructs		software quality
	132 / 174				static analysis
					type structure
					Logical relations
F					Distributed systems

F	4 136 / 170	F.1.1	Models of Computation		quantum states
		F.1.2	Modes of Computation		Rewriting logic
		F.2.2	Nonnumerical Algorithms and Problems		cellular automata
		F.2.3	Tradeoffs between Complexity Measur.		Automata
		F.3	Logics and meanings of programs		Membrane computing
		F.3.1	Specifying, Verifying and Reasoning		Formal languages
		F.3.2	Semantics of Programming Languages		Neural networks
		F.4.2	Grammars and Other Rewriting Syst.		finite automata
G Mathematics of Computing G	1 1781 / 081	G.1	Numerical analysis		65[B-Y]??
		G.1.1	Interpolation		Finite difference/element method
		G.1.10	Applications		Boundary element method
		G.1.3	Numerical Linear Algebra		Nonlinear equations/systems
		G.1.4	Quadrature and Numerical Different.		Stochastic optimization
		G.1.5	Roots of Nonlinear Equations		Convergence/ analysis
		G.1.6	Optimization		Matrix
		G.1.7	Ordinary Differential Equations		Numerical methods
		G.1.8	Partial Differential Equations		Differential equations
		G.1.9	Integral Equations		approximate solution/approxi-rnation
		G.1.m	Miscellaneous		Error estimates/analysis
		G.2	Discrete mathematics		Constrained optimization
		G.2.m	Miscellaneous		Simulation
		G.3	Probability and statistics		Delay system
		G.4	Mathematical software		Spectral method
					Fuzzy
					Adaptive control/representation
					Stability analysis

Klasa	Klaster II. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne	Słowa kluczowe (wg listy rankingowej)
G		G	Mathematics of Computing		Graph algorithm
		G.2.1	Combinatorics		approximation algorithms
		G.2.2	Graph Theory		Cycles
					Tree
					Hypercube, hypergraph
	2 496/ 670				Hamiltonian
					Path and cut enumeration
					Randomized/parallel algorithms
					Interconnection networks
					Fault tolerance
		G.1.2	Approximation		approximation algorithms
	3 219/300	G.2.1	Combinatorics		Wavelets
		G.2.2	Graph Theory		Fast Fourier transform
G					Scheduling

Mathematics of Computing

H Information Systems	1 1917 / 2747	H.1	Models and principles	Software psychology Access methods Statistical databases Clustering Query formulation Relevance feedback Retrieval models Search process Recovery and restart Dissemination Systems issues User issues Desktop publishing Spreadsheets Word Processing Natural language Prototyping Training, help, and documentation User-centered design Image databases	Data mining Human factors Human information processing Information filtering Scientific databases Interaction styles (e.g., commands, menus, forms, direct manipulation) Graphical user interfaces (GUI) Information browsers Information networks Performance evaluation User profiles and alert services Web-based services Animations, Video Audio input/output User interface management systems Artificial, augmented, and virtual realities Evaluation/methodology	database
		H.1.2	User/machine systems			Visualization
		H.2.2	Physical design			Data mining
		H.2.8	Database applications			information retrieval/seeking
		H.3	Information storage and retrieval			Web Services
		H.3.3	Information Search and Retrieval			query expansion/processing
		H.3.4	Systems and Software			ontology
		H.3.5	Online information services			context information/awareness
		H.3.7	Digital libraries			clustering
		H.4	Information systems applications			decision support/making
		H.4.1	Office Automation			search atrategy/engine
		H.4.3	Communications Applications			semantic web
		H.5.1	Multimedia Information Systems			HCI/human-robot-interaction
		H.5.2	User Interfaces			Multimedia
		H.5.4	Hypertext/Hypermedia			evaluation
		H.5.5	Sound and Music Computing			user study
						user interfaces
						Web search
						Knowledge Management
	2 449/597	H.1	Models and principles			database
		H.1.1	Systems and Information Theory	General systems theory Information theory Value of information Abstracting methods Dictionaries Indexing methods Linguistic processing Thesauruses	Decision support (e.g., MIS) Logistics Data description languages (DDL) Data manipulation languages (DML) Database programming languages Query languages	query processing
		H.2	Database management			XML
		H.2.1	Logical design			ontology
		H.2.3	Languages			Decision support
		H.2.4	Systems			indexing
		H.2.5	Heterogeneous databases			information retrieval
		H.2.m	Miscellaneous			semantic web
		H.3	Information storage and retrieval			Performance analysis
		H.3.1	Content Analysis and Indexing			
		H.4.2	Types of Systems			
H		H.4.m	Miscellaneous			

Klasa	Klaster Il. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne		Słowa kluczowe (wg listy rankingowej)
H		H.1.1	Systems and Information Theory	General systems theory Information theory Value of information Data models Ergonomics Natural language Theory and methods Schema and subschema Image databases Scientific databases Spatial databases and GIS Statistical databases	Abstracting methods Dictionaries Indexing methods Linguistic processing Thesauruses Decision support (e.g., MIS) Evaluation/methodology Graphical user interfaces (GUI) Interaction styles (e.g., commands, menus, forms, direct manipulation) Screen design (e.g., text, graphics, color)	Decision support system
		H.2	Database management			databases
		H.2.1	Logical design			XML
		H.2.5	Heterogeneous databases			Knowledge management
		H.2.8	Database applications			visualization
		H.3	Information storage and retrieval			Fuzzy sets
		H.3.1	Content Analysis and Indexing			ontology
	3	H.4.2	Types of Systems			Multicriteria decision
	303/471	H.5	Information interfaces and presentation			data warehouse
		H.5.2	User interfaces			Conceptual modeling
						Data models
		H	Information Systems	Evaluation/methodology Organizational design Asynchronous interaction Synchronous interaction Collaborative computing Theory and models	Computer-supported cooperative work Web-based interaction	collaboration/collaborative learning
		H.5.3	Group and Organization Interfaces			social computing
						Wikipedia
						knowledge management
	4					communication
	163/222					computer-mediated communication
						awareness
H						visualization

Information Systems

I		I.2	Artificial intelligence		classes, classification, classifier
		I.2.3	Deduction and Theorem Proving		Learning
		I.2.6	Learning		classification
		I.2.7	Natural language processing		claster analysis, clasterization
		I.3.3	Picture/image generation		fuzzy systems
		I.4.10	Image representation		recognition
		I.4.9	Applications		neural networks
	1	I.5.1	Models		Image recognition
	1239/1687	I.5.2	Design methodology		similarity
		I.5.3	Clustering		text mining
		I.5.4	Applications		SVM
		I.6	Simulation and modeling		feature extraction
		I.6.4	Model Validation and Analysis		genetic programming
		I.6.5	Model development		speech detection
					PCA
					pattern recognition
					datamining
Computing Methodologies		I.1.2	Algorithms		Image analysis/processing/ retrieval
		I.2	Artificial intelligence		learning/machine learning
		I.2.1	Applications and Expert Systems		fuzzy
		I.2.10	Vision and Scene Understanding		face detection/recognition
	2	I.2.3	Deduction and Theorem Proving		motion
	979 / 1510	I.2.6	Learning		robots/robotics
		I.2.9	Robotics		classification
		I.4	Image processing and computer vision		color analysis/processing/ images
I		I.4.1	Digitization and Image Capture		retrieval

Klasa	Klaster II. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne	Słowa kluczowe (wg listy rankingowej)
I		I.4.2	Compression (coding)		wavelet analysis
		I.4.3	Enhancement		video analysis
		I.4.4	Restoration		simulation/simulated
		I.4.5	Reconstruction		3D
		I.4.6	Segmentation		Texture analysis
		I.4.7	Feature measurement		
	2	I.4.8	Scene analysis		
	979 / 1510	I.5	Pattern recognition		
		I.5.2	Design methodology		
		I.5.4	Applications		
		I.6	Simulation and modeling		
		I.6.5	Model development		
		I.6.8	Types of Simulation		
		I.2	Artificial intelligence		Fuzzy control/distance
		I.2.1	Applications and Expert Systems		Learning
		I.2.11	Distributed artificial intelligence		Ontology/ontologies
		I.2.3	Deduction and Theorem Proving		Knowledge discovery/representation
		I.2.4	Knowledge Represent. Formal. Methods		Cognition/cognitive
	3	I.2.9	Robotics		Multi-agent
	660 / 960	I.4	Image processing and computer vision		Neural networks
		I.5	Pattern recognition		Multi-agent systems
		I.5.2	Design methodology		Uncertainty
		I.5.3	Clustering		Decision making/analysis
		I.5.4	Applications		Semanrtic web
I		I.6.5	Model development		mobile robots

Computing Methodologies

I		I.1.3	Languages and Systems		Genetic algorithm
		I.2.8	Problem Solv., Control Methods, Search		heuristics (multiheurisitcs)
		I.3.3	Picture/image generation		Adaptive control
		I.3.4	Graphics utilities		02.30.Yy ?
		I.3.5	Computation Geomet., Object Modeling		Scheduling
		I.5.1	Models		fuzzy
	4	I.5.2	Design methodology		Neural network
	471 / 564	I.5.4	Applications		Evolutionary algorithms/computation
		I.5.5	Implementation		Nonlinear system
		I.6.1	Simulation theory		Optimization
		I.6.3	Applications		Physically-based modeling
		I.6.5	Model development		Robust control
					Geometric modeling
		I.1.1	Expressions and Their Representation		images
		I.1.2	Algorithms		reconstruction
		I.1.4	Applications		shape
		I.2.6	Learning		Wavelet
		I.4.1	Digitization and Image Capture		Data fitting/hidding/assignment
		I.4.2	Compression (coding)		Feature extraction
	5	I.4.3	Enhancement		Image analysis
	255 / 367	I.4.5	Reconstruction		retrieval
		I.4.7	Feature measurement		JPEG
		I.5.2	Design methodology		Neural networks
		I.5.4	Applications		Texture analysis
		I.6	Simulation and modeling		Camera calibration
		I.6.4	Model Validation and Analysis		
		I.6.5	Model development		
I		I.6.8	Types of Simulation		

Computing Methodologies

Klasa	Klaster II. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne	Słowa kluczowe (wg listy rankingowej)
I Computing Methodologies I	6 184 / 293	I.1.1	Expressions and Their Representation		Virtual reality
		I.1.2	Algorithms		interaction
		I.2.6	Learning		visualization
		I.2.11	Distributed artificial intelligence		Texture design
		I.3.5	Computational Geometry, Object Model.		animation
		I.4.1	Digitization and Image Capture		3D modeling
		I.4.2	Compression (coding)		GPU computing
		I.4.5	Reconstruction		XML
		I.5.1	Models		Simulation
		I.5.4	Applications		
		I.6.4	Model Validation and Analysis		
		I.6.8	Types of Simulation		
		I.7.1	Document and Text Editing		
		I.7.2	Document preparation		
J Computer Applications J	1 554 / 931	J.0	Computer applications		health care
		J.3	Life and medical sciences		Medical Informatics Applications
		J.4	Social and behavioral sciences		genetic algorithm
		J.5	Arts and humanities		Grid computing
		J.7	Computers in other systems		bioinformatics
		J.m	Miscellaneous		Simulation
					Dynamic programming
					clustering
					image processing
					Wavelet analysis

J Computer Applications J	1 554 / 931				modelling
					Sequence
					Adaptation
					optimization
					Telemedicine
					Classification
					Finite element modeling
					Microarray
	2 387 / 463	J.2	Physical sciences and engineering		Lithography
		J.6	Computer-aided engineering		modeling
		J.m	Miscellaneous		Nanoimprint
					Numerical methods
					Finite element method
					Microfluidics
					Genetic algorithms
					Photonic crystals
					Surface modeling
					MEMS
					simulation
					Biosensors
					geometry
	3 76 / 145	J.1	Administrative data processing		algorithms
		J.3	Life and medical sciences		business process management
		J.4	Social and behavioral sciences		e-government
					modeling

Klasa	Klaster II. dok[1].	Podklasy klasyfikacji CCS		Deskryptory tematyczne	Słowa kluczowe (wg listy rankingowej)
K	1 322 / 633	K.3.1	Computer Uses in Education		IT/ICT
		K.4.3	Organizational impacts		Software development
		K.5.2	Governmental issues		knowledge management
		K.6.0	Management of computing and information systems		Decision support
		K.6.1	Project and People Management		education
		K.6.2	Installation management		collaboration
		K.6.3	Software management		project management
		K.7.1	Occupations		
Computing Milieux	2 366/610	K.0	Computing milieux		Learning
		K.1	The computer industry		Education
		K.3.0	Computers and education		Cooperative/collaborative learning
		K.3.1	Computer Uses in Education		e-learning
		K.3.2	Computer and Information Science Education		teaching
		K.3.m	Miscellaneous		Interactive learning environments
		K.4	Computers and society		pedagogy
		K.4.2	Social issues		distance education
		K.5.0	Legal aspects of computing		multimedia
		K.6.0	Management of computing and information systems		web-based education
		K.6.1	Project and People Management		evaluation
K		K.6.2	Installation management		Distributed learning environments

K		K.6.3	Software management		Computer-mediated communication
Computing Milieux		K.6.4	System management		Human-computer interface
		K.7.4	Professional ethics		
	3 góra 171 / 294	K.4.1	Public policy issues		security
		K.4.4	Electronic commerce		Privacy
		K.5.1	Hardware/software protection		authentication
		K.6.5	Security and Protection		trust management
		K.6.m	Miscellaneous		access control
		K.7.2	Organizations		e-commerce
	3 dół 119 / 230	K.2	History of computing		E-commerce
		K.3.m	Miscellaneous		Security
		K.4.1	Public policy issues		Web technology
		K.4.4	Electronic commerce		online shopping
		K.5	Legal aspects of computing		Privacy
		K.6.4	System management		mobile commerce
		K.6.5	Security and Protection		Social Networks
		K.7.4	Professional ethics		Case study
		K.7.m	Miscellaneous		Trust
		K.8.1	Application packages		Multi-agent systems
					phishing
K					Grid computing

[1] Całkowita ilość dokumentów w klastrze / ilość dokumentów ze słowami kluczowymi.

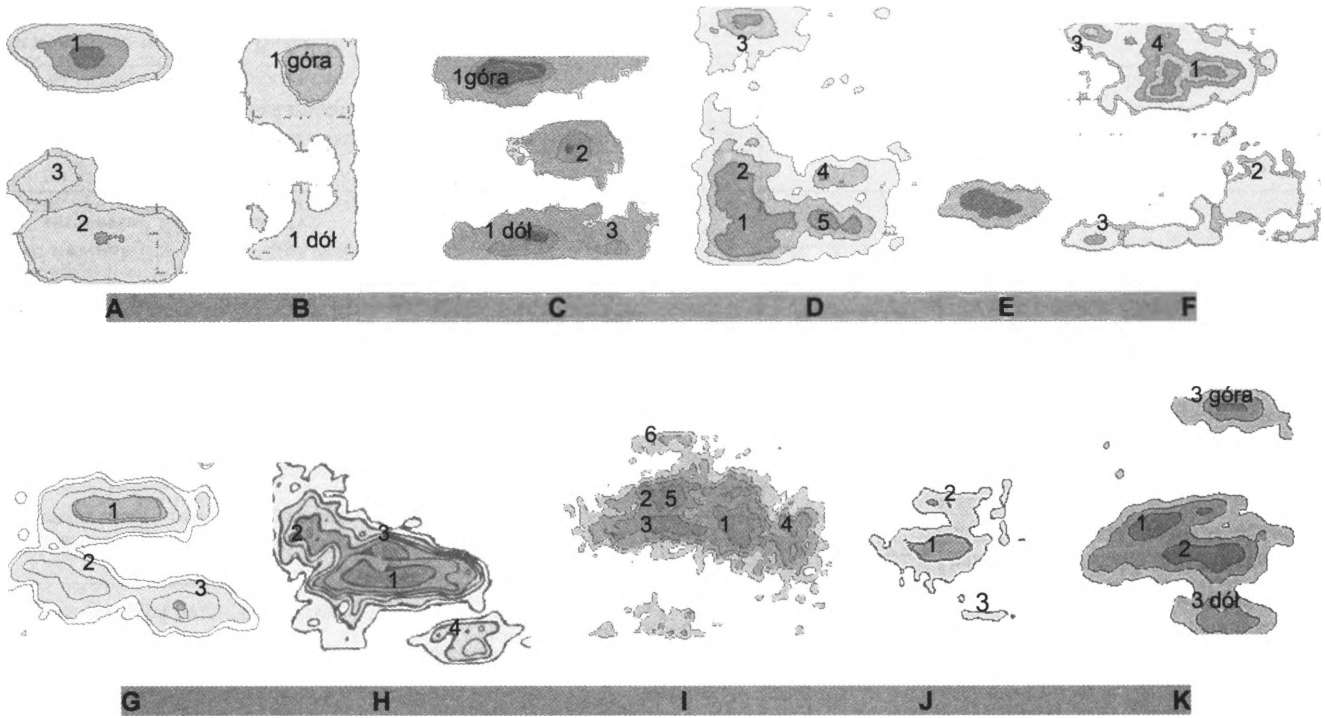

1
3
2
1 góra
1 dół
1góra
2
1 dół
3
3
2
4
1
5
3
4
1
2
3
A
B
C
D
E
F
1
2
3
2
3
1
4
6
2 5
3
1
4
2
1
3
3 góra
1
2
3 dół
G
H
I
J
K

Tabela B1. Charakterystyki artykułu nr 1622 i sąsiadujących z nim na mapie wizualizacji przy relacji klasyfikacji podstawowej i dodatkowych jako 0.6:0.4

Nr	Tytuł	Autor(zy)	Klasa gł	Klasy dodatkowe	Słowa kluczowe
1622	**Combining bibliometrics, information retrieval, and relevance theory, Part 2: Some implications for information science:**	Howard D. White	H.3.3	H.3.4, I.2.0	bibliometrics, cognitive models, information retrieval, query formulation, relevance
1426	**Parameterized pattern queries**	Casdric du Mouza et al.	H.3.3	F.2.2, H.2.4, H.3.4, I.5.2	optimization, Pattern matching, Representation and manipulation of data
1468	**Applying logistic regression to relevance feedback in image retrieval systems**	T. LeÄln, P. Zuccarello et al.	H.3.3	G.3, H.2.8, H.3.3, H.4.2, I.4.7	content-based image retrieval systems, Logistic regression, Low-level image descriptors, Visual information retrieval
1847	**ConnectA!: An Intelligent Search Engine based on Authors' Connectivity**	Hamidreza Baghi et al.	H.3.4	H.3.3, H.3.4, H.4.3, I.2.4	
1861	**Which factors explain the Web impact of scientists' personal homepages?**	Franz Barjak at al.	H.3.4	H.3.4, H.3.5	
2136	**Online forums supporting grassroots participation in emergency preparedness and response**	Leysia Palen **et al.**	H.3.5	H.3.4, K.4.2	
2166	**Analysis of navigability of Web applications for improving blind usability**	Hironobu Takagi **et al.**	H.3.5	H.5.2, H.5.4, K.4.2	accessibility, Web accessibility, online shopping, usability testing, voice browsers
2168	**Mobile computer Web-application design in medicine: some research based guidelines**	**Andreas Holzinger, Maximilian Errath**	H.3.5	H.5.2, J.3	Information interfaces and representation, Interface design, Internet applications, Life and medical sciences, Mobile computing
2674	**The Indicator Browser: A Web-Based Interface for Visualizing UrbanSim Simulation Results**	**Yael Schwartzman, Alan Borning**	H.4.3	I.6.7, J.7	
3127	**Key factors of heuristic evaluation for game design: Towards massively multi-player online role-playing game**	Seungkeun Song, Joohyeon Lee	H.5.2	H.1.2, K.8.0	design process, Heuristics evaluation, MMORPG game design, Usability

3231	**Weathergods: tangible interaction in a digital tabletop game**	**Saskia Bakker et al.**	H.5.2	H.5.2, K.8.0	digital tabletop gaming, interaction design, pervasive games, tangible interaction, tangible user interfaces
3053	**Dynamic personalization of web sites without user intervention**	**Ranieri Baraglia, Fabrizio Silvestri**	H.3.3	H.3.4, I.2.0	
1900	**Prediction of Information Sharing Behavior in China: Understanding the Cultural and Social Determinants**	**Jessica Pu Li et al.**	H.3.5	H.3.4, K.4.2	
1853	**Remix and Robo: sampling, sequencing and real-time control of a tangible robotic construction system**	**Hayes Raffle et al.**	H.5.2	H.5.2, K.8.0	
2198	**Agency, tangible technology and young children**	**Peta Wyeth**	H.5.2	H.5.2, K.8.0	
2198	**Managing collaborative activities in project management**	**Shaoke Zhang et al.**	H.5.2	H.5.2, K.8.0	activity centric computing, project management, prototype
2311	**Interactive Continuous Collision Detection Using Swept Volume for Avatars**	**Young J. Kim et al.**	I.3.6	F.2.2, H.5.1, I.3.5, I.3.7	
2310	**A constrained SLAM approach to robust and accurate localisation of autonomous ground vehicles**	Kwang Wee Lee et al.	I.2.9	G.2.2	localisation, Mobile robotics, SLAM Neighbourhood environments, Road map matching,
2317	**Mobile, hardware-accelerated urban 3D maps in 3G networks**	Antti Nurminen	I.3.6	H.4.3, H.5.2	3D maps, VRML, mobile computing, wireless networks
588	**Teaching IT in Health Care and Nursing ProgramExperiences**	**Tatjana Welzer et al.**	K.3.2	J.3, K.4.3	
1166	**Book review**		K.6.1	A.1, J.1	
1419	**Assessing the contributions of business and IT knowledge to the development of IT/business partnerships**	**GeneviÄ¨ve Bassellier, Izak Benbasat**	K.6.1	K.4, K.7	IS/business partnerships, business knowledge, information technology knowledge
1155	**A commodity market algorithm for pricing substitutable Grid resources**	**Gunther Stuer et al.**	K.6.0	K.1	dynamic pricing, Grid, Grid economics

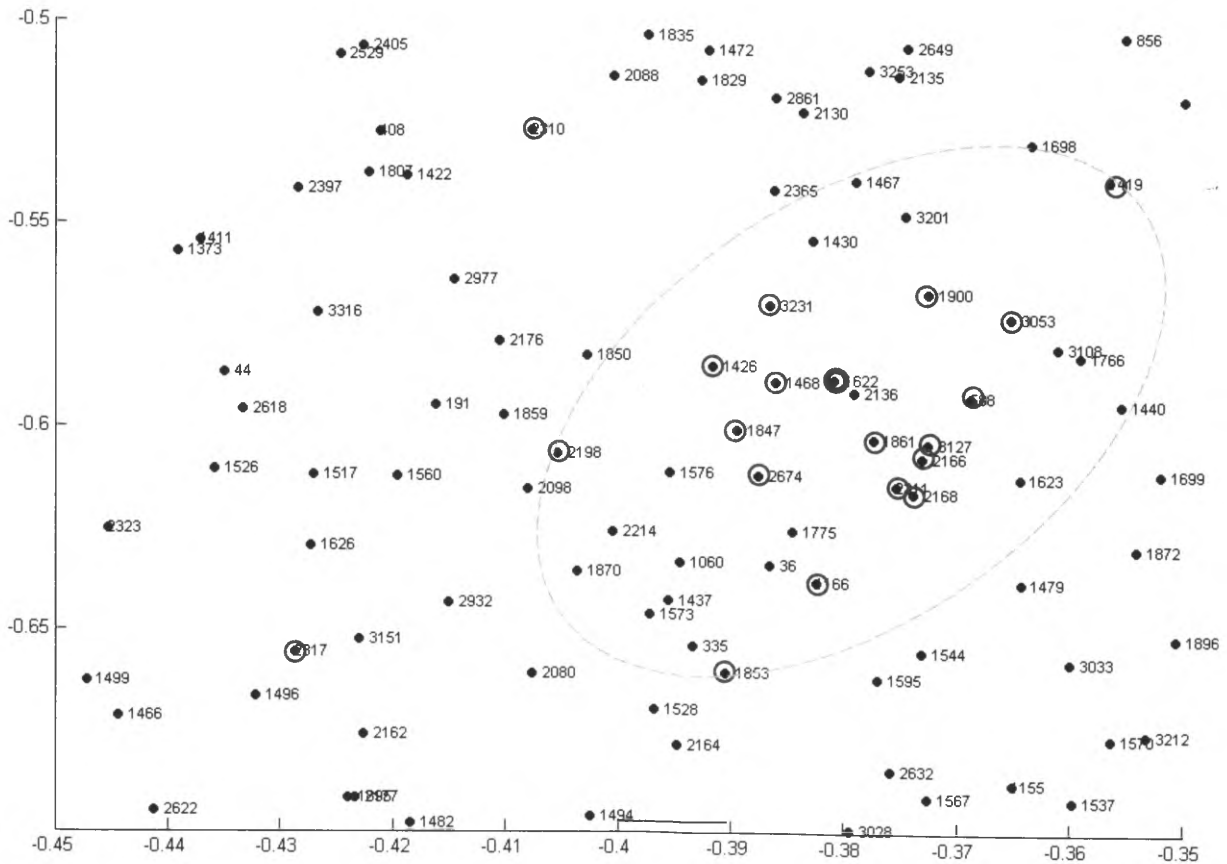

-0.5
-0.55
-0.6
-0.65
-0.45
-0.44
-0.43
-0.42
-0.41
-0.4
-0.39
-0.38
-0.37
-0.36
-0.35
2405
2529
1835
1472
2649
856
2088
1829
3253
2135
2861
2130
1698
1805
1422
2397
2365
1467
3201
1373
1430
2977
3316
3231
1900
3053
2176
1850
3108
1766
44
1426
1468
1622
2136
2618
191
1859
1847
1440
2198
1526
1517
1560
1576
2674
1623
1699
2098
2168
2323
1626
2214
1775
1872
1870
1060
36
66
1479
2932
1437
1573
3151
335
1544
1896
1499
2080
1853
1595
3033
1496
1528
1466
2162
2164
3212
2632
155
2622
1482
1494
1567
1537
3028

Tabela B2. Charakterystyki artykułu nr 1622 i sąsiadujących z nim na mapie wizualizacji przy relacji klasyfikacji podstawowej i dodatkowych jako 0.7:0.3

Nr	Tytuł	Autor(zy)	Klasa gł	Klasy dodatkowe	Słowa kluczowe
1622	**Combining bibliometrics, information retrieval, and relevance theory, Part 2: Some implications for information science: Research Articles**	Howard D. White Drexel University,	H.3.3	H.3.4, I.2.0	bibliometrics, cognitive models, information retrieval, query formulation, relevance
1430	Object identification and retrieval from efficient image matching. Snap2Tell with the STOIC dataset	**Jean-Pierre Chevallet et al.**	H.3.3	F.2.2, H.3.1, H.3.4, I.5.2	image retrieval, mobile information retrieval, object identification, test collection building
1468	**Applying logistic regression to relevance feedback in image retrieval systems**	T. LeÃłn, P. Zuccarello et al.	H.3.3	G.3, H.2.8, H.3.3, H.4.2, I.4.7	content-based image retrieval systems, Logistic regression, Low-level image descriptors, Visual information retrieval
1426	**Parameterized pattern queries**	Casdric du Mouza et al.	H.3.3	F.2.2, H.2.4, H.3.4, I.5.2	optimization, Pattern matching, Representation and manipulation of data
1576	The phrase-based vector space model for automatic retrieval of free-text medical documents	Wenlei Mao, Wesley W. Chu	H.3.3	H.3.1, J.3	Computing methodologies, Concept-based vector space model, Information storage and retrieval/methods, Information systems, Phrase-based vector space model, Unified medical language system, Vector space model
1895	Recommender Systems	Hannes Werthner et al.	H.3.4	H.4.2, K.4.4	Recommender systems give advice about products, information or services users might be interested in. They are intelligent applications to assist users in a decision-making process where they want to choose one item amongst a potentially overwhelming set of alternative products or services.

1560	Combining multimodal preferences for multimedia information retrieval	Eric Bruno et al.	H.3.3	H.3.1, H.3.3, I.2.6	RankBoost, multimedia indexing and retrieval, multimodal fusion
1623	Dynamic personalization of web sites without user intervention	**Ranieri Baraglia, Fabrizio Silvestri**	H.3.3	H.3.4, I.2.11, I.5.1	
1479	Social capital and the search for information: Examining the role of social capital in information seeking behavior in Mongolia: Research Articles	**Catherine A. Johnson**	H.3.3	H.1.1, H.3.4	developing countries, information seeking, social aspects, social networking, trust
1467	A clustering entropy-driven approach for exploring and exploiting noisy functions	**Shih-Hsi Liu et al.**	H.3.3	G.3, H.2.8	cluster, entropy, exploitation, exploration
1853	Web search engine based on DNS	Wang Liang et al.	H.3.4	H.3.3, H.3.5	Distributed system, Domain, Information retrieval, Search engine, Web-based service
3151	Semantic web HCI: discussing research implications	Duane Degler et al.	H.5.2	H.3.3, I.2.11	HCI, OWL, RDF, context, metadata, ontology, semantic web, user interaction, user interface, visualization
2316	ShapePalettes: interactive normal transfer via sketching	**Tai-Pang Wu et al.**	I.3.6	H.1.2, I.3.5, I.3.7	human-computer interaction, image-based modeling, interactive modeling
1318	Model compensation approach based on nonuniform spectral compression features for noisy speech recognition	Geng-Xin Ning et al.	I.2.7	E.4, H.5.5, I.5.2	
1319	A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation	**Mohammad H. Radfar**	I.2.7	E.4, H.5.5, I.5.2, I.5.4	
408	What can Children Learn through Game-based Learning Systems?	Masanori Sugimoto	K.3.1	K.8.0	
409	Information behavior of small groups: implications for design of digital libraries	**Nan Zhou, Gerry Stahl**	K.3.1	K.7.1	CSCL, digital libraries, information behavior

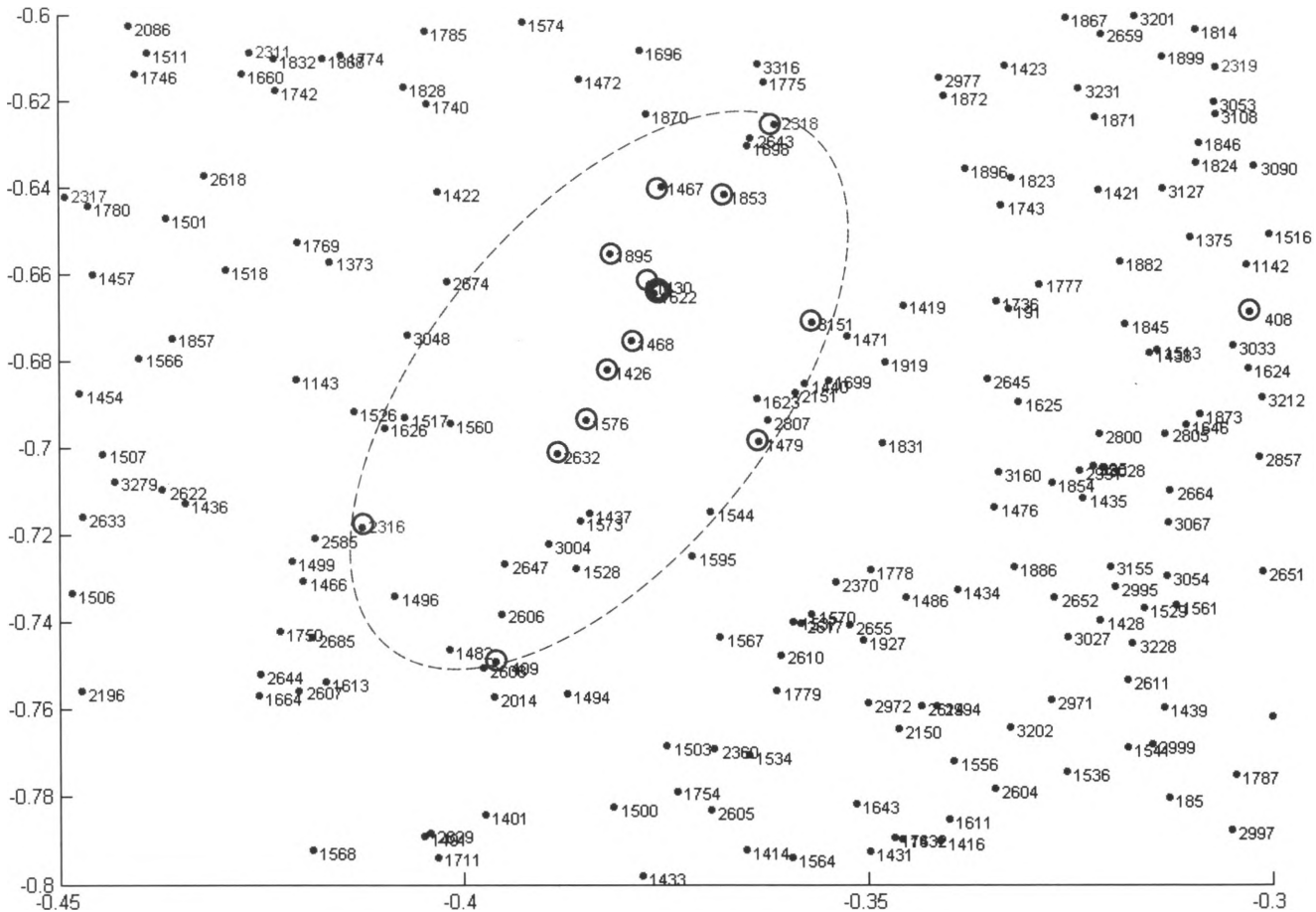

-0.6
-0.62
-0.64
-0.66
-0.68
-0.7
-0.72
-0.74
-0.76
-0.78
-0.8
-0.45
-0.4
-0.35
-0.3

Tabela B3. Charakterystyki artykułu nr 882 i sąsiadujących z nim na mapie wizualizacji przy relacji klasyfikacji podstawowej i dodatkowych jako 0.7:0.3

Nr	Tytuł	Autor(zy)	Klasa gł	Klasy dodatkowe	Słowa kluczowe
882	**Communication over Hypercomplex Kähler Manifolds: Capacity of Multidimensional-MIMO Channels**	Özgür Ertuğ	C.2.1	C.2.2, I.2.8	MIMO, differential geometry and topology, ergodic capacity, information theory, pattern diversity, polarization diversity, space divers.
505	**Autonomous Grid Computing**		C.2.0	C.4, F.2.2, H.3.0, I.2.0, K.6.0	
333	**60-GHz millimeter-wave radio: principle, technology, and new Results**	Nan Guo et al.	C.2.0	C.2.1, I.4.6, I.5.4	
1031	**Experiences from a wireless sensor network deployment in a petroleum environment**	Ian Johnstone et al.	C.2.1	C.3, J.2	embedded systems, heterogeneous networks, wireless sensor networks
1029	**Radio frequency identification: technologies, applications, and research issues: Research Articles**	Yang Xiao et al.	C.2.1	C.3, I.5.4	auto-identification, privacy, radio frequency identification, reader, tag, ubiquitous comp.
567	**Journal of Network and Computer Applications**		C.2.0	F.2.0, I.4.0, I.5.0, K.6.5	
332	**Fuzzy System for DOA Estimation in Mobile Communications using a FPGA**	Alberto Rochin Garcia et al.	C.2.0	C.2.1, I.2.3	
1304	**An Account of Implementing Applicative Term Rewriting**	**Muck van Weerdenburg**	F.4.2	F.2.2, G.2.2	efficient rewriting, nonlinear match trees, open terms
1312	**Electronic Notes in Theoretical Computer Science (ENTCS)**		F.4.2	F.4.1, F.4.3	
1301	**Fast congruence closure and extensions**	Robert Nieuwenhuis, Albert Oliveras	F.4.2	F.2.2, F.4.1, F.4.2	02.10.-v, 07.05.Bx, 84.30.Bv, 89.20.Ff, 95.75.Pq, Congruence closure, Decision procedures, Equational reasoning, Verification
249	**Arithmetic computation in the tile assembly model: Addition and multiplication**	**Yuriy Brun**	F.1.2	I.5.m	Tile assembly model, Adder, Crystal growth, Mol. computation, Multiplier, Self-assembly
4732	**Definition, modelling and simulation of a grid computing scheduling system for high throughput computing**	**Eddy Caron et al.**	I.6.7	C.2.4, D.4.1, I.6.8	Grid/global computing, High throughput computing, Meta-scheduling, Simulation
4733	**A Grid Based Simulation Environment for Mobile Distributed Applications**	**Dawit Mengistu et al.**	I.6.7	C.2.4, I.6.8	
1784	**Enhancing border security: Mutual information analysis to identify suspect vehicles**	**Siddharth Kaza et al.**	K.6.5	H.1.1, K.6.1	Border safety, Intelligence and security informatics, Mutual information

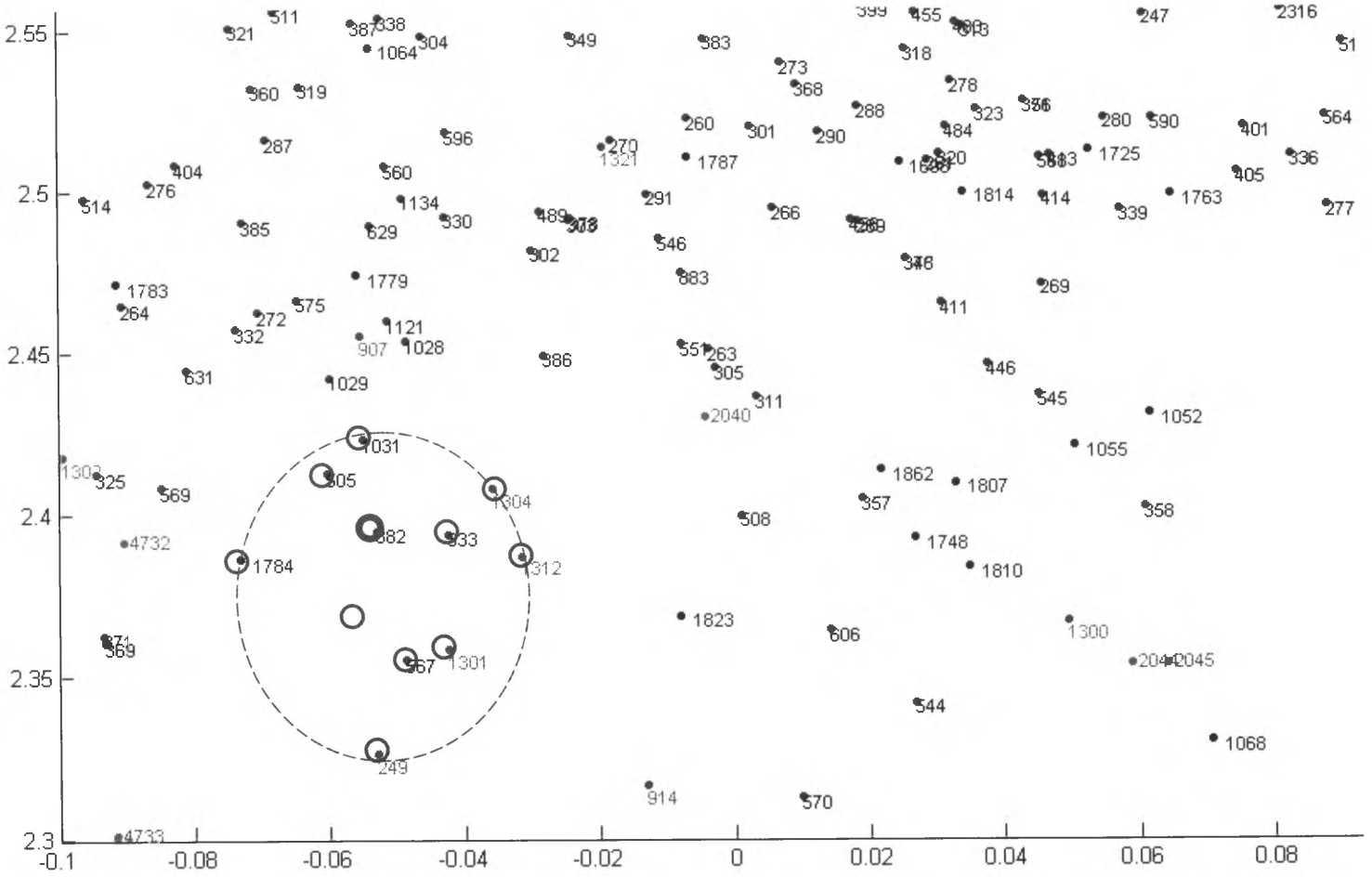
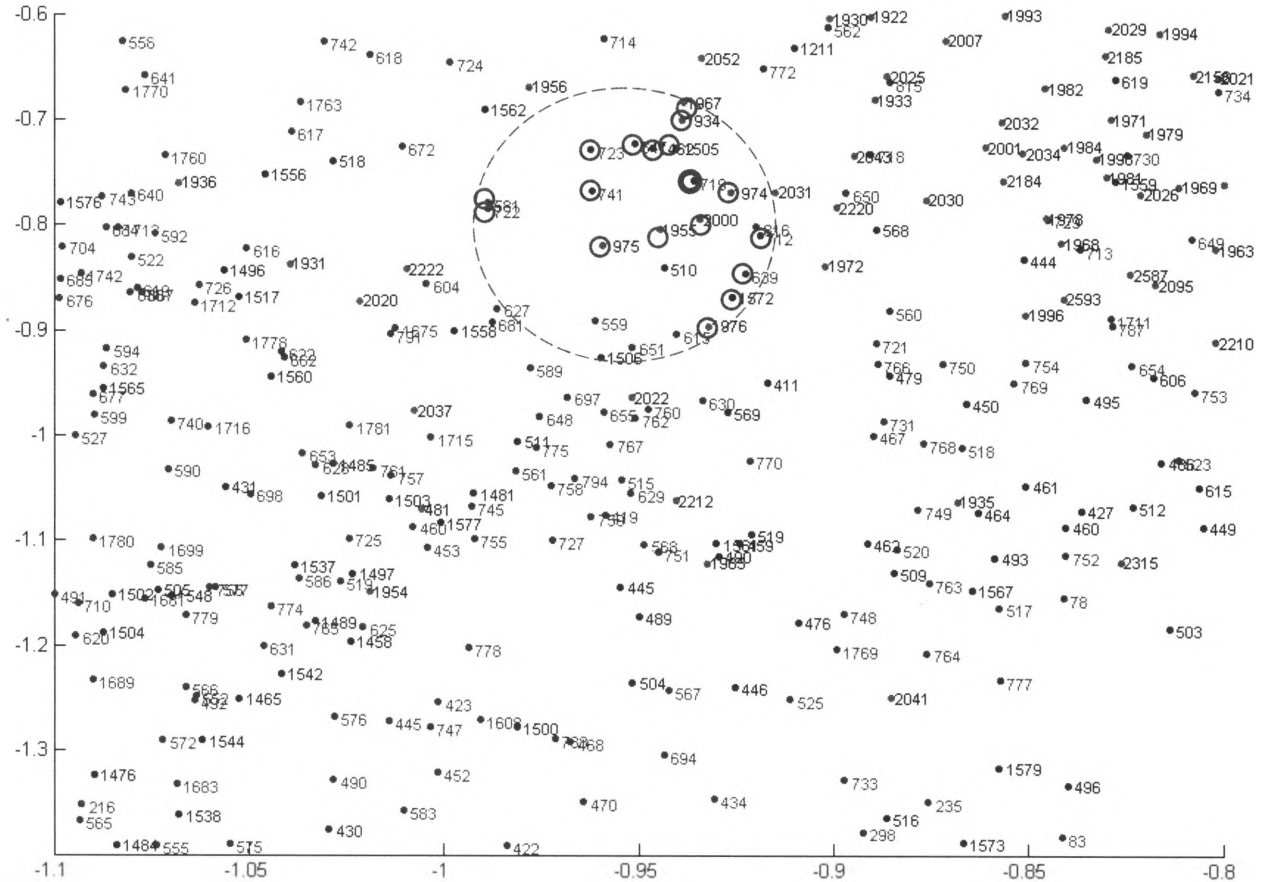

2.55
2.5
2.45
2.4
2.35
2.3
-0.1
-0.08
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08

Tabela B4. Charakterystyki artykułu nr 719 i sąsiadujących z nim na mapie wizualizacji przy relacji klasyfikacji podstawowej i dodatkowych jako 0.7:0.3

Nr	Tytuł	Autor(zy)	Klasa gł	Klasy dodatkowe	Słowa kluczowe
719	**CCLRC Portal infrastructure to support research facilities: Research Articles**	Asif Akram et al.	D.2.11	H.3.4, H.3.5	Grid, Web service, portlet
647	**Advances in Engineering Software**		D.2.11	D.2.2, D.2.3, G.1.8, I.6.5, I.6.6	
723	**A Context-Aware Architecture for Smart Space Environment**	E. Goh et al.	D.2.11	H.3.5, H.4.1	
581	**SLA-Driven Clustering of QoS-Aware Application Servers**	Giorgia Lodi et al.	D.2.11	C.5.5, D.2.5, H.3.5, I.5.3	Service Level Agreement, Quality of Service, QoS-aware application server, QoS-aware cluster, dynamic cluster configuration, monitoring, load balancing.
722	**At the forge: RJS templates**	Reuven M. Lerner	D.2.11	H.3.5	
714	**Situational computing: An innovative architecture with imprecise reasoning**	**C. B. Anagnostopoulos et al.**	D.2.11	H.1.2, I.2.3, I.2.4	Context awareness, Fuzzy and approximate reasoning, Human-computer interaction, Ontological modeling, Pervasive computing, Situational context representation
712	**Erratum: A highly modular and extensible architecture for an integrated IMS-based authoring system: the e-Aula experience**	José Luis Sierra et al.	D.2.11	H.1.1, K.6.4	
639	**An integration architecture for knowledge management systems and business process management systems**	**Jisoo Jung et al.**	D.2.11	D.2.13, H.2.8, H.3.2, H.3.5, J.1	Business process management, Knowledge management, Process knowledge management, Workflow
741	**Towards the architectural definition of the Health Watcher system with AO-ADL**	**Monica Pinto et al.**	D.2.11	J.3, K.6.3	

1955	**Measuring Reliability of Applications Composed of Web Services**	Hangjung Zo et al.	H.3.5	C.4, D.2.13	
1975	**Towards the theoretical foundation of choreogra-phy**	**Zongyan Qiu et al.**	H.3.5	D.2.1, D.3.1	choreography, dominant role, dominated choice, dominated loop, implementation, projection, semantics
1974	**Web Engineering Security: Essential Elements**	**William Bradley Glisson, Ray Welland**	H.3.5	D.2.0, K.6.3, K.6.5	
2000	**MDA-based Automatic OWL Ontology Development**	**Dragan Gašević et al.**	H.3.5	D.2.12, D.3.2, I.7.2	model Driven Architecture, OWL, Ontology development, UML Profile, XSLT
1934	**SenseWeb: An Infrastructure for Shared Sensing**	**Aman Kansal et al.**	H.3.5	C.2.1, C.2.3, H.3.5	sensor network, sensor gateway, Web2.0, and peer-produced systems.
1967	**Foundations of Microsoft Expression Web: The Basics and Beyond**	**Cheryl D. Wise**	H.3.5	D.1.0, D.2.6, I.7.2, K.6.3	Expression Web is a standard-compliant Web editor from Microsoft. A Web server with the ASP.NET 2.0 framework is required to use its advanced features. This tutorial book is a practical guide for designing Web sites using Expression Web.
1976	**Adaptive web service composition**	**Fernando Antônio Aires Lins et al.**	H.3.5	D.2.1, F.3.1	
1462	**Dimensionality Reduction for the Control of Powered Upper Limb Prostheses**	**Klaus Buchenrieder**	K.6.3	C.3, I.5.2, I.5.3, I.5.4	
1505	**A Case Study: Applying Lyra in Modeling S60 Camera Functionality**	**Jukka Honkola e al.**	K.6.3	D.2.2, I.6.5	
1572	**Collaborative Process Improvement: With Examples from the Software World (Practitioners)**	**Celeste Yeakley, Jeff Fiebrich**	K.6.3	K.4.3	Organizations willing to initiate a process improvement effort will still need information about how to do so. This book seeks to provide that information.


-0.6
-0.7
-0.8
-0.9
-1
-1.1
-1.2
-1.3
-1.1
-1.05
-1
-0.95
-0.9
-0.85
-0.8

Spis ilustracji

Ilustracja 1. Fragment macierzy podobieństwa 87

Ilustracja 2. Zrzut ekranowy aplikacji z wynikami wizualizacji 92

Ilustracja 3. Mapa wizualizacji wszystkich klas (wyświetlone symbole dla klas o dużej ilości dokumentów) 94

Ilustracja 4. Podstawowa mapa wizualizacji dokumentów (wraz z klasami) 94

Ilustracja 5. Mapa wizualizacji dokumentów klas A, B, C, D 95

Ilustracja 6. Mapa wizualizacji zmodyfikowanego zestawu danych (bez klasy I) 97

Ilustracja 7. Mapa wizualizacji dla relacji klasyfikacji głównej do klasyfikacji dodatkowej jako 0.5:0.5 98

Ilustracja 8. Mapa wizualizacji dla relacji klasyfikacji głównej do klasyfikacji dodat-kowej jako 0.7:0.3 98

Ilustracja 9. Mapa wizualizacji po obróbce graficznej 100

Ilustracja 10. Mapa słów kluczowych 101

Ilustracja 11. Mapy słów kluczowych kombinacji różnych klas 101

Ilustracja 12. Mapy klasyfikacji CCS z roku 1968 116

Ilustracja 13. Mapy klasyfikacji CCS z roku 1978 117

Ilustracja 14. Mapy klasyfikacji CCS z roku 1988 118

Ilustracja 15. Mapy klasyfikacji CCS z roku 1988 dla wybranych kombinacji klas 121

Ilustracja 16. Mapy klasyfikacji CCS z roku 1998 124

Ilustracja 17. Mapy klasyfikacji CCS z roku 1998 dla wybranych kombinacji klas 125

Ilustracja 18. Identyfikacja wyszukanych obiektów na mapie wizualizacji 134

Indeks rzeczowy

A

ACM 57

C

CCS 57
cloud computing 59
co-citation analysis 38
co-descriptors 43
co-keywords 43
Computing Classification System – Patrz CCS
co-words 43

D

dendrogram 28
desaturacja 110

F

faseta 9
fisheye 34
focus+context 34
fraktal 106

G

glyf 26
graf 28

I

infografika 16
Infovis – Patrz wizualizacja informacji
instancja 66
inżynieria ontologiczna 66

K

ko-klasa 77

L

lakunarność 111

M

mapowanie nauki 38
MDS

N

naukografia – Patrz mapowanie nauki

P

paradygmat nauki 39

R

renderowanie 17

S

semantyka wektorowa 33
systemy eksperckie 66

T

treemap 29

W

wektor cech 33
Wizualizacja edukacyjna 20
wizualizacja informacji 20
wizualizacja nauki 38
wizualizacja naukowa 16
wymiar fraktalny 106

Wizualizacja informacji (Information Vizualization – Infoviz) staje się nowym polem badawczym w informacji naukowej, gdyż nowoczesne systemy informacyjne muszą dostarczać coraz bardziej wymagającym użytkownikom alternatywnych sposobów prezentacji danych w czytelnej i kognitywnej formie graficznej, jak również efektywnego ich wyszukiwania i/lub uporządkowania.

Autorka traktuje wizualizację jako narzędzie i metodę interpretacji danych i na tej postawie generuje obrazy ze zbiorów danych wielowymiarowych. Kolorowe wersje wszystkich map wizualizacyjnych i wybranych rysunków z tekstu czytelnik może obejrzeć pod adresem www.umk.pl/~wieo/Infoviz.

Na polskim rynku brakuje przeglądowych pozycji literaturowych z zakresu wizualizacji informacji. Infoviz – to tematyka interdyscyplinarna, obejmująca jednocześnie wiele dziedzin wiedzy, dotycząca zarówno aspektów poznawczych, badań w zakresie informacji naukowej, statystyki, elementów inteligencji obliczeniowej, metod klasyfikacji i analizy skupień oraz metod wizualizacji wielowymiarowych danych i analizy informacji tekstowych.

„...(pracę) można uznać za jedną z niewielu w polskiej literaturze przedmiotu zawierającą tak obszerną i dogłębną analizę tej problematyki. Wychodzi ona naprzeciw od dawna podnoszonym w gronie specjalistów informacji naukowej postulatom zajęcia się zagadnieniami dotyczącymi prezentacji informacji w tych systemach. Autorce udało się dokonać trafnej analizy oraz precyzyjnie opisać i ocenić wartość stosowanych obecnie metod wizualizacji. Novum pracy stanowi niewątpliwie propozycja nowej metody wizualizacji wypełniająca istniejącą lukę w tym zakresie".

dr hab. Wiesław Babik

Seria wydawana przez Wydawnictwo
STOWARZYSZENIA BIBLIOTEKARZY POLSKICH
we współpracy
Z INSTYTUTEM INFORMACJI NAUKOWEJ
I STUDIÓW BIBLIOLOGICZNYCH
UNIWERSYTETU WARSZAWSKIEGO