PROBLEMY–BADANIA–PRZEGLĄDY

Veslava OSIŃSKA

Uniwersytet Mikołaja Kopernika, TORUŃ

Visual mining czyli eksploracja informacji za pomocą graficznych reprezentacji

W odpowiedzi na zalew informacji, w szczególności medialnej, użytkownicy chętnie preferują jej formy wizualne ze względu na właściwości percepcyjne i analityczne. Trudno o systematykę technik wizualizacyjnych, których zróżnicowanie wynika z faktu, że ta metodologia mająca korzenie w naukach komputerowych wykształciła się z wielu kierunków badawczych, m.in.: analiza danych, statystyka, data mining, grafika komputerowa, interakcja człowiek-komputer, kognitywistyka itp. Na rozwój wizualizacji wpłynęły wzrost mocy obliczeniowej komputerów, włączając najnowsze rozwiązania wydajnych kart graficznych oraz nieliniowy przyrost zasobów sieciowych wolnego dostępu. W artykule przedstawione są wyniki wieloaspektowych analiz zbioru dokumentów na podstawie wizualizacji metadanych. Autorka pokazuje, jak zastosowanie metod wizualnych dostarcza nowych perspektyw w analizie i interpretacji danych naukowych, jak mapy wizualizacyjne wspomagają eksplorację, jak również wyszukiwanie badanych dokumentów. Zaprojektowany odpowiednio interfejs aplikacji zapewnia sprzężenie zwrotne, co umożliwia użytkownikowi sterowanie procesem wizualizacji.

***Visual mining as data exploration using visualization maps**. Currently users prefer visual information in order to deal with the flood of information, particularly regarding to medial communication. Visual form of data has to facilitate their perception and analysis. It is difficult to classify of information visualization and visual data mining techniques because it evolved from interdisciplinarity. Having roots in computer science, visualization overlaps with data analysis, data mining, statistics, computer graphics, human-computer interaction, cognitive science. Recent achievements in information science and computer science together with the increased availability of digital scholarly data and computing resources accelerate visualization development. In this paper the results of multifaceted visual analysis by means of metadata mapping are presented. The author exemplifies the potential of visual analysis, especially in data exploration and retrieval. Application's interface allows to realize feedback with user and thus control the visualization process.*

Wstęp

Problem powszechnego nadmiaru informacji jest znany i dyskutowany w różnych środowiskach: naukowych, inżynierskich, biznesowych, dziennikarskich, literackich oraz licznych forach internetowych. Jak sobie zatem radzimy w praktyce? Przyzwyczailiśmy się, że w naszej ocenie źródła sieciowe wymagają porów-

nania, zestawienia, weryfikacji i ostatecznie odfiltrowania lub odrzucenia. Można określić taką kolejność zachowań, jako naturalną – użytkownicy sami przystosowują się do aktualnego stanu rozwoju Internetu. Według przewidywań naukowców, m.in. twórcy usługi World Wide Web Tima Bernersa Lee, taki stan nieuporządkowania miał trwać do 2010 roku[1], po którym sieć miała ewoluować do sieci typowo semantycznej [1, 2].

Na płaszczyźnie teoretycznej powstają różne koncepcje i szkoły, stawiające sobie za cel szybkie wydobycie relewantnych danych ze strumienia informacji. Istnieją metody wykorzystujące idee sieci semantycznych, strukturalizowanych zasobów sieciowych oraz serwisy kontrolowanej jakości (*subject gateways*). Są one wyraźnie uzależnione od stanu rozwoju technologii sieciowych. Inne podejście do tego problemu polega na edukacji szerokiego grona użytkowników sieci polegające na właściwym zastosowaniu technologii i kontroli edytorskiej w systemach wyszukiwawczych. Specjaliści informacji, zarówno teoretycy jak i praktycy, posiadający doświadczenie w obsłudze zasobów bibliologicznych i bibliotecznych, proponują rozwiązania oparte na odpowiednich dla danej problematyki modelach zarządzania informacją i wiedzą. Zakładając, że uda się w ten sposób uporządkować Internet, czytelność posegregowanej i pogrupowanej informacji prawdopodobnie pozostanie równoległym problemem do rozwiązania.

Ogrom informacji obserwujemy nie tylko w przestrzeni sieciowej. Zalewają (czasownik ten jest najlepszą metaforą przeładowania informacją) nas strumienie danych, produkowane codziennie przez systemy medialne. Przemysł wydawniczy, który zawdzięcza swój szybki rozwój technologiom DTP, technologiom sieciowym i dostępności elektronicznych czytników od dekady generuje ponad milion książek rocznie. Dla porównania: w całym XX wieku szacunkowa ogólnoświatowa ilość wydanych książek wyniosła 8 mln. Wzrost produkcji piśmienniczej jest szczególnie zauważalny we współczesnej nauce. Jeśli w 1950 roku wydawano na świecie 100 czasopism naukowych, to dzisiaj ta liczba przekroczyła jeden milion [4, s 10]. Gwałtownie rozrastają się również globalne specjalistyczne bazy danych, indeksujące coraz większe zasoby piśmiennictwa naukowego.

Takie zestawienia wielkoskalowych danych statystycznych, szczególnie w odniesieniu do dynamicznych zmian, wymagają prezentacji w postaci wykresów i map. Tabelaryczna forma, z punktu widzenia odbiorcy, w przypadku porównania obszernego zbioru dużych wartości liczbowych jest nieergonomiczna. Zagadnienie wydajności sposobów prezentacji informacji należy do obszaru badań nad ludzką percepcją i zrozumieniem (*perception and cognition*)[2] wzorców wizualnych – temat ten jest rozwijany poniżej.

Wizualizacja i wizualna analiza danych

W użytkowaniu dużych baz danych metody wizualne mogą w widoczny sposób wpływać na ich zrozumienie i interpretację. Generalnie struktury tabelaryczne (np. bazy danych) służą do tego, aby dane przedstawić wieloaspektowo: właściwości danych – rekordów w tabeli – opisywane są pomocą wielu pól. W przeniesieniu wartości tych pól na płaszczyznę reprezentacji ekranu lub wydruku przeszkadza właśnie ta nadmiarowa ilość właściwości. Jednym z rozwiązań może być matematyczne „rozciągnięcie" przestrzeni reprezentacji i uzyskanie w wyniku przestrzeni hiperbolicznej. Przy włączeniu mechanizmów powiększania (*zooming*), możliwe jest stosowanie klasycznej techniki wizualnej analizy – *focus plus context*[3].

W nauce wizualizacja informacji (w literaturze często używa się też terminu Infoviz) utorowała ścieżkę od końca lat 80-tych. Pojęcie to zdefiniowano w pracy trzech autorów *Reading in Information Visualization: Using Vision to Think* [10, s. 7-25]. Wizualizację informacji wyodrębnili oni na tle innych zastosowań wizualizacji jako metodologię naukową i praktyczną dotyczącą analizy danych abstrakcyjnych. W odróżnienie od wizualizacji naukowej, która zajmuje się zjawiskami naturalnymi i procesami fizycznymi, zarówno na zewnątrz, tak i wewnątrz człowieka, obiekty badań Infoviz mają naturę czysto abstrakcyjną: na przykład dane statystyczne, charakterystyki ruchu sieciowego, notowania giełdowe itp. Zgodnie z tą definicją dotyczą one bardziej procesów i zachowań jakościowych, niż ilościowych. W Infoviz konieczne jest zastosowanie wizualnej reprezentacji zamiast reprezentacji liczbowo-tekstowej. Kolejną niezbędną

[1] W 2008 roku Tim Berners Lee w swoim wystąpieniu na konferencji TED 2009 ogłosił, że „Web semantyczny już nadchodzi".

[2] W angielskojęzycznej literaturze fachowej te dwa pojęcia występują nierozłącznie przy omawianiu procesów kognitywnych i przetwarzania wizualnego (*Vision*) [30].

[3] Focus plus context – zasada projektowania interfejsu wizualizacyjnego, który umożliwia równolegle: widok całości rozkładu danych oraz ich szczegółów np. w powiększonym oknie [14].

cechą wizualizacji informacji jest właściwe wykorzystanie ludzkiej percepcji. Poza tym aplikacje do zastosowań wizualnych muszą zapewniać dwustronną interakcję z użytkownikiem, aby mógł on dobierać najbardziej pasujące do danego zadania charakterystyki układów graficznych: zagęszczenie, powiększenie, kolorystykę, ostrość glifów, oznakowanie i grupowanie badanych obiektów.

W wykształceniu pojęcia wizualizacji pomoże rozbudowana definicja wskazująca, iż nie jest to jedynie odtwarzanie danych, ale również wspomaganie ich zrozumienia i interpretacji [13, s. 1-9]. W wizualizacji informacji wielu badaczy widziało narzędzie nie tylko do analizy ale także do uruchomienia potencjału wnioskująco-poznawczego, wymuszającego zdobycie wiedzy o wzajemnych relacjach i podobieństwach grup danych [tamże, s. 9]. Jednocześnie interakcja wizualizacji zapewnia metaforyczną komunikację idei.

Według Edwarda Tufie – autora klasycznego podręcznika: *The Visual Display of Quantitative Information*, gdzie zostały określone zasady projektowania dobrego interfejsu wizualizacyjnego, wyłącznie w obrazach, a nie w liczbach znajdziemy najefektywniejszy sposób opisu, analizy i zestawień dużych zbiorów danych ilościowych [29, s. 12-35]. Należy wizualizację potraktować jako alternatywę dla rozbudowanych tabel, które komunikują odseparowane ciągi liczbowe. W analizie pojedynczych wartości tabela jest pomocna, lecz we wnioskowaniu, kiedy potrzeba wykrycia trendów, relacji, dynamiki i wzorców, już nie wystarcza.

Techniki wizualizacji wykorzystywano w nauce już dużo wcześniej. Zastosowano je w ramach eksploracyjnej analizy danych, którą po raz pierwszy określił i rozwinął amerykański statystyk John Tukey w 1977 roku *Exploratory Data Analysis*. Eksploracja danych (*dataminig*[4]), możliwa dzięki rozwojowi systemów komputerowych, jako jeden z etapów praktycznego odkrywania wiedzy o danych, służy do wynajdywania ukrytych zależności, podobieństw i trendów w grupach danych przy wykorzystaniu dużych repozytoriów i hurtowni danych (*warehouse*). Metody obliczeniowe eksploracyjnej analizy danych obejmują zarówno proste statystyki opisowe, jak i bardziej zaawansowane, wywodzące się z obszaru badań nad sztuczną inteligencją.

W sposób naturalny wizualizacja za pomocą wykresów i map stała się jednym z podstawowych rozwiązań nowoczesnego datamining, charakteryzującego się interfejsem przyjaznym dla specjalistów, analityków, maklerów biznesowych oraz użytkowników „inteligentnego oprogramowania" na rozmaitych poziomach. J. Tukey pierwszy dostrzegł i podkreślił znaczenie wizualizacji w eksploracyjnej analizie danych masowych [13, s. 15]. Cytując znane chińskie przysłowie „obraz wart jest tysiąca słów", dodał, że graficzne reprezentacje są wyjątkowo wydajne w szybkim przekazywaniu dużych ilości różnorodnej informacji numerycznej [tamże, s. 16; 30, s. 353]. Informacje te pomimo kompleksowej natury, z założenia są komunikowane w sposób czytelny i efektywny. Niezbędna jest tu wiedza o zdolnościach ludzkiego mózgu w odczytywaniu ukrytych wzorców.

Analiza wieloaspektowych baz danych w pierwszym przybliżeniu powinna zatem wykorzystywać narzędzia wizualizacyjne. W przypadku użytkowania rozbudowanych baz danych duży rozmiar, złożoność i konieczność ciągłej aktualizacji wymagają zastosowania wydajnych metod analitycznych. Wtedy użycie wizualizacji, jest metodą z wyboru.

Percepcja wzrokowa a zrozumienie

Dlatego więc naukowcy, wykorzystujący wizualizację winni również zainteresować się ludzką percepcją. W postrzeganiu obrazów wydawałoby się, że udział bierze przede wszystkim jeden ze zmysłów, angażując do tego narząd wzroku, a przetwarzanie informacji zachodzi dopiero po dotarciu sygnału do kory wzrokowej. Jeśli natomiast przyjrzymy się, jak jest zbudowana siatkówka ludzkiego oka, to odkryjemy, że obraz jest analizowany już na etapie postrzegania – percepcji. Siatkówka to składająca się z kilku warstw komórek nerwowych tkanka, pokrywająca wklęsłą wewnętrzną powierzchnię oka. Jej funkcja polega na przekształceniu wpadającego do oka światła w impulsy elektryczne przekazujące informacje do kory wzrokowej w mózgu. W siatkówce oka ułożone są trzy osobne warstwy receptorów; złożoność tę uzupełnia pięć rodzajów komórek: pręciki, czopki, komórki zwojowe, amakrynowe i horyzontalne [30, s. 53]. Każda warstwa i każdy typ komórek są odpowiedzialne za składowe ludzkiego widzenia, takie jak kontrast, krawędzie, jasność, korelacje barwne itp. oraz za przesyłanie informacji na zewnątrz oka (czyli do mózgu). Taka „specjalizacja zadań" umożliwia niezależny, szybki przekaz impulsów w głąb mózgu i równoległe przetwarzanie informacji o obrazach. Odpowiedni obszar

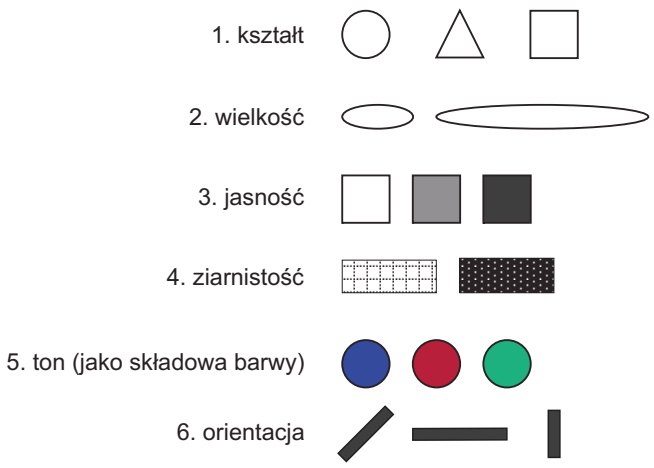
[4] *datamining* – w dosłownym tłumaczeniu oznacza „kopanie danych".

w mózgu, przeznaczony do takiej współpracy, usytuowany jest w korze wzrokowej. Podsumowując, oko plus kora wzrokowa tworzą potężny procesor równoległy o wysokim stopniu przepustowości i bezpośrednio sprzężony z naszymi ośrodkami poznawczymi. Te cechy świadczą o tym, że w poznawaniu otaczającego świata widzenie i rozumowanie ściśle współpracują, dlatego te dwa procesy są punktem odniesienia w kognitywnych badaniach nad wizualizacją.

Należy tu jeszcze wspomnieć, że lateralizacja mózgu sprawia, iż dwa niezależne kanały informacyjne, biegnące od lewego i prawego oka, również mają „własne" ośrodki przetwarzania, zlokalizowane na przeciwległych półkolach. Ale dzięki tej „komplikacja" możliwe jest widzenie stereoskopowe.

Możemy postrzegać obrazy przedstawione jedynie w określony sposób, i zupełnie ich nie dostrzegać w innej wizualizacji. Ta właściwość znajduje zastosowanie w grach percepcyjno-kognitywnych, wykorzystujących złudzenia optyczne. Jeśli zrozumiemy jak działa percepcja, to wiedzę tę można zastosować do wyświetlenia informacji. To co widzimy jako obiekty, to jest efekt przetworzenia i łączenia wizualnych cech, z których się buduje podstawowe elementy widzenia. O tych elementach po raz pierwszy napisał Jacques Bertin – francuski psycholog – w książce *Semiology of Graphics*, gdzie spróbował usystematyzować znaczenia znaków graficznych. Wyróżnił on sześć podstawowych cech glifów[5], które decydują o widzeniu [30, s. 145-159; 23, R. 1], przedstawione na Rysunku 1.

Najbardziej rozpoznawalnymi kształtami glifów na mapach informacji są koła, kwadraty, romby. Różnicowanie ich wartości uzyskuje się za pomocą kolorów, tonów (np. odcieni szarości w skali biały – czarny)



Rys. 1. Podstawowe cechy glifów wykorzystywanych w wizualizacji informacji.

i wielkości (grubość, wysokość). W odwzorowaniu dynamicznych zmian przydatne jest zaznaczanie orientacji glifów.

Działanie percepcji polega na nadawaniu arbitralnych wartości obserwowanym w układzie wizualizacyjnym obiektom. W szczególności: wyższym słupkom, dłuższym kreskom i liniom, powiększonym lub ciemniejszym kółkom intuicyjnie przypisujemy większe wartości. Wyróżniające się kolorem lub kształtem glify świadczą o odmienności danego obiektu względem całości. W ten sposób na poziomie percepcji możemy już grupować i kategoryzować dane względem podobieństwa cech, ułatwiając sobie dalszą wielostronną eksplorację i zgłębianie wiedzy o badanych obiektach.

Reguły percepcji wzrokowej mogą również pomóc w doborze kształtu glifów. Dlatego w zestawieniu zróżnicowanych wartości wydajny jest wykres słupkowy, wymyślony jeszcze w XVIII wieku[6]. Naturalnym ruchem gałek ocznych jest przemiatanie wzrokiem w kierunku góra-dół. Dlatego wykres słupkowy jest bardziej ergonomiczny niż inny, używany w statystyce, np. „tortowy". Wykres kołowy stwarza tę trudność, iż mylnie szacujemy wartości ostrych i mocno rozwartych kątów oraz ich ocena zależy od pionowego lub poziomego ułożenia segmentu oraz rzutu aksjometrycznego (wstaw przypis) obiektu.

Wizualizacja domen wiedzy

Wizualizacja informacji *stricte* naukowej, czyli pochodzącej z bibliograficznych i bibliometrycznych baz danych znacząco rozwinęła się w ostatnim dziesięcioleciu, pomimo tego, że pierwsza mapa nauki, nakreślona ręcznie powstała w latach 60-tych, a wygenerowana komputerowo – w latach 70-tych. Wizualizacja zaczyna więc budować solidną pozycję w metodologii nauk. Uznaje się, że służy ona do wykrywania aktualnych trendów tematycznych, dominujących obszarów w nauce oraz dynamiki zmian w historii rozwoju badań. Tematykę tę dyskutowali i dyskutują na łamach prasy biblio- i naukometrycznej (w kolejności chronologicznej): Eugene Garfield [15-17], Henry Small i Henry White [29], Chaomei Chen [11, 12], Kevin Boyack [9], Katy Börner [4-8]. W polskiej literaturze

[5] Kształt znaku graficznego. Termin znany w poligrafii.

[6] Pomysłodawcą wykresu słupkowego był szkocki inżynier William Playfair. W 1786 roku przedstawił on w ten sposób dane ekonomiczne w *Commercial and Political Atlas*. Był również autorem wykresu kołowego.

fachowej też są dostępne prace, dotyczące tej problematyki [21-23].

W wizualizacji nauki najczęściej wykorzystywane są bazy WoS, Medline, Scopus. Badacze na takich mapach mają możliwość całościowego spojrzenia na rozwój interesującej tematyki, sfery badań, grupy badawczej albo nauki w skali lokalnej, krajowej i globalnej. Mapy nauki są publiczne dostępne za pomocą serwisów dedykowanych. Jednym z nich jest wystawa posterowa on-line *Places&Spaces*[7], utworzona przez naukowców z Uniwersytetu w Indiana.

Przykłady map on-line, a także przytoczone poniżej dowodzą, iż wizualizacja domen wiedzy zawiera duży potencjał analityczny, m.in. umożliwiając:

— ujawnienie społecznej struktury dyscypliny/nauki na podstawie danych o współautorstwie, współcytowaniach;
— badanie rozwoju dziedzin nauki i ewentualne prognozowanie przyszłych trendów naukowo-badawczych;
— wspomaganie wyszukiwania informacji, służąc jako graficzny interfejs wyszukiwawczy;
— określenie kierunków i polityki finansowania określonych obszarów nauki.

Formalizmy analizy wizualnej w przykładach

Jednostki analizy – są to zazwyczaj metadane dokumentów, składające się na takie pola, jak: tytuł, autor, abstrakt, słowa kluczowe, źródło pochodzenia, dziedzina badań, liczba cytowań oraz pozycje bibliograficzne. W zależności od postawionego celu badawczego, mapy wizualizacyjne służą do reprezentacji graficznych podobieństw i wzajemnych relacji pomiędzy artykułami, czasopismami, autorami i/lub osobami, powołującymi się na dane prace, jak również instytucjami specjalizującymi się w danej dziedzinie. Namnożenie się metod, technik, zróżnicowanie zadań wizualizacji od dawna stwarza konieczność usystematyzowania i zastosowania formalizmu w opisie tej nowej dyscypliny.

Jak zasugerowały Katy Börner i Angela Zoss [7], w wytypowaniu grup materiału badawczego można wyróżnić trzy poziomy analizy: micro, mezo i macro. Zastosowanie najniższego poziomu (micro) – indywidualnego, oznacza mapowanie metadanych, charakteryzujących aktywność, działanie, mobilność, rozwój konkretnej osoby. W ten sposób możemy zobaczyć z kim dany naukowiec współpracuje, jak i czy zmienia się podejmowana przez niego problematyka badań, w jak przebiega jego kariera naukowa itp. Drugi poziom – mezo – odnosi się do grup społecznych. Pytania stawiane w tego rodzaju analizie dotyczą rozpiętości współpracy danego zespołu, granic i zmienności zainteresowań, obszarów naukowo-badawczych instytucji, jednostek edukacyjnych. Na poziomie macro otrzymuje się mapy dziedzin wiedzy dla danego kraju, kontynentu lub w skali globalnej. Na najwyższym poziomie agregacji danych[8] otrzymuje się wizualne konfiguracje wybranych dziedzin naukowych lub struktury całej nauki.

W nowoczesnym, coraz częstszym podejściu oprócz klasycznych metod i jednostek wykorzystuje się statystyki zachowań użytkowników bibliotek wolnego dostępu i repozytoriów literatury naukowej [3]. Takie informacje jak logi odwiedzających, schematy zachowań, wpisywane hasła są bardzo wartościowym materiałem do badań nad popularnością, czytelnością i cytowalnością dokumentów. Zaprojektowane w serwisach sieciowych mechanizmy społecznego tagowania mogą wnieść istotne modyfikacje do wyjściowej reprezentacji.

Etapy procesu wizualizacji włączają: określenie jednostek analizy i skompletowanie danych, wybranie odpowiedniej miary i przeprowadzenie obliczeń i następnie uruchomienie algorytmów mapowania przestrzennego [5]. W definiowaniu miary podobieństwa zawsze musi być procedura liczenia wspólnych cech obiektów. Najczęściej używa się iloczynu skalarnego (zwykłego przemnożenia wartości). Alternatywnie są wybierane na przykład modele wektorowe słów w tekście [21], korelacje Pearson'a lub zaawansowane algorytmy lingwistyczne.

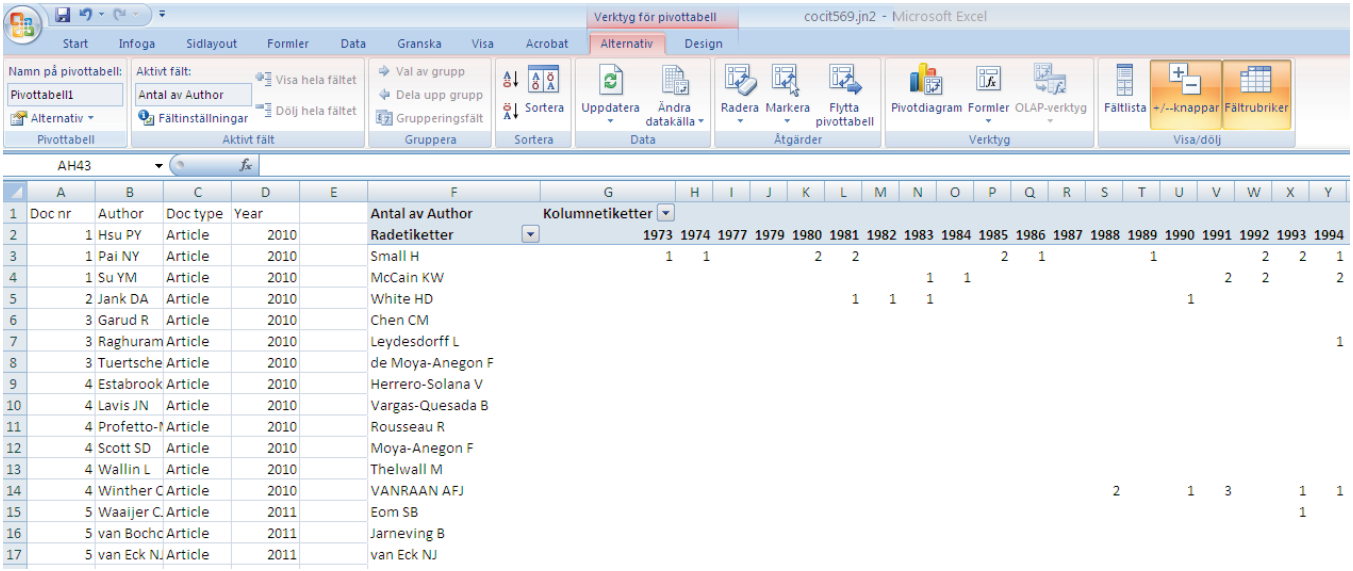
Kolejne podejście w systematyce analiz wizualnych wymaga określenie perspektywy mapowania. Zaproponowano więc istniejące metody Infoviz sprowadzić do czterech fundamentalnych perspektyw mapowania: czasowej, geograficznej, tematycznej i sieciowej [7].

1. W skali czasu (kiedy)

Dane bibliograficzne są mapowane w określonym okresie bądź okresach czasu. Powstały wzór obrazuje dynamikę zmian w strukturze i organizacji danego obszaru wiedzy zazwyczaj za pomocą osi czasu. Eugen Garfield nazwał takie reprezentacje naukografami [17]. Właściwym pytaniem badawczym tu jest: „Kie-

[7] www.scimaps.org

[8] Termin, używany też w statystyce w celu określenia określonej procedury obliczeniowej. W kontekście natomiast chodzi o najwyższy poziom organizacji metadanych w odniesieniu do kategorii tematycznych badań.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Doc nr	Author	Doc type	Year		Antal av Author	Kolumnetiketter																		
2						Radetiketter		1973	1974	1977	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
3	1	Hsu PY	Article	2010		Small H		1	1			2	2				2	1			1			2	2	1
4	1	Pai NY	Article	2010		McCain KW									1	1							2	2		2
5	2	Su YM	Article	2010		White HD							1	1	1							1				
6	3	Jank DA	Article	2010		Chen CM																				
7	3	Garud R	Article	2010		Leydesdorff L																				1
8	3	Raghuram	Article	2010		de Moya-Anegon F																				
9	3	Tuertsche	Article	2010		Herrero-Solana V																				
10	4	Estabrook	Article	2010		Vargas-Quesada B																				
11	4	Lavis JN	Article	2010		Rousseau R																				
12	4	Profetto-I	Article	2010		Moya-Anegon F																				
13	4	Scott SD	Article	2010		Thelwall M																				
14	4	Wallin L	Article	2010		VANRAAN AFJ														2		1	3		1	1
15	4	Winther C	Article	2010		Eom SB																			1	
16	5	Waaijer C.	Article	2011		Jarneving B																				
17	5	van Bochc	Article	2011		van Eck NJ																				
	5	van Eck N.	Article	2011																						

Rys. 2. Wizualizacja historiograficzna artykułów na temat analizy współcytowań.

dy?". Tak możemy prześledzić genealogię współczesnej nauki na podstawie metadanych 39 mln. artykułów naukowych opublikowanych w latach 1817-2010 [19]. Atrakcyjności tej mapy dodaje źródło pochodzenia danych – baza Scopus, która ma około dwukrotnie większą objętość tytułów w porównaniu z WoS. Wyraźnie widać na mapie, że w drugiej połowie XX wieku gwałtowny rozwój należy nauk medycznych i przyrodniczych. Na początku wymienionego okresu czasu dominuje fizyka i astronomia, w środku skali (koniec XIX wieku po lata 20-te XX w.) – matematyka. Łatwe wytłumaczenie znajdziemy w historii rozwoju nauk matematycznych, wynikającego również z uwarunkowań geopolitycznych. XIX wiek – to czas formowania pojęć algebry klasycznej. Do matematycznych kierunków na początku XX w. zalicza się także słynna lwowska szkoła Stefana Banacha [28]. W czasie 1. wojny światowej skonstruowano mechaniczne urządzenia szyfrująco-deszyfrujące – są to początki szybkiego rozwoju kryptografii i kryptoanalizy.

W naukografach oś czasu nie zawsze jest wymogiem. Dynamikę zmian można przedstawić w postaci serii zmieniających się obrazów. Tak sześć map pokazuje zmiany w organizacji specjalistycznej klasyfikacji literatury informatycznej z cyfrowej biblioteki ACM od 1968 do 2009 [20]. Ogólnodostępny program Excel również nadaje się do obrazowania zmian. Mapa demonstruje historię analizy współcytowań na przestrzeni lat 1973 do dzisiaj (Rys.2). Jest także graficznym dowodem, iż Henry Small, Henry White, Katherine McCain byli pionierami w tej dziedzinie.

2. Geograficzna (gdzie)

Jeśli dane biblio- i naukometryczne przetworzyć w ten sposób, aby odfiltrować informację o lokalizacji instytucji, z którą powiązana jest badana grupa osób, to przy wykorzystaniu geograficznych map danego regionu można otrzymać geograficzne reprezentacje z góry zdefiniowanej aktywności ludzi. Tego typu wizualizacje dostępne są na stronie *Places&Spaces* w kategorii „*Cartographic*". Można przeanalizować jak obraz świata utworzony na podstawie liczby logów użytkowników gry sieciowej[9] różni się od rzeczywistego, gdzie są w USA wolne miejsca pracy dla naukowców i jakiej kategorii[10] albo skupić się na historycznych wizualizacjach, np. marsz armii Napoleona na Moskwę, klęskę i odwrót[11].

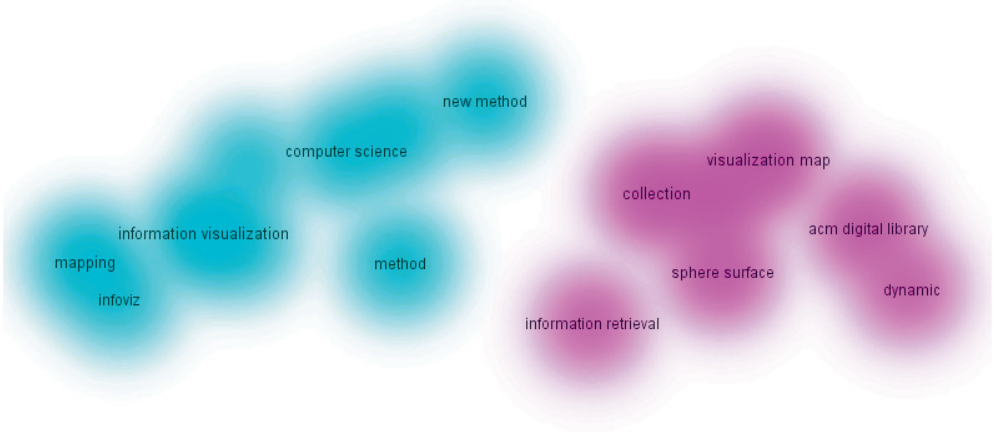
3. Tematyczna (co)

Perspektywa „co" ukierunkowana jest na przedstawienie zestawień tematycznych analizowanych danych literatury. W graficznych reprezentacjach powstającym klastrom (grupom) artykułów lub ich twórców przypisuje się nazwy opisowe, które generuje się statystycznie albo nadaje się ręcznie. Te nazwy identyfikują tematyczne obszary badań naukowych i w ten sposób dokonuje się mapowania nauki. Najbardziej rozpoznawalną, obiektywną (bo wykorzystującej aż 7 mln artykułów z baz WoS, Scopus) i aktualną jest

[9] http://www.scimaps.org/maps/map/logicaland_participa_74/

[10] http://www.scimaps.org/maps/map/us_job_market_where__122/

[11] http://www.scimaps.org/maps/map/napoleons_march_to_m_9

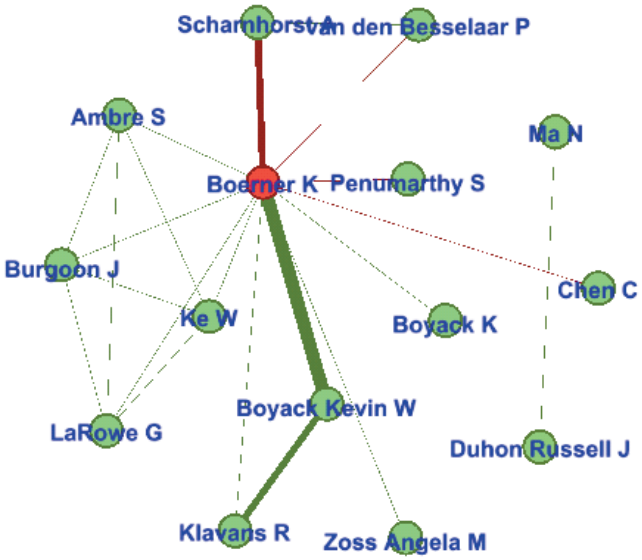


Rys. 3. Tematyczna wizualizacja obszarów zainteresowań autorki.

mapa z 2007[12] autorstwa K.W. Boyacka [9], pokazująca aktualne i przyszłe trendy w nauce światowej. Rys. 3 zawiera mapę obszarów zainteresowań autorki utworzonej na podstawie tytułów, słów kluczowych i abstraktów własnych prac angielskojęzycznych. Użyto tu mapy tzw. energetycznej, która ilustruje powstanie dwa główne klastry tematyczne, odnoszące się metod wizualizacji/mapowania oraz badań nad literaturą informatyczną.

4. Sieciowa (z kim)

Za pomocą sieci możemy wizualizować więzi społeczne, co jest wykorzystywane w formalizmie „z kim". To pole badań nazywa się „Analizą sieci społecznych" (*SNA – Social Network Analysis*), znajdujące zastosowania praktyczne w marketingu, biznesie i nauce. Powstające graficzne reprezentacje współautorów, współpracowników, badaczy odwołujących się do tych samych źródeł generowane w ten sposób aby najlepiej odwzorować intensywność, stopień relacji pomiędzy badanymi osobami. Do tego celu używa się grafów, składające się z węzłów (wierzchołków) i krawędzi (połączeń). Węzły w takiej sieci identyfikują osoby, a krawędzie o zróżnicowanych długości, szerokości – relacje pomiędzy nimi. Rysunek 4 przedstawia mapę współautorów znanej badaczki i popularyzatorki Infoviz – Katy Borner. Trzon tej współpracy należy do grona: Boyack, Klavans i Scarhnhorst, co odpowiada rzeczywistości. Mapa została utworzona z wykorzystaniem wolnego oprogramowania Gephi[13].



Rys. 4. Mapa współautorstwa Katy Borner.

Podsumowanie możliwości analizy wizualnej – visual mining

Wizualizacja, którą odkryto jako narzędzie wspomagające eksploracyjną analizę danych masowych, staje się na naszych oczach samodzielną dyscypliną o praktycznych zastosowaniach w nauce, biznesie i neuromarketingu. Metodologia wizualizacji informacji znajduje się w fazie rozwoju. Brakuje tu jednak solidnych podstaw teoretycznych, usystematyzowania i opisów formalistycznych [12, s. 42-55].

W artykule zostało przedstawione wieloperspektywiczne podejście do problematyki Infoviz. Załączone i wskazane przykłady dowodzą, iż wizualizacja umo-

[12] http://www.scimaps.org/maps/map/maps_of_science_fore_50/

[13] www.gephi.org

żliwia wieloaspektową analizę kolekcji dokumentów na podstawie ich metadanych. Organizację tych jednostek analizy proponuje się sprowadzić do trzech poziomów: indywidualnego (mapy aktywności konkretnych osób, rozwoju indywidualnej kariery naukowej), grupowego (wizualizacja współpracy społeczności lokalnej lub sieciowej) i krajowego (mapy rozwoju badań w danym kraju, globalne mapy nauki). Inny formalizm polega na wyodrębnieniu czterech fundamentalnych perspektyw mapowania: czasowej (kiedy), geograficznej (gdzie), tematycznej (co) i sieciowej (z kim). Takie podejście z pewnością dostarcza nowych możliwości w analizie, interpretacji i wnioskowaniu o kompleksowej strukturze danych. Niepodważalną cechą jest to, iż mapy wizualizacyjne stymulują poznawanie współczesnego stanu wiedzy. Niosą również wartość edukacyjną, ponieważ tego typu aplikacje zawierają mechanizmy interakcji. W projektowaniu interfejsów wizualizacyjnych są wykorzystywane wówczas wyniki badań nad ludzką percepcją i zrozumieniem.

Autorka pokazuje, że nowoczesne mapy generowane za pomocą algorytmów wizualizacyjnych są swego rodzaju arkuszami graficznymi, umożliwiającymi wielostronną (czyli wielowymiarową oraz obiektywną) analizę danych. Wykorzystywane od dawna w zaawansowanych metodach analizy eksploracyjnej, obecnie są niezastąpionym elementem i etapem procesu datamining. Istniejące określenie visualmining dobrze odzwierciedla swoje przeznaczenie naukowo-empiryczne oraz formę interakcji z użytkownikiem, włączając tak ważne w zastosowaniach sieciowych sprzężenie zwrotne.

Literatura cytowana

1. Berners-Lee T.: *The Semantic Web.* "Scientific American" 2001. [on-line]. [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.
2. Berners-Lee T.: *The next Web of open, linked data.* 2008. [on-line]. Zasoby youtube. [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://www.youtube.com/watch?v=OM6XIICm_qo
3. Bollen Johan i in.: Clickstream Data Yields High-Resolution Maps of Science. "PLoS ONE" [on-line] 2009, Vol. 4, no. 3 [dostęp 20 stycznia 2013] Dostępny w World Wide Web: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0004803.
4. Börner K.: *Atlas of Science*, MIT Press, 2010.
5. Börner K., Chen Ch., Boyack K.W.: *Visualizing Knowledge Domains.* W: B. Cronin (red.). *Annual Review of Information Science & Technology.* "Information Today" 2005 Vol. 37 s. 179-255.
6. Börner K., Scharnhorst A.: *Visual Conceptualizations and Models of Science.* "Journal of Informetrics" 2009 No. 3(3) s. 161-172.
7. Börner K., Zoss A.: *Evolving and Emerging Populations and Topics Extracted from NSF Awards.* "Virtual Presentation to NSF" 2000 no. 7 [on-line]. [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://scienceofsciencepolicy.net/system/files/attachements/2010-borner-zoss-nsf.pdf
8. Bourner K., Klavans R. i in.: *Design and Update of a Classification System: The UCSD Map of Science.* [on-line]. Scimaps portal. [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://www.scimaps.org.
9. Boyack K. i in.: *Mapping the Backbone of Science.* "Scientometrics" 2005 Vol. 64 No 3s. 351-374.
10. Card S. K., Mackinlay J. D., Shneiderman B.: *Reading in Inforation Visualization: Using Vision to Think.* USA, CA 1999.
11. Chen Ch., Kuljis J.: *The rising landscape: a visual exploration of superstring revolutions in physics.* "Journal of the American Societyfor Information Science and Technology" 2003 Vol. 54 No. 5 s. 435–446.
12. Chen Ch.: *Information Visualization: Beyond the Horizon.* Wyd. 2, Springer, 2006.
13. Few S.: *Now you see it. Simple Visualization techniques and Quantitative Analysis.* CA. 2009.
14. *Focus-plus-Context.* [on-line]. Portal Infovis-wiki-net. [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: Dostępny w World Wide Web: http://www.infovis-wiki.net/index.php/Focus-plus-Context
15. Garfield E.: *Essays/Papers on „Mapping the World of Science"* [on-line]. E. Garfield, Ph. D. Home Page [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://garfield.library.upenn.edu/ mapping/mapping.html
16. Garfield E.: *From the science of science to scientometrics visualizing the history of science with HistCite.* [on-line]. "Proceedings of ISSI" 2007 Vol. 1 No. 21-26. [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://garfield.library.upenn.edu/papers/issiprocv1p21y2007.pdf
17. Garfield E.: *Scientography: Mapping the tracks of science.* W: "Current Contents: Social & Behavioural Sciences" 1994 nr 7(45) s. 5-10.
18. Marszakowa-Szajkiewicz I.: *Bibliometryczna analiza współczesnej nauki.* Katowice 1996, s.32-38.
19. Mosher D.: *Data as Art: 10 Striking Science Maps.* 2011. [on-line]. [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://www.wired.com/wiredscience/2011/03/best-science-maps/

20. Osińska V., Bala P.: *New Methods for Visualization and Improvement of Classification Schemes: The Case of Computer Science*. "Knowledge Organization" 2010 nr 37 s. 157-172.
21. Osińska V.: *Przybliżenie semantyczne w wizualizacji informacji w Internecie i bibliotekach cyfrowych*. „Biuletyn EBIB" [on-line] 2006, nr 7 (77) [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://www.ebib.info/2006/77/osinska.php.
22. Osińska V.: *Rozwój metod mapowania domen naukowych i potencjał analityczny w nim zawarty*. W: *Zagadnienia Informacji Naukowej*. Warszawa 2010, s. 15-16.
23. Osińska V.: *Wizualizacja informacji*. Warszawa 2010.
24. *Places&Spaces. Mapping Science*. Wystawa on-line. [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://www.scimaps.org.
25. Scharnhorst A.: *Complex Networks and the Web: Insights From Nonlinear Physics*. [on-line]. "Journal of Computer-Mediated Communication" 2003, Vol. 8 No.4, [dostęp 10 lipca 2012]. Dostępny w World Wide Web: http://jcmc.indiana.edu/vol8/issue4/scharnhorst.html
26. Skalska-Zlat M.: *Cybermetrics, Netometrics, Webometrics – nowe pojęcia i zadania informetrii*. W: *Przestrzeń informacji i komunikacji społecznej*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego, 2004, ss. 159-168.
27. Small H.: *Co-citation in the scientific literature: A new measure of the relationship between two documents*. "Journal of the American Society for Information Science" 1973 No. 24 s. 265–269.
28. Stefan Banach – matematyk stulecia. „Dziennik Związkowy. Polish Dayly News" 27 Kwietnia 2012. [on-line]. [dostęp 20 stycznia 2013]. Dostępny w World Wide Web: http://www.dziennikzwiazkowy.com/wspomnienia/20590-stefan-banach--matematyk-stulecia.html
29. Tufte E.: *Envisioning Information*. USA: Graphic Press, 1990.
30. Ware C.: *Information Visualization. Perception for Design*. USA, CA 2004.

Dr Veslava OSIŃSKA – Uniwersytet Mikołaja Kopernika w Toruniu. Instytut Informacji Naukowej i Bibliologii. Adres: 87-100 Toruń, Bojarskiego 1; e-mail: wieo@umk.pl.