TORUŃSKIE STUDIA BIBLIOLOGICZNE
2008, nr 1

Veslava Osińska
Instytut Informacji Naukowej i Bibliologii
Uniwersytet Mikołaja Kopernika w Toruniu
e-mail: wieo@umk.pl

Wizualizacja i mapowanie przestrzeni danych w bibliotekach cyfrowych

Otaczająca nas rzeczywistość i procesy w niej zachodzące zdeterminowane są przez informację cyfrową, generowaną przez komputery, liczniki, sterowniki i inne urządzenia elektroniczne. Szybki rozwój technologii sieciowych w ostatniej dekadzie spowodował zwiększenie zdolności gromadzenia i przetwarzania rozproszonych zasobów informacyjnych. Mamy do czynienia nie tylko z ogromną ilością danych, lecz także większą ich złożonością. Dobrym przykładem są dane wykorzystywane w diagnostyce medycznej, szczególnie wszystkie metody obrazowania, takie jak tomografia komputerowa, funkcjonalny rezonans magnetyczny[1] czy wieloparametryczne bazy danych klinicznych, powiązane często z innymi równie licznymi systemami archiwizacji i analizy danych. Inne przykłady to zbiory danych mikro- i makroekonomicznych oraz modele dynamicznych zmian w strukturach gospodarczych i ekonomicznych.

Jeśli dane powstające w wyniku różnych operacji numerycznych i analiz statystycznych charakteryzują się wieloma parametrami (cechami), które pozwalają opisać ich zależności, to mówi się o wielowymiarowości danych. Wizualizacja wielowymiarowych danych staje się jednym z ważniejszych czynników decydujących o właściwym zrozumieniu da-

[1] Funkcjonalny rezonans magnetyczny służy do obrazowania czynnościowego, ujawniając patologię na poziomie komórki, a nawet genomu.

nych, informacja w postaci liczb nie jest bowiem naturalna dla ludzkiego systemu percepcji.

Mózg człowieka ze swoim systemem percepcji wizualnej jest zawsze „stacją końcową" wszystkich metod wizualizacji. To właśnie w nim odbywa się najważniejsza część analizy danych, tak aby była jak najbardziej przydatna w kolejnych procesach kognitywnych[2]. Wizualizacja dwuwymiarowa jest najbardziej naturalnym sposobem przedstawiania wyników, gdyż zazwyczaj umieszcza się ją na dwuwymiarowej kartce papieru lub płaszczyźnie monitora komputerowego. Jednak tabele z dużą liczbą kolumn zawierających liczby są 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 ma wyspecjalizowane obszary do analizy krawędzi poziomych, pionowych i pochyłych, to już wizualizacje przedstawiające rozproszone zbiory punktów, bez ich wcześniejszej konglomeracji, stają się często nic nieznaczącym szumem. Większy problem pojawia się w sytuacji, gdy chcemy na dwuwymiarowym wykresie przedstawić zmienne wielowymiarowe. Wówczas bez specjalnego treningu i często skomplikowanego opisu pomocniczego nie możemy jednoznacznie określić korelacji między cechami. Z pomocą przychodzą nam metody wizualizacji oparte na bezpośrednim przetwarzaniu obrazów[3]. Można wtedy uzyskać wizualizację 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 – można otrzymać bardzo dobre własności percepcyjne prezentowanych wizualizacji. Jeśli zostaną do tego dołączone metody rekonstrukcji trójwymiarowej, to powstają już dwa zbiory niezależnych parametrów – wymiary przestrzenne i kolory, które można z powodzeniem wykorzystać do prezentacji zależności korelacyjnych między dużymi zbiorami danych.

Termin *wizualizacja informacji* rozumiany jest jako wizualna prezentacja przestrzeni informacyjnych i struktur w celu ułatwienia ich szybkiego przyswojenia i zrozumienia. W rzeczywistej (nieabstrakcyjnej) repre-

[2] W. Duch, *Umysł, świadomość i działania twórcze* [on-line]. Kognitywistyka.net [dostęp 30 września 2008]. Dostępny w World Wide Web: http://www.kognitywistyka.net/artykuly/wd-ust.pdf.

[3] C. Ware, *Information Visualization. Perception for Design*, San Francisco 2004, s. 5–22.

zentacji 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. Dlatego wizualizację danych, która jest skoncentrowana wokół danych mierzalnych, takich jak np. wyniki medycznych badań ludzkiego ciała lub dane geograficznych systemów informacyjnych (ang. *Geographic Information System*, GIS), należy odróżniać od wizualizacji informacji, zajmującej się danymi nierzeczywistymi, czyli np. tekstem lub strukturami hierarchicznymi. InfoVis (skrót popularnie stosowany w literaturze naukowej i biznesowej od ang. *Information Visualization*) to dyscyplina, która poszukuje nowych metafor graficznych w celu przedstawienia informacji niemającej naturalnej i oczywistej reprezentacji. Wizualizacja informacji jest stosunkowo młodą dyscypliną badawczą o 10-letniej historii, ale szybko się rozwija. InfoVis wykorzystuje osiągnięcia takich pokrewnych dziedzin, jak: data mining, analiza danych, interakcja człowiek–komputer (ang. *Human–Computer Interaction*, HCI) i grafika komputerowa.

„Proces wizualizacji wiąże dwa najpotężniejsze systemy przetwarzania informacji – ludzki mózg oraz współczesny komputer”[4]. Oddziaływanie między człowiekiem i komputerem pojawia się na poziomie interfejsu użytkownika, aktualnie obserwuje się więc wzmożony rozwój graficznych interfejsów użytkownika (ang. *Graphic User Interface*, GUI). Celem projektów wizualizacyjnych wydajnych GUI jest realizacja takich zadań, jak: obserwacja, wyszukiwanie, nawigacja, rozpoznanie, filtrowanie, odkrywanie, rozumienie oraz interakcja z dużymi zbiorami danych.

„Oko wypatruje podobne obiekty, aby je porównać, dokonuje ich analizy pod różnym kątem i perspektywy, aby dopasować ich elementy składowe”[5]. Wizualizacja może być jednym z etapów procesu analitycznego, jeśli pozwala na szybkie wykrycie związków pomiędzy poszczególnymi cechami lub ujawnienie nieprawidłowych wartości cech. Taka analiza wizualna koncentruje się na procesach rozumowania i odkrywania sensu danych[6]. Techniki wizualizacji są stosowane jako jedna ze skuteczniejszych form eksploracji danych (ang. *data mining*).

---

[4] V. Osińska, *Przybliżenie semantyczne wizualizacji informacji w Internecie*. Biuletyn EBIB [on-line] 2006, nr 7 (77) [dostęp 30 września 2008]. Dostępny w World Wide Web: http://www.ebib.info/2006/77/osinska.php.

[5] K. Narojczyk. *Komputerowa wizualizacja danych historycznyc*h, [w:] *Megabajty dziejów. Informatyka w badaniach, popularyzacji i dydaktyce historii*, pod red. R. T. Prinke, Poznań 2007, s. 79–95.

[6] Tamże.

Prezentacja danych jest już pełnoprawną częścią sztuki filmowej i reklamy. Wystarczy spojrzeć na filmy science fiction lub szpiegowskie i migające w nich komunikaty, kolorowe okienka, wykresy. Nowoczesne środki grafiki komputerowej pozwalają na estetyczną, przystępną, często zaskakującą prezentację danych, informacji i wiedzy. W związku z tym powstał nowy termin (synonim InfoVis) *infografika,* czyli grafika informacyjna. Stosowane w aplikacjach techniki wizualizacyjne uwarunkowane są przede wszystkim rodzajem informacji, jej wymiarem i poziomem abstrakcji. Do klasycznych technik zalicza się podstawowe formaty wizualne pośredniczące w przekazywaniu wiedzy, np. diagramy, listy, wykresy, tabele i macierze.

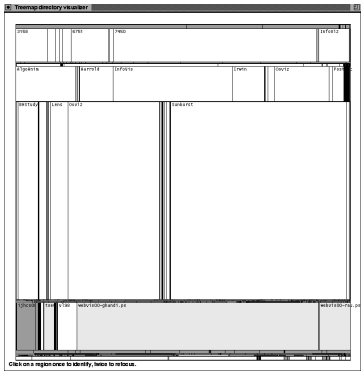
Najprostszym typem informacji jest informacja liniowa, składająca się z sekwencji liter i cyfr. Dane w postaci różnego rodzaju list i tabel, powszechne w historii piśmienniczej i obliczeniowej ludzkiej działalności, znane są jeszcze z czasów starożytnych. Znaki alfanumeryczne trudno jest przedstawić w innej formie niż tekst. Nie przeszkadzało to jednak, aby w latach 90. 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 i w ten sposób powstawały kolorowe spektra, przedstawiające zależności co najwyżej dwóch wartości.

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

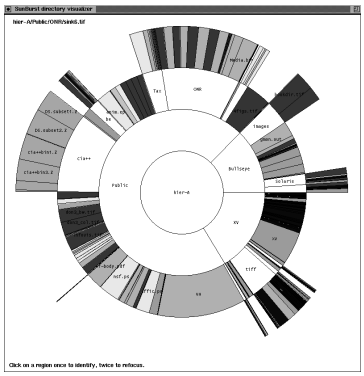
Na początku lat 90. szybkość procesorów nie nadążała za dynamiką zwiększania zasobów na twardych dyskach. Dlatego inżynierowie i naukowcy intensywnie poszukiwali nowych, wydajnych metod wizualizacji struktur katalogowych dla systemów unixowych. Drzewa hierarchiczne przedstawiano nie w postaci gałęzi, lecz map – topologię jednowymiaro-

wą poszerzono do dwóch wymiarów. Generację oprogramowania służącego do takich zadań, zapoczątkowanego przez Bena Shneidermana, nazwano Treemap[7]. Idea opierała się na zagnieżdżaniu prostokątów mniejszymi prostokątami o polach proporcjonalnych do pojemności zasobów folderów (rys. 1a). Kolejnym pomysłem na przeniesienie struktury drzewa katalogowego na dwuwymiarową przestrzeń jest schemat hierarchii kreślony za pomocą koncentrycznych pierścieni, jak np. w programie SunBurst autorstwa Johna Staska[8]. Katalog główny znajduje się w środkowym kole mapy. Segmenty kolejnych kół reprezentują podkatalogi z ich zawartością. Ogólna pojemność katalogu i typ pliku identyfikowane są odpowiednio za pomocą kąta segmentu i koloru (rys. 1b).

a.



b.



Rysunek 1. Strategia prostokątna (a) i pierścieniowa (b) wizualizacji zasobów katalogowych

Źródło: J. Stasko, R. Catrambone, M. Guzdial, K. McDonald, *An evaluation of space-filling information visualizations for depicting hierarchical structures*, „International Journal of Human–Computer Studies" 2000, vol. 53, iss. 5, s. 667–668.

W odróżnieniu od dendrogramu, w strukturach sieciowych powiązania istnieją nie tylko w kierunku góra–dół, lecz także pomiędzy węzłami równorzędnymi. Systemy danych geograficznych, a również wiele domen rzeczywistości powszechnie przedstawia się za pomocą węzłów i wekto-

[7] *Treemap Home Page* [on-line]. The Human-Computer Interaction lab. University of Maryland. Institute of Advanced Computer Studies [dostęp 30 września 2008]. Dostępny w World Wide Web: http://www.cs.umd.edu/hcil/treemap/.

[8] *SunBurst Page* [on-line]. Georgia Institute of Technology. Information Interfaces Research Lab [dostęp 30 września 2008]. Dostępny w World Wide Web: http://www.gvu.gatech.edu/ii/sunburst/.

rów, czyli grafów. Ich 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.

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 statystyczno-lingwistycznych algorytmów uczenia się maszynowego i sztucznych sieci neuronowych, aby na bieżąco określić tematyczne kategorie zasobów[9]. Przy modulowaniu reprezentacji semantycznych w zadaniach filtrowania i wyszukiwania informacji wykorzystywany jest wektorowy model przestrzeni wielowymiarowej. Dokumenty są przedstawiane w sposób formalny przy użyciu wektorów cech, za które mogą posłużyć np. słowa kluczowe, sekwencje słów, odległość pomiędzy wyrazami, występowanie spójników, topologia obiektów, formaty i rozmiary plików itp.

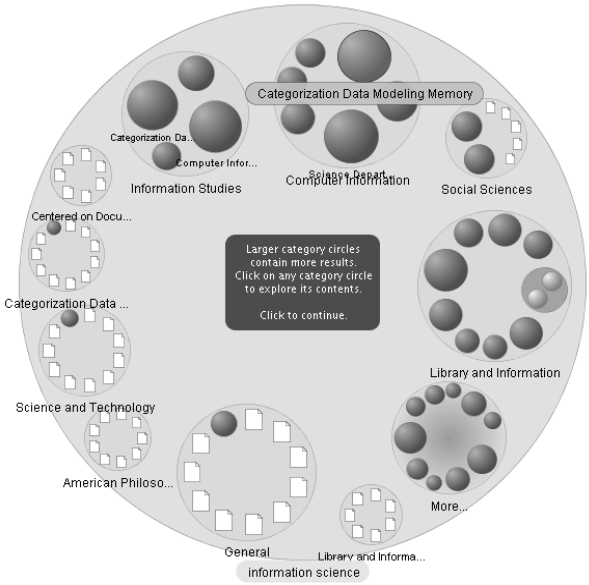
W przeglądaniu i wyszukiwaniu danych dużą rolę w aplikacji odgrywa przestrzeń eksploracyjna, którą ogranicza okno monitora. Sposobem na jej poszerzenie jest reprezentacja hierarchicznych struktur w przestrzeni hiperbolicznej. Pierwsze aplikacje, które wykorzystały technikę tzw. rybiego oka (ang. *fisheye*), to 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, czyli ze wzrostem odległości mamy eksponencjalne powiększenie przestrzeni).

Z chwilą sukcesu Google firmy komercyjne intensywnie rozwijające oprogramowanie wizualizacyjne dla zasobów Internetu, takie jak: KartOO, Groxis, Medialab Solutions, The Brain Technologies, Vivisimo, zaczęły profilować swoje przeglądarki w kierunku integracji zadań wyszukiwania, filtrowania i nawigacji[10]. W zależności od koncepcji projektantów i zastosowanych metafor wizualizacji użytkownik ma zapoznać się nie

[9] J. Kozłowski i in., *Wspomaganie wyszukiwania dokumentów mapami samoorganizującymi*. W: *Materiały z III Krajowej Konferencji „Multimedialne i Sieciowe Systemy Informacyjne"* [on-line]. Wrocław 2002 [dostęp 30 września 2008]. Dostępny w World Wide Web: http://www.zsi.pwr.wroc.pl/zsi/missi2002/pdf/s507.pdf.

[10] V. Osińska, dz. cyt.

z listą rankingową, lecz z wielowymiarową przestrzenią nawigacyjną. Zgodnie z założeniem większej swobody w nawigacji powinien on również mieć możliwość kolekcjonowania wyselekcjonowanych elementów. W takich wielowymiarowych mapach odrębne znaczenie przyjmują kolor, kształt, rozmiar, pozycja oraz połączenia obiektów (por. rys. 2).



Rysunek 2. Wygenerowana mapa skojarzonych tematycznie obszarów z wyrażeniem „information science" w wyszukiwarce Grokker

Źródło: *Information science* [on-line]. Grokker – Enterprise Search Management and Content Integration [dostęp 30 czerwca 2008]. Dostępny w World Wide Web: http://live.grokker.com/grokker.html?query= information%20science&Yahoo=true&Wikipedia=true&numResults=250.

Przeglądarki semantyczne nowej generacji, takie jak KartOO, TheBrain, Grokker, AquaBrower, ThinkMap, są projektowane z uwzględnieniem ostatnich osiągnięć w kognitywistyce i neuroscience. W projektowaniu interfejsów wizualizacyjnych jest przydatne badanie zależności pomiędzy ludzką percepcją a semantycznym wyszukiwaniem i przeglądaniem informacji.

W automatycznej klasteryzacji pozycjonowanie dokumentów zachodzi w kierunku dół–góra: od najniższego poziomu do najwyższego. Próba wizualizacji takiej struktury w przestrzeni trójwymiarowej prowadzi

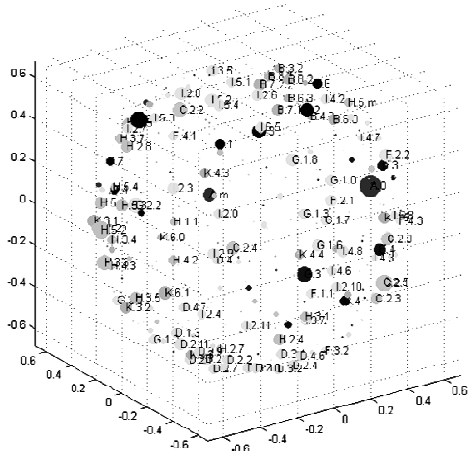
do umieszczenia głównego węzła w centrum, a węzłów podrzędnych we wszystkich kierunkach wokół środka (program Walrus[11]). Lecz można spotkać rozwiązania całkiem odmienne. Ponieważ w przypadku badań korelacyjnych same wartości liczbowe mają mniejsze znaczenie, można z nich zrezygnować i wykorzystać naturalne metryki odległości zamiast parametrów numerycznych. Spowoduje to stworzenie samych już tylko wielobarwnych map rozłożonych w przestrzeni, których percepcja wykorzystuje najbardziej naturalne drogi przetwarzania w ludzkim mózgu, używając również najstarszego i z tego względu najbardziej stabilnego, limbicznego szlaku przetwarzania informacji. Prace nad takim systemem warto rozpocząć od metod wizualizacji danych wielowymiarowych w postaci kolorowych map na powierzchni sfery o zadanej metryce przestrzennej. Pozwoli to w naturalny sposób wykorzystać własności percepcyjne kory wzrokowej w ludzkim mózgu i ograniczyć jednocześnie obszar dozwolonego rozpraszania informacji do powierzchni kuli. Wbrew pozorom taka metoda wizualizacji jest bardzo pojemna topologicznie, gdyż pozwala również na wstępne ukrywanie nieistotnych w danym procesie analizy danych związków korelacyjnych pod powierzchnią kuli, bez tracenia jakości opracowywanych danych. W przyszłości te niewykorzystane parametry będzie można użyć do uzupełnienia wizualizacji poprzez dodanie parametrów ruchu całego obiektu, obrotów kuli – lub tylko elementów składowych wizualizacji – kolorowych map na powierzchni. Taka właściwość, jak symetria sfery, wspomagająca interaktywną nawigację dokumentów, stanowiła czynnik decydujący przy wyborze tej metody do przeprowadzenia indywidualnych badań.

Dane testowe pochodziły z biblioteki cyfrowej The ACM Digital Library[12] (ACM od Association for Computing Machinery). Artykuły o tematyce informatycznej zostały podzielone za pomocą systemu klasyfikacji komputerowej (ang. *Computing Classification System*, dalej: CCS). Jeśli podklasy różnych klas głównych tematycznie się pokrywały, to dokument występował w obu węzłach. Założono, że liczba przypisanych kategorii dla danego dokumentu jest proporcjonalna do podobieństwa pomiędzy odpowiednimi klasami. Im bliższe podobieństwo semantyczne pomiędzy

[11] *Walrus – Graph Visualization Tool* [on-line]. Cooperative Association for Internet Data Analysis (CAIDA) [dostęp 30 września 2008]. Dostępny w World Wide Web: http://www.caida.org/tools/visualization/walrus/.

[12] *ACM Digital Library* [on-line]. The ACM Portal [dostęp 30 września 2008]. Dostępny w World Wide Web: http://portal.acm.org/dl.cfm.

klasami (podklasami), tym więcej zawierają one wspólnych dokumentów. I odwrotnie – dokumenty nie powtarzały się w klasach o odmiennej tematycznie zawartości. Obliczenie, a następnie normalizacja powtarzających się dokumentów (czyli elementów najniższego poziomu w klasyfikacji) dla każdej możliwej pary klas i podklas dały możliwość szybkiego skonstruowania macierzy podobieństwa o rozmiarze równym liczbie wszystkich występujących klas i podklas w kolekcji dokumentów, czyli 353.



A. Literatura ogólna
B. Hardware
C. Zorganizowane systemy komputerowe
D. Software
E. Dane
F. Teoria obliczeń
G. Matematyczne metody obliczeń
H. Systemy informacyjne
I. Metodologie obliczeniowe
J. Zastosowanie komputerów
K. Środowisko obliczeniowe

Rysunek 3. Klasy główne klasyfikacji CCS oraz ich graficzna reprezentacja na sferze (wg metody własnej)

Źródło: opracowanie własne.

Następnie, aby zmniejszyć liczbę wymiarów macierzy do trzech i obserwować ułożenie klas na powierzchni sfery, użyto wykresu skalowania wielowymiarowego (ang. *multidimensional scaling*, MDS)[13]. Taka reprezentacja (por. rys. 3) pozwoliła na graficzne odwzorowanie takich atrybutów klas, jak: indeks, poziom, liczebność i stopień podobieństwa za pomocą koloru, rozmiaru, położenia i przezroczystości klastrów na powierzchni sfery. Wyniki wizualizacji pokazują, że wybrany sposób reprezentacji na powierzchni sfery klas i podklas klasyfikacji CCS jest właściwy, ponieważ z dostateczną precyzją odwzorowuje odległości tematyczne pomiędzy klasami. Metryka semantyczna drzewa klasyfikacji została zmapowana do przestrzeni sfery. W miarę ewolucji podklas prze-

[13] Skalowanie wielowymiarowe – technika statystyczna mająca na celu wykrycie niewidocznych na pierwszy rzut oka korelacji pomiędzy danymi.

strzeń dokoła odpowiedniego klastra będzie gęściej wypełniana. Środowisko aplikacji pozwala na takie operacje, jak powiększenie fragmentów przestrzeni eksploracyjnej, obracanie sfery czy podgląd indeksów klas. Topologię klas i dokumentów obserwować można z różnej perspektywy, co stwarza możliwość natychmiastowego ich porównania i wykrycia korelacji pomiędzy nimi. Osobnym zagadnieniem jest już sama analiza map kolorów powstałych na powierzchni kuli, która może zostać wykorzystana również w dalszych procesach analizy poprzez obliczenie innych cech dodatkowych, takich jak wymiar korelacyjny lub fraktalny.

W niniejszym artykule starano się wykazać, że nowoczesne techniki wizualizacji są skutecznie implementowane w interfejsach aplikacji służących zarówno do przeglądania, nawigacji, wyszukiwania dużych zbiorów niejednorodnych pod względem formatu, struktury i języka danych, jak i zarządzania nimi. Semantyczna reprezentacja rozwiązuje także problemy obecne od wieków w lingwistyce, tj. synonimię i polisemię.

Konieczność unifikacji i formalizacji dotychczasowej zelektronizowanej wiedzy stała się na tyle wyraźna, że obecnie obserwujemy, jak kolejna generacja Internetu rozwija się nie tylko w kierunku semantyki, lecz także analizy wizualnej[14]. Wyzwaniem dla specjalistów InfoVis jest wynalezienie nowych, intuicyjnych form metafor w reprezentacji informacji. Dzięki nowatorskim technikom wizualizacji możliwe będzie głębsze rozumienie zasobów sieciowych w połączeniu z możliwością wnioskowania na podstawie relewantnych faktów i powiązań.

Methods of Scientific Information Visualization in Digital Libraries

Abstract

The article presents main methods of visualization in the last decade. InfoVis *(Information Visualization*) techniques must assure user the data browsing and retrieval as well as an easy navigation, recognizing, exploring and filtering data. The human perceptual processing model was also described. In order to visualize the hierarchy of Computing Classification System we used own method based on data mapping into a sphere.

[14] Por. *Visualizing the Semantic Web.XML – Based Internet and Information Visualization*, ed. by V. Geroimenko and Ch. Chen, London 2006.