Podtyp dziedziczy wszystkie pola typu nadrzędnego (włącznie z wartościami domyślnymi, listą dopuszczalnych wartości, objaśnieniami).

Powszechnie stosowanym sposobem tworzenia dokumentów półustrukturalizowanych stało się we współczesnych systemach informacyjnych wypełnianie masek wspomagających wprowadzanie danych (porównaj na przykład rysunek 228, rysunek 229 czy rysunek 230). W obu opisanych sytuacjach dopuszczano wprowadzenie niezaplanowanych treści przez projektowanie dodatkowych pól tekstowych, na przykład uwagi, komentarze.

Formalnie modeluje się dokumenty półustrukturalizowane z wykorzystaniem grafów oznakowanych [G. Grahne i inni, 2003]. Współcześnie dokumenty półustrukturalizowane utożsamia się często z dokumentami budowanymi z wykorzystaniem języków znaczników (porównaj na stronie 60), które są oznakowanymi drzewami, będącymi specjalnym przypadkiem grafów oznakowanych. Granica pomiędzy półustrukturalizowanymi dokumentami a danymi organizowanymi według dobrze określonego schematu danych może być trudna do uchwycenia [A. Laender i inni, 2002].

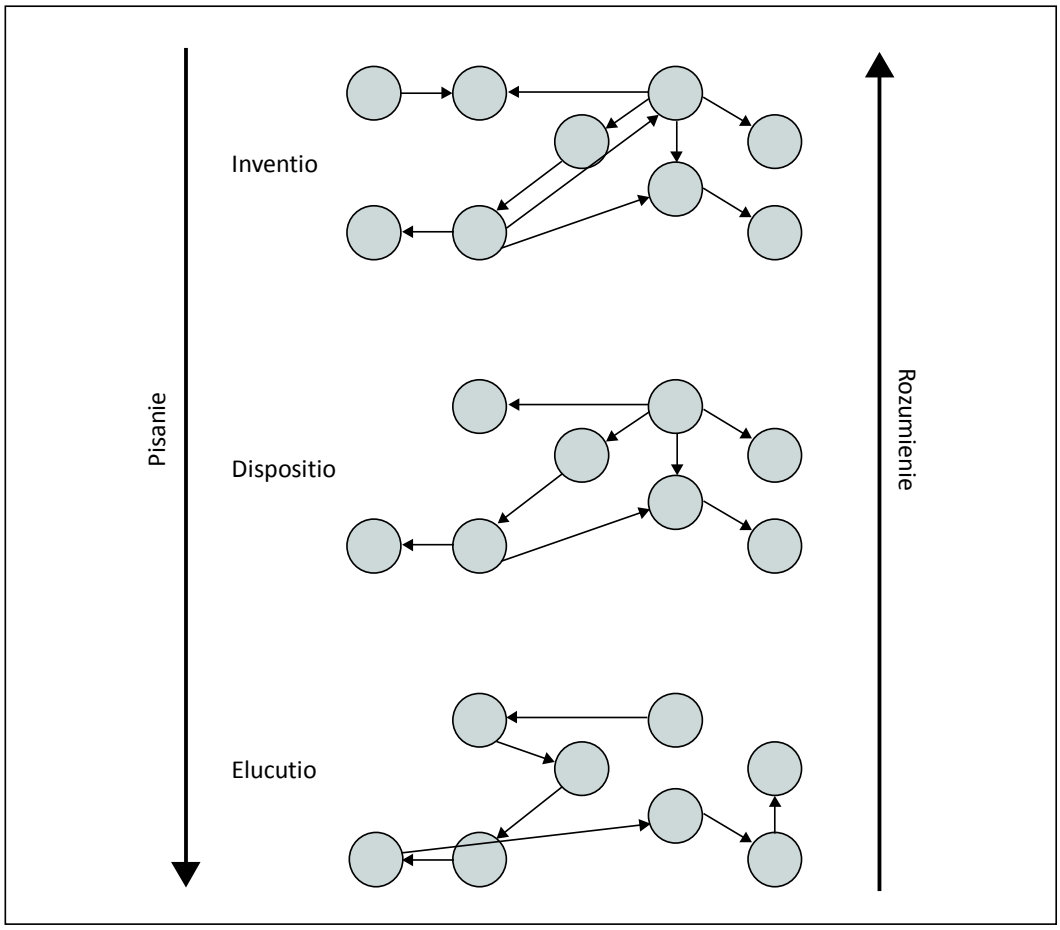
Metadane dokumentów

Dokumenty – w systemach informacyjnych, które je gromadzą – opatrzone są często metadanymi. W ogólności metadanymi dokumentu nazywać będziemy atrybuty, mówiące o jego treści, strukturze i źródle pochodzenia. Często wyliczane są listy pożądanych metadanych dokumentów. Metadane rozumie się na przykład jako informacje określające:

- twórców i współtwórców dokumentu,
- rozmiar dokumentu (liczbę znaków, ilustracji, rozmiar w bajtach itp.),
- datę i czas powstania dokumentu,
- status dokumentu (na przykład roboczy, wstępnie zatwierdzony, zatwierdzony),
- datę i czas zatwierdzenia dokumentu,
- osobę/instytucję upoważnioną do zatwierdzenia,
- tytuł dokumentu,
- powiązania i relacje z innymi dokumentami,
- tematykę dokumentu (porównaj na stronie 90),
- język dokumentu,
- cel powstania dokumentu (w szczególności adresatów dokumentu),
- dostępność treści dokumentu (ograniczenia dotyczące czasu lub osób upoważnionych do odczytania treści),

Odwrotność procesu pisania i rozumienia przedstawia rysunek 22. Koła przedstawiają węzły, strzałki powiązania. Na górnym rysunku widzimy węzły powiązane zgodnie z *inventio*, z nich tworzona jest hierarchia, która przekształcana jest w listę.

Ważny dla nas związek pomiędzy retoryką a hipertekstem w inny sposób ilustruje rysunek 23.



Rysunek 22. Pisanie i czytanie w hipertekście inspirowanym retorykom

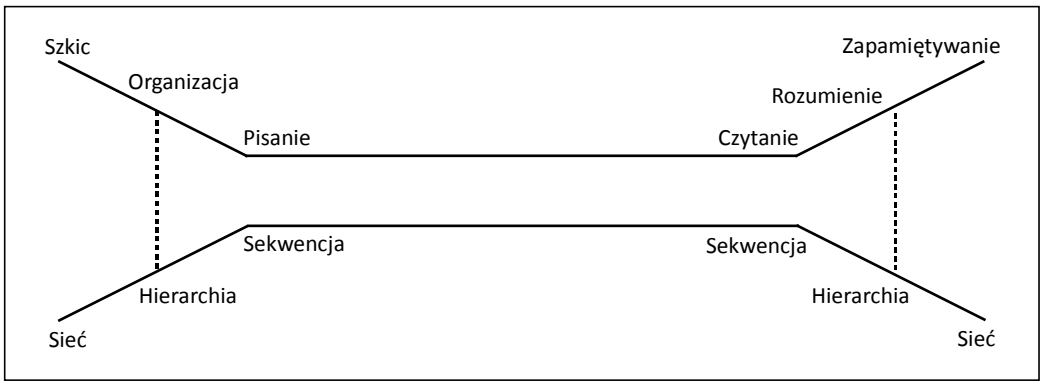
HDM - model projektowania hipertekstu

Model projektowania hipertekstu (*Hypertext Design Model*) jest związany z hipertekstowymi systemami autorskimi (*hypertext authoring*) [B. Schröcksnadl i inni, 1992], [K. Meusel i inni, 1993]. Model ten nawiązuje także do modelu relacyjnego [F. Garzotto i inni, 1988]. Interesujące jest rozgraniczenie pomiędzy rolą projektującego strukturę hipertekstu (*authoring-in-the-large*) a rolą autora treści poszczególnych węzłów (*authoring-in-the-small*). Projektujący strukturę tworzy schematy (*empty skeleton*), wykorzystywane przez autorów poszczególnych węzłów. Odpowiada to

proponowanym przez nas strukturom, wykorzystującym filtrowanie informacji z Internetu dla tworzenia instancji przepływów pracy [W. Abramowicz i inni, 2001a, W. Abramowicz i inni, 2001f].

Grafy dwudzielne

W grafach dwudzielnych wyróżnia się dwa rodzaje węzłów [G. Chartrand i inni, 2004]. Często, wykorzystując je w modelowaniu, przypisuje się rodzajom węzłów określoną semantykę. Nie inaczej jest w przypadku hipertekstu. W modelach hipertekstu postrzegamy dwudzielność jako podział pomiędzy częścią deklaratywną i strukturalizującą hipertekstu [W. Abramowicz, 1984a], [W. Abramowicz, 1990a]. Oba rodzaje węzłów mogą być typizowane oraz opatrywane atrybutami odpowiadającymi metadanym technicznym i biznesowym.



Rysunek 23. Pisanie i czytanie a retoryka [B. Shneiderman, 1987]

Sieci Petriego

Sieci Petriego są grafami dwudzielnymi, w których podziała może przebiegać pomiędzy deklaratywnością (*places*) i proceduralnością (*transitions*) [P.H. Starke, 1980]. Są one także wykorzystywane do modelowania hipertekstu [R. Furuta i inni, 1989a], [R. Furuta i inni, 1989b], [P.D. Stotts i inni, 1988], [P.D. Stotts i inni, 1989b], [P.D. Stotts i inni, 1989a]. W takim modelowaniu szczególnie podkreśla się operacje wykonywane na hipertekście.

Te dwa spojrzenia na modelowanie hipertekstu jako grafu dwudzielnego wskazują podstawowe problemy w zarządzaniu nim, które należy rozwiązać: jak modelować węzły i strukturę pomiędzy nimi, jak modelować operacje na nich, jak prezentować struktury i operacje użytkownikom.

Hipertekst jako automat skończony

[P. Schnupp, 1992] modeluje hipertekst jako automat skończony. W każdym stanie hipertekstu G_z mamy wyróżniony węzeł zwany węzłem aktualnym k_i. Możemy wówczas uporządkowaną piątkę Meisera przedstawić następująco:

$$G_Z = [K, E, F, I, A, k_i], \tag{7}$$

gdzie:

- K jest skończonym zbiorem węzłów,
- E jest zbiorem skierowanych, opatrzonych atrybutami ze zbioru A krawędzi,
- I jest zbiorem informacji,
- Funkcja F przyporządkowuje informacjom węzły (F: I → K).

Każda dopuszczalna operacja na hipertekście o_i przekształca go w hipertekst o nowym węźle aktualnym k_j:

$$[K, E, F, I, A, k_i] \xrightarrow[o_i]{} [K, E, F, I, A, k_j]. \tag{8}$$

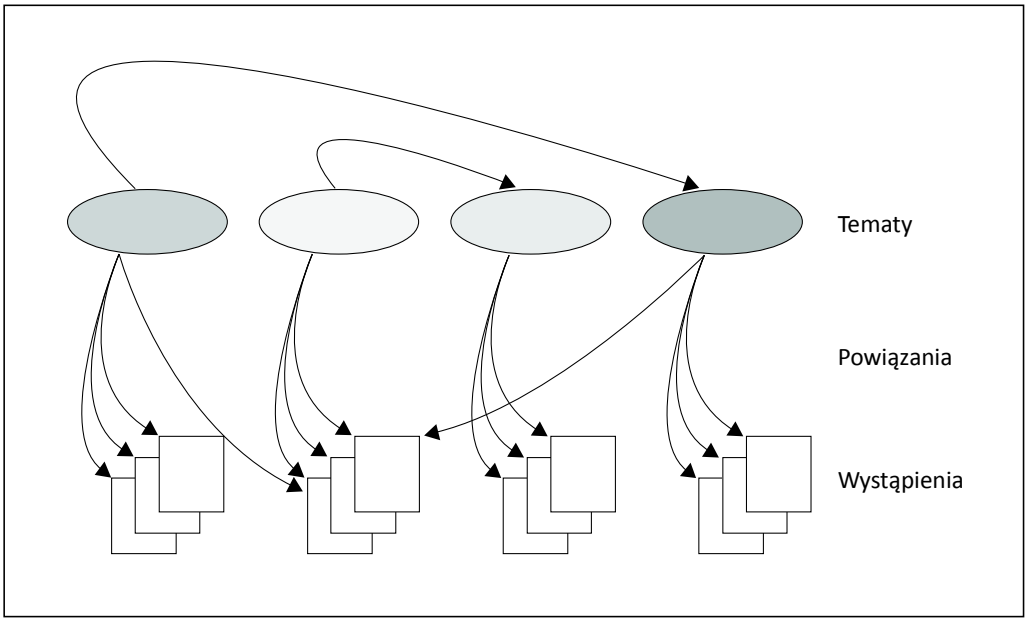
Modelowanie z wykorzystaniem hipergrafów

Modelowanie wykorzystujące teorię grafów ograniczone jest możliwością modelowania relacji dwuargumentowych. Tego ograniczenia nie ma modelowanie wykorzystujące teorię hipergrafów[106], gdzie modelowane są relacje n argumentowe [C. Berge, 1989]. W pracy [W. Abramowicz, 1990a] przedstawiliśmy uzasadnienie dla modelowania relacji n-argumentowych hipertekstu. Hipergrafy pozwalają łatwo modelować na przykład wszystkie obiekty mające te same wartości atrybutów. Hiperłącza wiążą wówczas takie obiekty. Ważną cechą hipergrafów jest możliwość przynależności dowolnego obiektu do dowolnej liczby hiperłączy. W konsekwencji wartości atrybutów mogą stanowić o przynależności obiektów hipertekstu do różnych klas obiektów. Własność ta może być wykorzystywana do modelowania nawigacji, wyszukiwania i filtrowania obiektów z hipertekstu [W. Abramowicz, 1990c]. Mechanizm ten pozwala także na modelowanie struktur temporalnych [K. Sandkuhl i inni, 1992]. Przedstawienie relacji n argumentowych może stanowić ciekawą drogę rozwoju Internetu. Jednak wiąże się to z koniecznością badań nad nowymi modelami adresowania (porównaj na stronie 332). Wiele tych modeli stanowiło podstawę do licznych implementacji systemów hipertekstu. W publikacji [J. Whitehead, 2000], można znaleźć referencje do kilkudziesięciu takich systemów stworzonych w latach 1968 – 2000.

[106] Wystąpienie przedrostka hiper- w nazwie hipertekstu i hipergrafów jest przypadkowe.

Mapy tematów i umiejętności

Mapy tematów (*topic maps*) są próbą strukturalizacji informacji w Internecie nawiązującą do hipertekstu. Podlegają one standaryzacji ISO[107]. Mapy tematów składają się z trzech elementów: tematów (*topics*) będących rzeczami, osobami czy pojęciami; abstrakcyjnych powiązań (*associations*); wystąpień (*occurences*). Stąd bywają nazywane TAO (*topics, associations, occurences*).



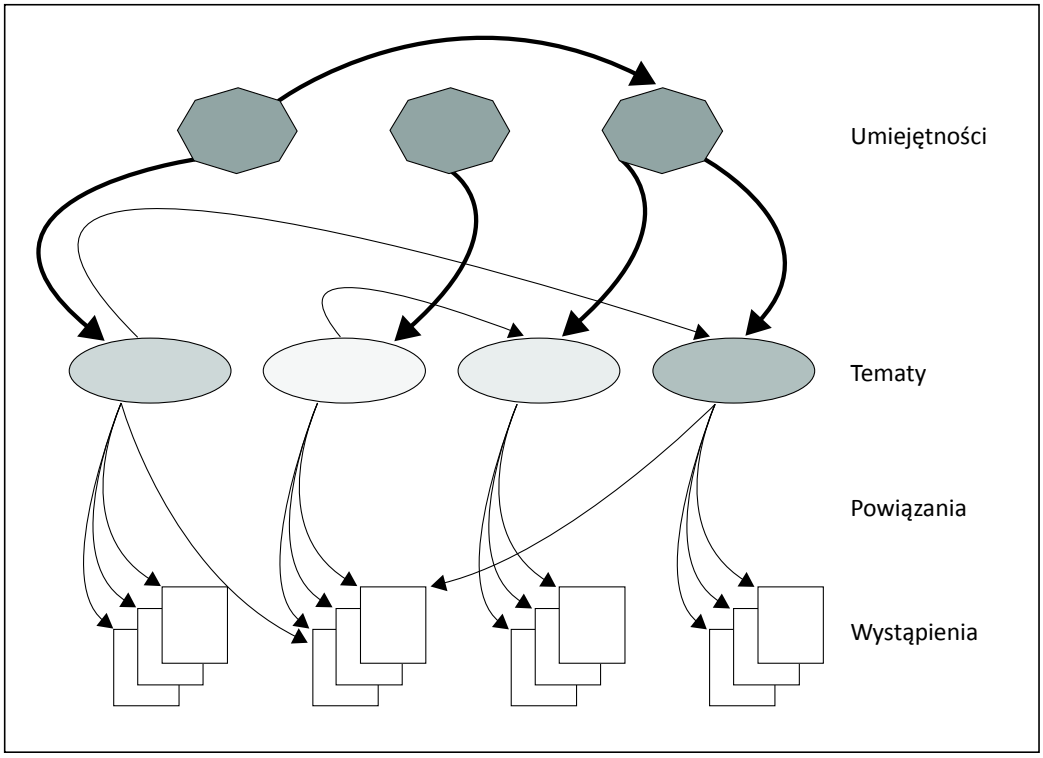
Rysunek 24. Struktura map tematów [T. Jakubowski, 2004]

Tematy charakteryzują się nazwą, wystąpieniem i rolą w powiązaniu. Każdy temat może mieć trzy nazwy: nazwę podstawową (tylko ta nazwa jest obligatoryjna, *base name*), nazwę prezentowaną (na przykład użytkownikowi w aplikacji, *display name*) i nazwę stanowiącą kryterium sortowania (*sort name*). Temat może być powiązany z jednym lub wieloma źródłami, nazywanymi wystąpieniami tematu. Wystąpienie może na przykład wyjaśniać temat. Z drugiej strony wystąpienie może być relewantne dla tematu. Stwarza to możliwość kategoryzacji wystąpień przez tematy. Powiązania mogą być typizowane. Typ powiązania może wskazywać semantykę powiązania pomiędzy tematem a wystąpieniem. Wykracza to poza koncepcję nietypizowanych powiązań w Internecie, ale nawiązuje do typizacji powiązań w hipertekście (porównaj na stronie 76).

[107] http://www.y12.doe.gov/sgml/sc34/document/0058.htm

Mapy tematów mogą być budowane automatycznie, na przykład z wykorzystaniem metod taksonomicznych [W. Abramowicz i inni, 2004c], współwystąpień [W. Abramowicz i inni, 2003e] oraz odkrywania wiedzy [W. Abramowicz i inni, 2004e].

Mapy tematów prócz reprezentacji ontologii, mogą być wykorzystane w zarządzaniu obiektami trudno strukturalizowanymi, na przykład wielowątkowymi dokumentami bez jawnej struktury wewnętrznej [W. Abramowicz i inni, 2002b], obiektami będącymi przedmiotem obrotu na rynkach elektronicznych [W. Abramowicz i inni, 2002d], reprezentacją materiałów dydaktycznych [W. Abramowicz i inni, 2003f], oraz wspomaganiem efektywności procesu dydaktycznego [W. Abramowicz i inni, 2003g]. Do budowy map tematów można wykorzystać opisane przez nas metody fragmentacji i podziału tematycznego dokumentów (porównaj na stronie 90). Wystąpienia nie muszą dotyczyć całych dokumentów, ale rozpoznanych w nich fragmentów tematycznych.



Rysunek 25. Struktura map umiejętności [W. Zalech, 2004]

Właśnie ten ostatni obszar zastosowań zainspirował nas do rozszerzenia koncepcji map tematów o czwarty element – umiejętności (*skills*). Umiejętności mogą być przypisane ludziom lub organizacjom. Prezentują zdolność

do wykonania działania, opanowanie wiedzy. Tę nową strukturę nazwaliśmy mapami umiejętności (*skill maps*).

Mapy umiejętności mogą być stosowane do zarządzania wiedzą w podmiotach gospodarczych [W. Abramowicz i inni, 2004d] oraz do komunikowania wiedzy w kooperujących środowiskach [W. Abramowicz i inni, 2003d]. Budowanie odpowiednich ścieżek przez umiejętności może sterować dynamicznie zmieniającym się przepływem pracy (*workflow*) w procesie dydaktycznym [W. Abramowicz i inni, 2002b].

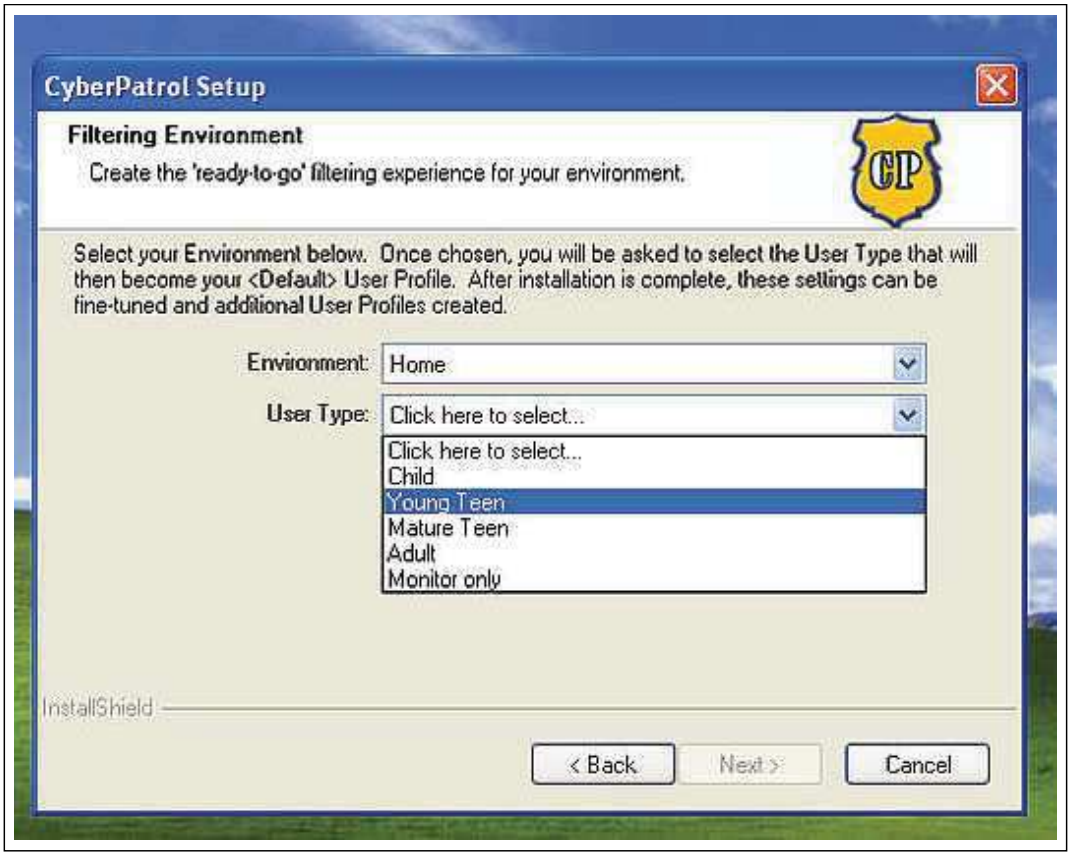
Zarówno mapy tematów, jak i mapy umiejętności mogą być opisywane przez języki znaczników, na przykład XML [J. Park i inni, 2002].

Delinearyzacja i linearyzacja

Pokazując związki hipertekstu z retoryką (porównaj na stronie 81) wskazaliśmy na teksty jako częste źródło tworzenia hipertekstu. Proces ten nazywamy delinearyzacją tekstu w hipertekst lub jego fragmentaryzacją – pojęcia te będziemy używali wymiennie. Istnieje wiele rekomendacji dla delinearyzacji. Jedne zakładają interaktywność tego procesu [R. Kuhlen, 1991], inne dążą do jego automatyzacji [M.H. Chignell i inni, 1991], [Y. Hara i inni, 1991]. My wyróżniamy w nim cztery etapy [T. Tomaszewski, 1999]. Na początku określamy węzły według wskazanych uprzednio zasad (porównaj na stronie 75). Często inspiracją jest sama struktura tekstu. Węzłami stają się na przykład rozdziały. Drugim etapem jest stworzenie powiązań pomiędzy węzłami. Jeżeli inspiracją do tworzenia węzłów była struktura tekstu, to możemy ją teraz wykorzystać do pokazania związków strukturalnych pomiędzy poszczególnymi częściami tekstu. W ten sposób powstają powiązania strukturalne. Stosując opisane później metody indeksowania (porównaj na stronie 163), możemy zbudować powiązania semantyczne. Ponieważ nie wszystkie odkryte powiązania semantyczne będą wykorzystywane, a ich nadmiar może doprowadzić do złożoności hipertekstu, która uniemożliwi jego wykorzystywanie, stosuje się różne techniki badania przydatności rozpoznanych powiązań semantycznych dla antycypowanych potrzeb informacyjnych użytkowników (porównaj na stronie 212). Ostatnim, interaktywnym, etapem delinearyzacji jest dodawanie przez autorów hipertekstu subiektywnych powiązań, służących na przykład pragmatykom wykorzystania hipertekstu [W. Abramowicz, 1993a].

W opisanym powyżej procesie założyliśmy udział autorów hipertekstu w jego tworzeniu. Często jednak – choćby ze względu na objętość tekstów źródłowych – udział człowieka jest niemożliwy lub nieuzasadniony ekonomicznie. Przedmiotem naszych badań były dokumenty prawne, które ze

może także wskazać profil stereotypowy, którym użytkownik ma się posługiwać. Można dopuścić adaptację profili stereotypowych przez użytkownika do jego indywidualnych potrzeb lub nie. Profile stereotypowe mogą być indywidualizowane na podstawie pewnych charakterystyk socjologicznych każdego z użytkowników [B. Shapira i inni, 1997]. Szczególnym przypadkiem profili stereotypowych są opisane niżej profile korporacyjne.



Rysunek 81. Wykorzystanie stereotypów profili do filtrowania treści z Internetu

Profile korporacyjne

Organizacja może budować profile, wyrażające potrzeby informacyjne jej członków. Przykładem takich profili są profile korporacyjne. Dokumenty filtrowane na ich podstawie mogą być udostępniane wszystkim pracownikom korporacji. Mogą one jednak podlegać specjalizacji. Ogólny profil korporacyjny jest na tyle szeroki, że odpowiada potrzebom informacyjnym wszystkich pracowników. Twórca takiego profilu zakłada, że wszyscy pracownicy powinni otrzymywać dokumenty, zawierające pewne minimum informacji. Ogólne profile korporacyjne uzupełniane są w poszczególnych jednostkach organizacyjnych. Nie zakłada się jednak usuwania z nich składników profilu korporacyjnego. Takie specjalizowane profile nazywa-

my wydziałowymi. Specjalizacja profili może dotyczyć nie tylko struktur organizacyjnych, ale także struktur procesowych. Nie odnosi się wówczas do pracowników określonych jednostek organizacyjnych, lecz uczestników poszczególnych procesów. Specjalizacja profili wydziałowych może być wieloetapowa, tak jak wielopoziomowa może być struktura organizacyjna korporacji. Ostatnim etapem specjalizacji profilu korporacyjnego może być specjalizacja dokonywana przez pojedynczego pracownika. W jej wyniku otrzymujemy profil stanowiska – w przypadku specjalizacji funkcjonalnej, lub profil roli – w przypadku specjalizacji procesowej. Oczywiście dla poszczególnych pracowników specjalizacja może osiągnąć różny poziom w szczegółowości. Na przykład kasjerzy bankowi mają zapewne w banku wspólne profile, jeżeli ich zakres działania nie podlega różnicowaniu. Analitycy giełdowi muszą specjalizować swoje profile indywidualnie, ponieważ tylko w bardzo dużych bankach są grupy analityków zajmujących się dokładnie tym samym obszarem działania. Wówczas profil stanowiska w jej części specjalizowanej nie różni się od profilu formułowanego przez indywidualnego użytkownika filtru [B.D. Sheth, 1994].

Specjalizowane profile tworzą hierarchię w organizacjach zorganizowanych funkcjonalnie. W organizacjach procesowych struktura profili jest heterarchią, jeżeli poszczególni pracownicy uczestniczą w ogólności w realizacji różnych procesów. Konieczne jest wówczas wprowadzenie reguł dziedziczenia profili z poszczególnych gałęzi hierarchii. Jest to szczególnie istotne, gdy następują sprzeczności w dziedziczeniu.

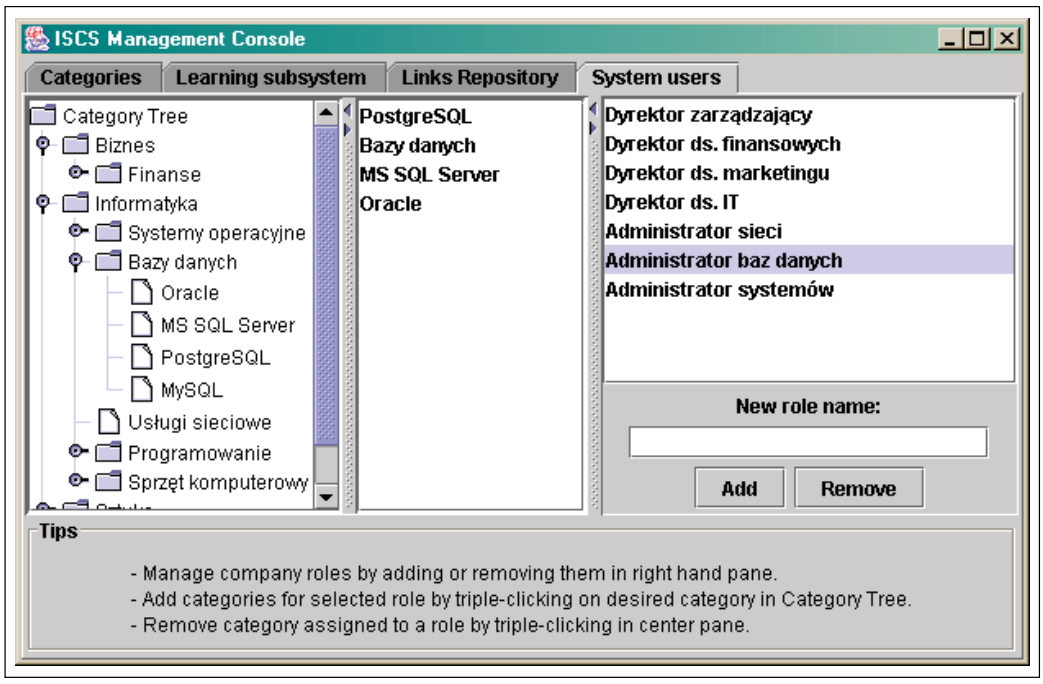
W przygotowanym przez nas ISC – Internet Sources Categorization System – poszczególnym pracownikom można przyporządkować kategorie dokumentów, które powinny być przedmiotem ich nadzwyczajnego zainteresowania (porównaj rysunek 82). Wówczas tworzenie stereotypowych profili i profili korporacyjnych obywa się przez wskazywanie kategorii w reprezentacji informacji filtru informacyjnego. W koncepcji UIH (*user interest hierarchy*) do poszczególnych kategorii w hierarchii dodano zbiory dokumentów relewantnych dla określonego poziomu kategorii. Dokumenty przyporządkowywane są przez budowanie odpowiednich skupień (DHC – *divisive hierarchical clustering*) [H.-R. Kim i inni, 2007]. Klasyfikacja wykorzystywana jest także jako pierwszy etap w budowaniu profili dla środowisk tak dynamicznie zmieniających się jak blogi internetowe [M. Nakatsuji i inni, 2006].

Prócz profili kognitywnych rozróżniamy także inne rodzaje profili.

Profile społeczne

Profile społeczne określają role, w jakich znajduje się użytkownik, na przykład: *jestem zainteresowany dokumentami od mojego szefa i tymi, które*

krążą w firmie a dotyczą mnie bezpośrednio. Nawiązują one do koncepcji indeksów pragmatycznych (porównaj na stronie 185). Profile społeczne różnią się od stereotypów profili – o ich wyborze nie decydują w pierwszej kolejności rzeczywiste potrzeby informacyjne, lecz przynależność do określonej społeczności.



Rysunek 82. Przypisywanie kategorii poszczególnym stanowiskom w przedsiębiorstwie w ISC [D. Rutkowski, 2003]

Ciekawym podejściem jest połączenie profili społecznych z kognitywnymi dla sieci *peer to peer*. Najpierw użytkownicy przyporządkowani są do *Personalized Term Pattern trees*, odpowiadających profilom społecznym. W tak określonych strukturach używa się profili kognitywnych. Rozwiązanie takie może znacznie ograniczyć przesyłanie duplikowanych danych w sieciach *peer to peer* [H.-N. Kim i inni, 2004].

Fenomen feromonowy

Rola profili społecznych będzie rosnąć w *Web 2.0* w zależności od wpływu fenomenu nazywanego przez nas feromonowym. Fenomenem feromonowym określać będziemy sytuacje premiujące budowanie profili odróżniających ich właścicieli od innych członków ich społeczności. Społecznie korzystne jest pozyskiwanie informacji wyróżniającej jej użytkowników od innych. Zwiększa to szanse na tworzenie unikatowej wiedzy. Obecne tendencje w budowaniu profili społecznych nie są zgodne z tym fenomenem: sprzyjają pozyskiwaniu informacji już znanych społeczności właściciela

profilu społecznego [K.-Y. Jung i inni, 2003b]. Przykładem takiego zastosowania jest edukacja. Proces dydaktyczny może być wsparty profilem społecznym. Jest skuteczny wtedy, gdy kształcimy masowo [M.M. Recker i inni, 2003]. Jesteśmy przekonani, że w przypadku fenomenu feromonowego profile społeczne wspomagałyby różnicowanie nauczanych treści, obniżając przy tym koszty kształcenia. Świadomość fenomenu feromonowego musi wpływać na działania Obywateli Tworzących produkty i usługi cyfrowe służące zaspokajaniu potrzeb Obywateli Informujących się, Komunikujących się i Uczących się.

Profile ekonomiczne

Ze względu na rosnącą świadomość kosztu relewancji informacji (porównaj na stronie 364) wprowadźmy profil ekonomiczny użytkownika, uwzględniający różne miary ekonomiczne związane z filtrowaniem informacji, na przykład związane z funkcją nakładów i użyteczności. Możemy w profilu ekonomicznym definiować funkcję, która zmienia wartość relewancji w zależności od kosztu dokumentu (mierzonego na przykład kosztem jego pozyskania: opłata za pobranie dokumentu, opłata za filtrowanie dokumentu, koszt transmisji, czas czytania i oceny dokumentu – porównaj na stronie 365). Z naszych prac wynika, że profile ekonomiczne mogą być sterowane pomiarami telemetrycznymi, będącymi źródłem informacji o koszcie dokumentów [T. Jankowski, 2002]. W tak oczywisty sposób relacji pomiędzy kosztem dokumentu i jego relewancją nie formułuje się w systemach wyszukiwawczych. W profilu ekonomicznym umieszczamy wiele elementów kontekstów fizycznych. Pewne z nich są trudnomierzalne, na przykład wpływ wielkości monitora urządzenia mobilnego na czas i w konsekwencji na koszt oceny dokumentu. Profil ekonomiczny może być istotnym narzędziem dla Obywateli Uczących się (porównaj na stronie 214), chcących stosować miernik *return of investment in knowledge* – ROIK) [W. Abramowicz i inni, 1997c]. Profil ekonomiczny użytkownika wspomaga obliczenie relewancji ekonomicznej (porównaj na stronie 49).

Konteksty

Kontekstem jest każda informacja opisująca sytuacje, w jakiej znajdują się osoby i inne obiekty opisywane przez systemy informacyjne [G.D. Abowd i inni, 1999]. Wskazuje się na różnorakie konteksty. Przykładowo kontekstem użytkownika jest miejsce, w jakim się znajduje, ludzie w jego otoczeniu oraz sytuacja społeczna. Wydajność sieci komputerowej, z której korzysta użytkownik, dostępne urządzenia i koszt ich użytkowania nazywa się kontekstem sieciowym. Wyróżnia się także kontekst ludzki, składający się z informacji o użytkowniku, jego wiedzy, przyzwyczajeniach, otoczeniu społecznym oraz realizowanych zadaniach [A. Schmidt i inni, 1998]. [W.N. Schilit, 1995] wyróżnia:

- kontekst zasobów informatycznych –znajdujące się w pobliżu drukarki, stacje robocze, połączenia sieciowe, ich przepustowość oraz koszty itp.,
- kontekst użytkownika – osoby w pobliżu, sytuację socjalną,
- kontekst fizyczny – oświetlenie, poziom hałasu, temperatura, zatłoczenie na ulicach [W.N. Schilit, 1995],
- kontekst czasu rozumiany jako pora dnia, tygodnia czy roku pozwalająca na odwzorowanie kontekstów: informatycznego, użytkownika oraz fizycznego na przestrzeni czasu.

Zmiany kontekstów mogą być zapisywane. W ten sposób otrzymujemy *historię kontekstów*, która może się okazać przydatna w badaniu ich zmian [G. Chen i inni, 2000].

Systemy zorientowane na kontekst

Postuluje się, aby systemy informacyjne uwzględniały konteksty użytkowników. Systemy spełniające ten postulat nazywa się zorientowanymi na kontekst (*context-aware*) [P.J. Brown i inni, 2001]. Postulujemy uwzględnianie kontekstów w profilach. Zorientowanie na kontekst może dotyczyć także samej prezentacji dokumentów użytkownikom, może powodować automatyzacje realizacji pewnych usług, może w końcu oznaczać przyporządkowanie kontekstu użytkownika do dokumentu, powodując jego przetwarzanie odniesione do warunków konkretnego użytkownika [G.D. Abowd i inni, 1999]. W pracy [M. Korkea-aho, 2000] można znaleźć analizę takich systemów.

Postulat zorientowania na kontekst jest szczególnie trudny do spełnienia dla systemów mobilnych, ponieważ wiele składników kontekstu zmienia się dla nich szybciej niż w systemach stacjonarnych, na przykład kontekst sieciowy lub kontekst społeczny użytkowników. Zmienia się także szybciej niż profil kognitywny, będący wyrazem długoterminowych potrzeb informacyjnych użytkownika. Problem ten omówimy na przykładzie koncepcji mobilnego filtrowania informacji – *mobileIF* (porównaj na stronie 421).

Profil czasowy

Profil czasowy użytkownika umożliwia obliczenie relewancji czasowej (porównaj na stronie 47). Dla skutecznej realizacji tego celu model czasu używany w budowaniu indeksów czasu musi być spójny z modelem czasu wykorzystywanym przez profile czasowe (porównaj na stronie 190). Użyteczne jest takie zbudowanie języka zapytań dla profili czasowych, aby użytkownicy mogli pytać o punkty i interwały czasowe (porównaj na stronie 190). Zaawansowane języki zapytań powinny dopuścić uwzględnienie zarówno ciągłości, jak i dyskretności czasu (porównaj na stronie 193).

Wyszukiwarka *Google* pokazuje, jak działają zapytania, uwzględniające czas (porównaj rysunek 63, rysunek 65 oraz rysunek 83). W podobny sposób wykorzystuje się profil czasowy.

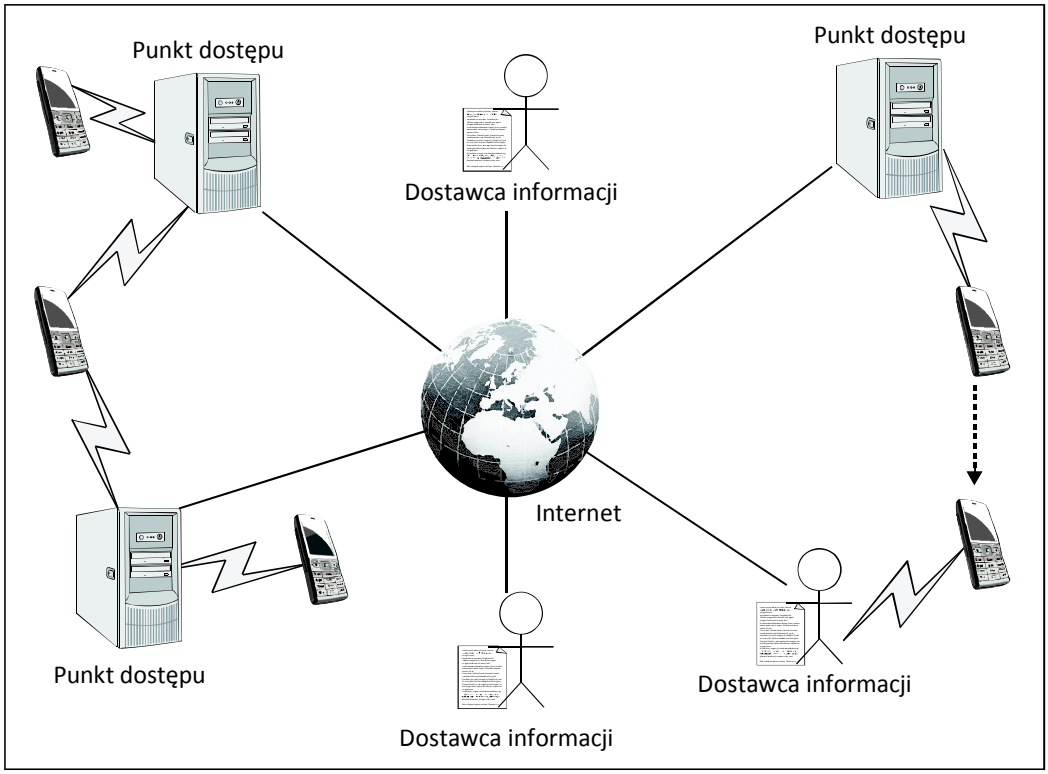
Profil czasowy może wskazywać na czas oddziaływania dokumentu, najkorzystniejszy czas wykorzystania informacji z dokumentu (porównaj na stronie 190). Takie podejście może być wykorzystane w sposób opisany w profilu filtru mobilnego (porównaj na stronie 251).

Profil osobowy użytkownika

Profil osobowy użytkownika służy do określenia relewancji osobowej (porównaj na stronie 48). Jest odpowiednikiem kontekstów społecznych, które mogą być odniesione do osób fizycznych, podmiotów gospodarczych, organizacji, produktów i usług.

Profil lokalizacyjny użytkownika

Profil lokalizacyjny użytkownika pomaga wyliczyć relewancję lokalizacyjną (porównaj na stronie 48). Język zapytań wykorzystywany do formułowania profili lokalizacyjnych musi uwzględniać model opisu przestrzeni geograficznej, uwzględnianej w filtrowanych dokumentach (porównaj na stronie 202).



Rysunek 84. Dostawcy informacji niezwiązani z lokalizacją [C. Panayiotou i inni, 2004]

Spam

Najpowszechniej stosowanym przykładem profili negatywnych są profile wykorzystywane w filtrach antyspamowych. Przez spam, dla naszych potrzeb, będziemy rozumieć przesyłki poczty elektronicznej przekazywane osobom wbrew ich woli[346]. Najczęściej spam kierowany jest do bardzo dużej liczby odbiorców.

Przesyłki te mogą mieć różnych charakter: komercyjny, społeczny, obyczajowy czy pornograficzny. Spam może obejmować dokumenty, które przesyłane są w formie niespersonalizowanej do przypadkowych odbiorców lub spersonalizowanej, wynikającej z obserwacji zachowań odbiorców lub z analizy informacji przekazywanej przez nich albo też wyszukiwanej w Internecie. Odebranie przesyłki pocztowej będącej spamem może być nie tylko związane ze stratą czasu przeznaczanego na jej przetworzenie, ale także z przekazaniem informacji o odbiorcy przesyłki do nadawcy. Informacjami tymi mogą być logi, ale może to być też przekazanie haseł dostępu do zasobów nimi chronionymi, na przykład do systemu bankowości elektronicznej.

O znaczeniu filtrów antyspamowych niech świadczy ilość spamu oraz dynamika jego wzrostu. W czerwcu 2005 roku było to *tylko* 30 miliardów obiektów dziennie, a w lutym 2007 roku szacowano już liczbę przesyłanego spamu na 90 miliardów obiektów dziennie[347]. Ocenia się, że 80-85% elektronicznych przesyłek pocztowych stanowi spam[348]. Rynek oferuje bardzo dużo narzędzi przeciwdziałających otrzymywaniu spamu, na przykład *Symantec Antispam*[349], *Horton AntiSpam*[350].

---

[346] Pojęcie spam nawiązuje do sprzedawanej w USA mielonej szynki w puszkach (SPAM – Shoulder Pork and hAM). Była przedmiotem intensywnej reklamy (Well, we have spam, tomato & spam, egg & spam, egg, bacon & spam... Spam spam spam spam, spam spam spam spam, lovely spam, wonderful spam...) stając się synonimem natrętnego przekazywania komunikatu reklamowanego. Po polsku czasami mówi się o informacji niezamówionej. Wobec powszechności użycia będziemy posługiwali się jednak pojęciem spamu.

[347] http://en.wikipedia.org/wiki/E-mail_spam

[348] http://www.maawg.org/about/FINAL_1Q2006_Metrics_Report.pdf

[349] http://www.symantec.com/region/pl/plprod/bright_antispam.html

[350] http://www.symantec.com/pl/pl/home_homeoffice/products/overview.jsp?pcid=is&pvid=nas2005

Profile negatywne filtrów antyspamowych mogą odfiltrowywać przesyłki pocztowe od nadawców umieszczonych na DNSBL (*DNS Blacklist*), będących listą adresów IP nadawców spamu. Wobec ciągłego zmieniania przez nadawców spamu swoich adresów IP i wynikających z tego trudności w zarządzaniu aktualną DNSBL, skuteczniejszymi filtrami antyspamowymi są filtry, posługujące się GL (*Graylist*). W tym wypadku filtr odrzuca przesyłki pocztowe, pochodzące od nieznanych nadawców. Poprawnie pracujące serwery poczty po kilku godzinach ponownie przesyłają przesyłkę. Źródła spamu z reguły tego nie robią. Kosztem takiego działania jest opóźnienie przekazywania podejrzanych przesyłek o czas potrzebny na ponowne przesłanie przez źródło.

Innym mechanizmem jest budowanie negatywnych filtrów, wykorzystujących w swoich profilach pojęcia często stosowane w spamie, takie jak: *winner*, *pills* czy *viagra*. W profilu takim mogą być także umieszczone adresy nadawców poczty, które nie są źródłem spamu kierowanego do dużej liczby odbiorców, ale są nimi z subiektywnego widzenia konkretnego odbiorcy, na przykład sprzedawca polis ubezpieczeniowych kierujący po raz kolejny niechcianą ofertę do utraconego klienta.

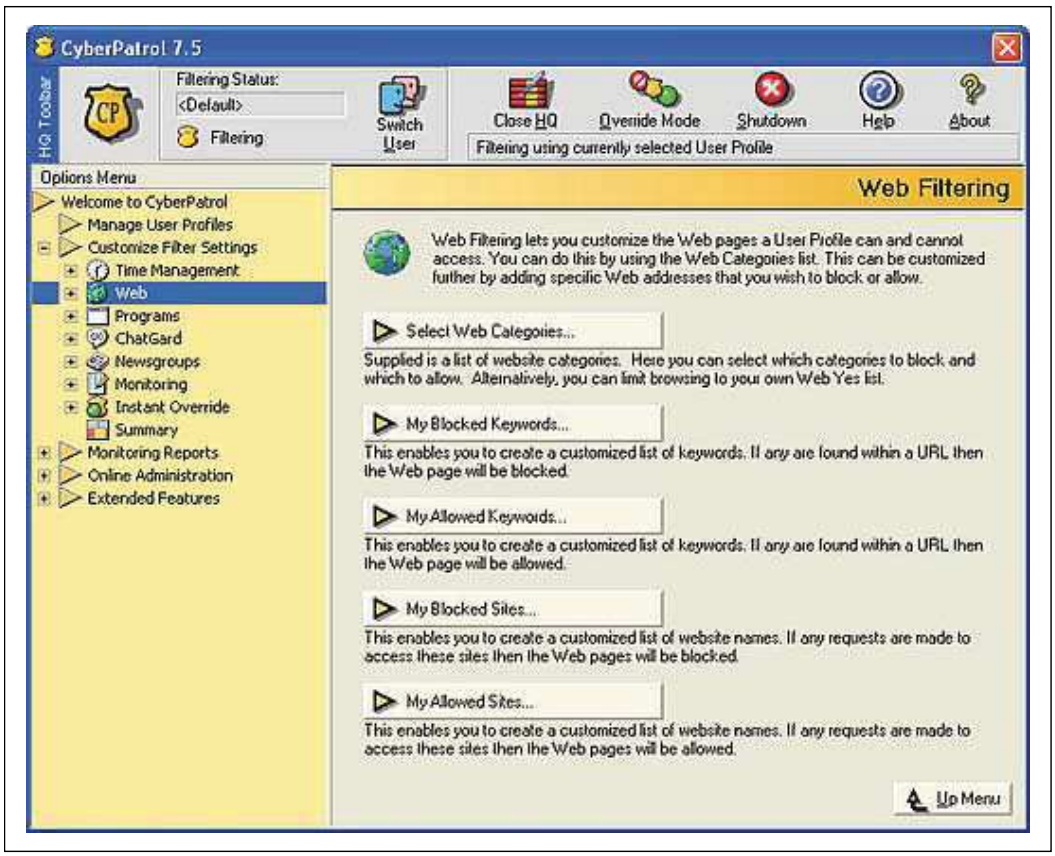
Filtrowanie i blokowanie dostępu do Internetu

Powszechność Internetu powoduje, że nie wszystkie treści w nim dostępne są przez wszystkich akceptowane. Część użytkowników nie chce o tych treściach nic wiedzieć, dążąc do ich zablokowania przez wyłączenie w trakcie wyszukiwania, wertowania lub nawigowania. Treści może nie tylko blokować sam użytkownik, ale także inni – wpływając na treści przez niego odbierane. Przykładami takich sytuacji mogą być:

- rodzice ograniczający treści dostępne dla swoich dzieci przez zakaz dostępu do informacji promujących przemoc,
- politycy usiłujący zakazywać rozpowszechniania treści pornograficznych dzieciom (filtry w szkołach) lub obywatelom (filtrowanie prawie wszystkim mieszkańcom Chin określonych informacji politycznych),
- instytucje państwa utrudniające działania przestępcze (rozpowszechnianie narkotyków czy handel bronią),
- pracodawcy zakazujący dostępu pracownikom do określonych źródeł w pracy (statystki pokazują, że w wielu krajach najczę-

ściej w trakcie pracy odwiedzane strony, które zawierają treści pornograficzne[351]).

Przyczyny ograniczania dostępu do informacji mogą być zatem różne: moralne i etyczne, wynikające z norm prawnych, polityczne, gospodarcze, wynikające z kodu kulturowego. Nie ma oczywiście uniwersalnych norm, określających treści niepożądane lub zakazane. Dlatego wiele ze środowisk – poczynając od rodziny, a kończąc na blokach państw – chce budować narzędzia kontroli dostępu do Internetu na własny sposób.

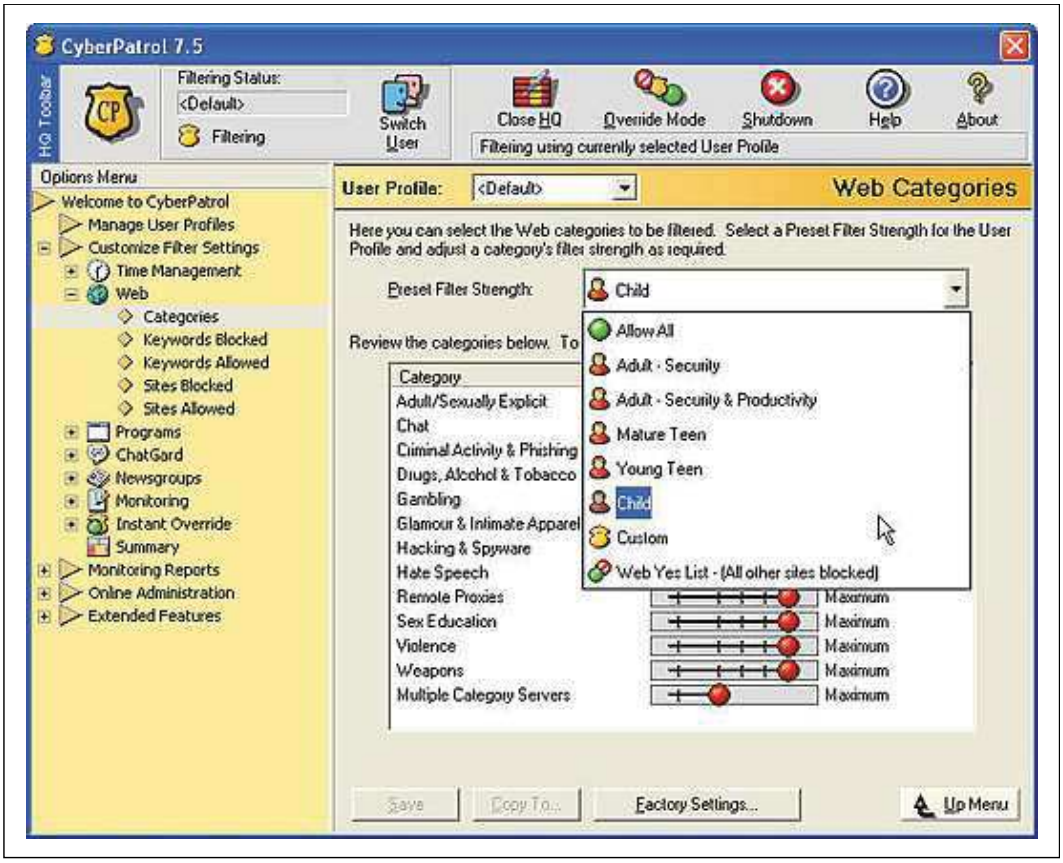


Rysunek 145. *CyberPatrol* filtrujący treści z Internetu[352]

Najprostszym sposobem ograniczenia dostępu do niepożądanych treści jest wyliczenie niepożądanych stron internetowych w wyszukiwarce internetowej. Wiele wyszukiwarek oferuje takie możliwości. Sposób ten jest jednak pracochłonny i nieskuteczny, dlatego że trzeba wyliczyć każdą ze stron.

[351] http://healthymind.com/s-porn-stats.html

[352] http://www.cyberpatrol.com/



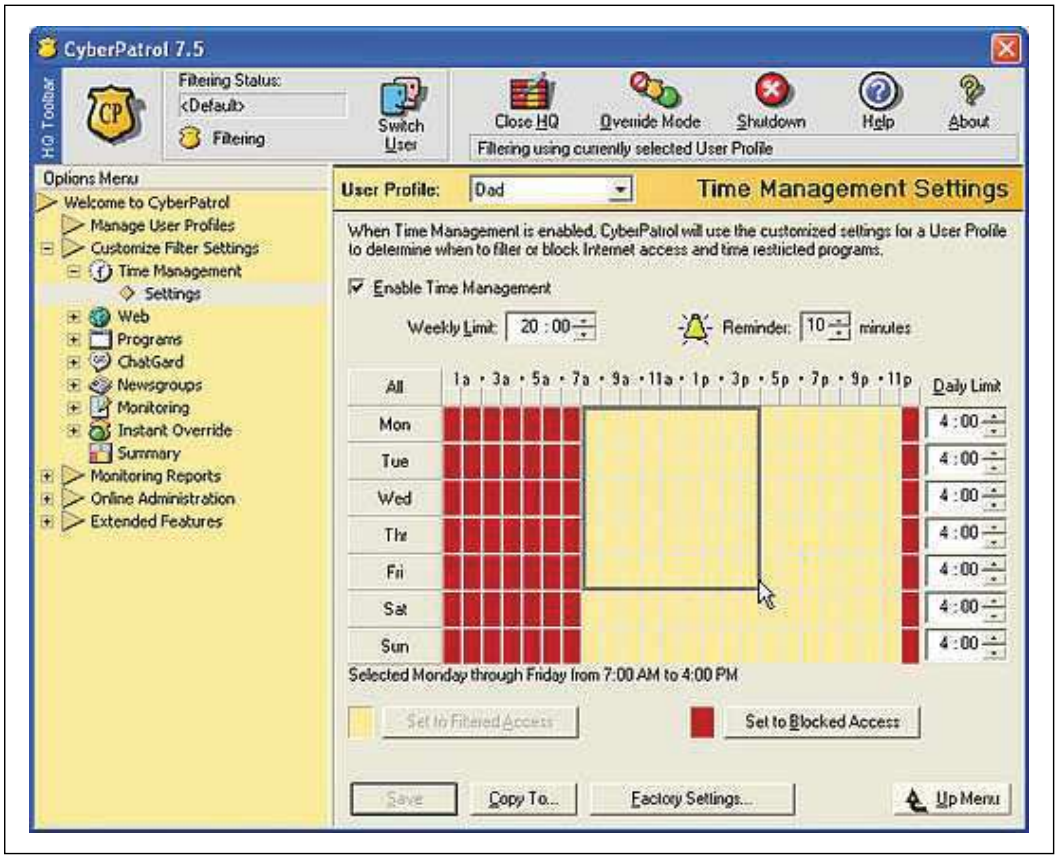
Rysunek 146. Język symboli graficznych ułatwiający różnicowanie restrykcji w dostępie do Internetu na przykładzie systemu *CyberPatrol*

Negatywne filtry informacyjne są często stosowanym narzędziem blokowania dostępu do określonych treści w Internecie. W celu ich skutecznego wykorzystania określa się standardy treści nieakceptowanych, formułowane przez różne środowiska. Budują je na przykład społeczności rodzicielskie, określające programy kontroli rodzicielskiej (*parental control programs*). *The SafeSurf Internet Rating Standard* proponuje systematykę treści nieakceptowanych, począwszy od stron posługujących się wulgarnym językiem, poprzez materiały dotyczące seksu i przemocy podane w sposób wulgarny, a skończywszy na pochwałach nielegalnego używania narkotyków[353].

World Wide Web Consortium promuje standard techniczny PICS (*Platform for Internet Content Selection*), wykorzystujący metadane do oznaczania treści rozpowszechnianych na poszczególnych stronach [P. Resnick i inni,

[353] http://www.safesurf.com/ssplan.htm

1996][354]. Ważnym założeniem standardu nie jest ingerowanie w treści dostępne w Internecie, lecz ograniczanie ich rozpowszechniania wśród określonego grona odbiorców. PICS nie formułuje polityki restrykcji dostępu, ale stara się dostarczyć narzędzia do jej realizacji. Wadą takiego rozwiązania jest konieczność zbudowania kategoryzacji, wykorzystywanej przez metadane [V.S. Jacob i inni, 2007].



Rysunek 147. Działanie filtru różnicującego czas dostępu do Internetu na przykładzie systemu *CyberPatrol*

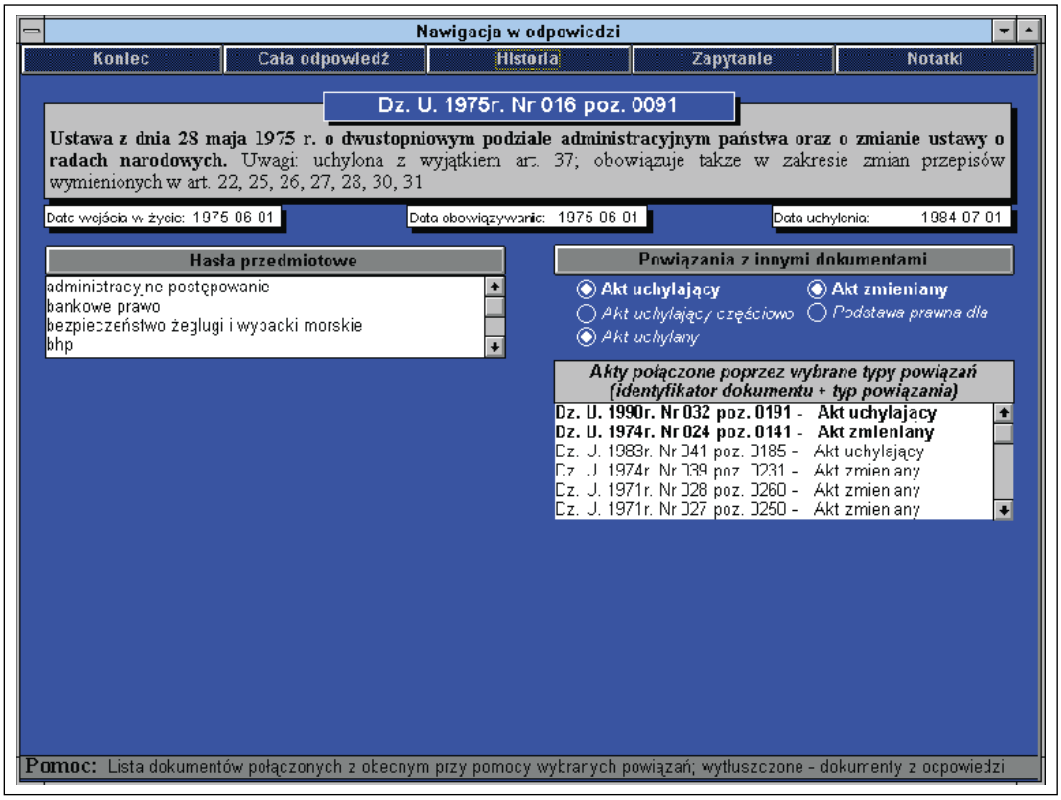
CyberPatrol[355] jest przykładem narzędzia wykorzystującego taką kategoryzację. Prócz opisu metadanych można podać słowa kluczowe oraz adresy stron o niepożądanych treściach (porównaj rysunek 145). Znaczenie poszczególnych ograniczeń można różnicować przez nadanie im wag. System *CyberPatrol* stosuje symbole graficzne dla ułatwienia różnicowania restrykcji (porównaj rysunek 146). Działanie filtru może być różnicowane

[354] http://www.w3.org/PICS/

[355] http://www.cyberpatrol.com/

Z zasobów bibliotek elektronicznych można korzystać także w sposób mobilny. Profil technologiczny powinien być wykorzystany dla kierowania odpowiednich dokumentów z biblioteki do urządzeń użytkownika, na których wygodnie będzie mu się z nimi zapoznać (porównaj na stronie 249)[418].

Filtry społeczne mogą wspomagać budowanie społeczności czytelniczej, pogłębiającej jakość wyszukiwania i filtrowania przez stosowanie sprzężenia zwrotnego relewancji w celu doskonalenia profili (porównaj na stronie 373) [U. Rohini i inni, 2005].



Rysunek 180. Wspomaganie nawigacji relacjami pomiędzy aktami prawnymi w systemie *HyperThemis*

Filtry informacyjne mają szczególne zastosowanie dla bibliotek elektronicznych gromadzących dokumenty, które powinny być możliwie szybko przekazane użytkownikom. Dobrym przykładem takiego zastosowania jest informacja prawna, polegająca na rozpowszechnianiu dokumentów zwią-

[418] O wymiarze tego rynku niech świadczą już teraz tłumy pasażerów w metrach japońskich i chińskich, cztających *tankōbon* w telefonach komórkowych. Z myślą o nich tworzone są czytniki *e-books*, efektownie prezentujące grafikę (porównaj rysunek 178).

zanych z prawem stanowionym, czyli aktami ustawodawczymi, do których zaliczamy konstytucję, ustawy i inne akty wydawane z mocą ustawy, oraz akty wykonawcze czyli rozporządzenia, zarządzenia i uchwały (dla ułatwienia pomijamy rozważania o aktach prawnych indywidualnych oraz orzecznictwie). Należy zauważyć, że dokumenty te stanowią strukturę hierarchiczną. Wyliczyliśmy już uprzednio relacje pomiędzy aktami prawnymi (porównaj rysunek 31). Powinny być one uwzględnione w procesie wyszukiwania w bibliotece. W zbudowanym przez nas systemie informacji prawnej *HyperThemis* zaprojektowaliśmy zapytania, wykorzystujące tę strukturę (porównaj rysunek 79) [W. Abramowicz i inni, 1998b]. Wspomaga ona także nawigację w odpowiedzi na zapytanie (porównaj rysunek 180).

Uwzględnienie tej struktury w profilach powodowałoby niestety wielokrotne przesyłanie tych samych dokumentów, na przykład ustaw stanowiących podstawę prawną dla wielu aktów prawnych. Skalę zagadnienia ilustruje rozkład liczby odwołań i referencji wychodzących z aktów prawnych (porównaj rysunek 181).

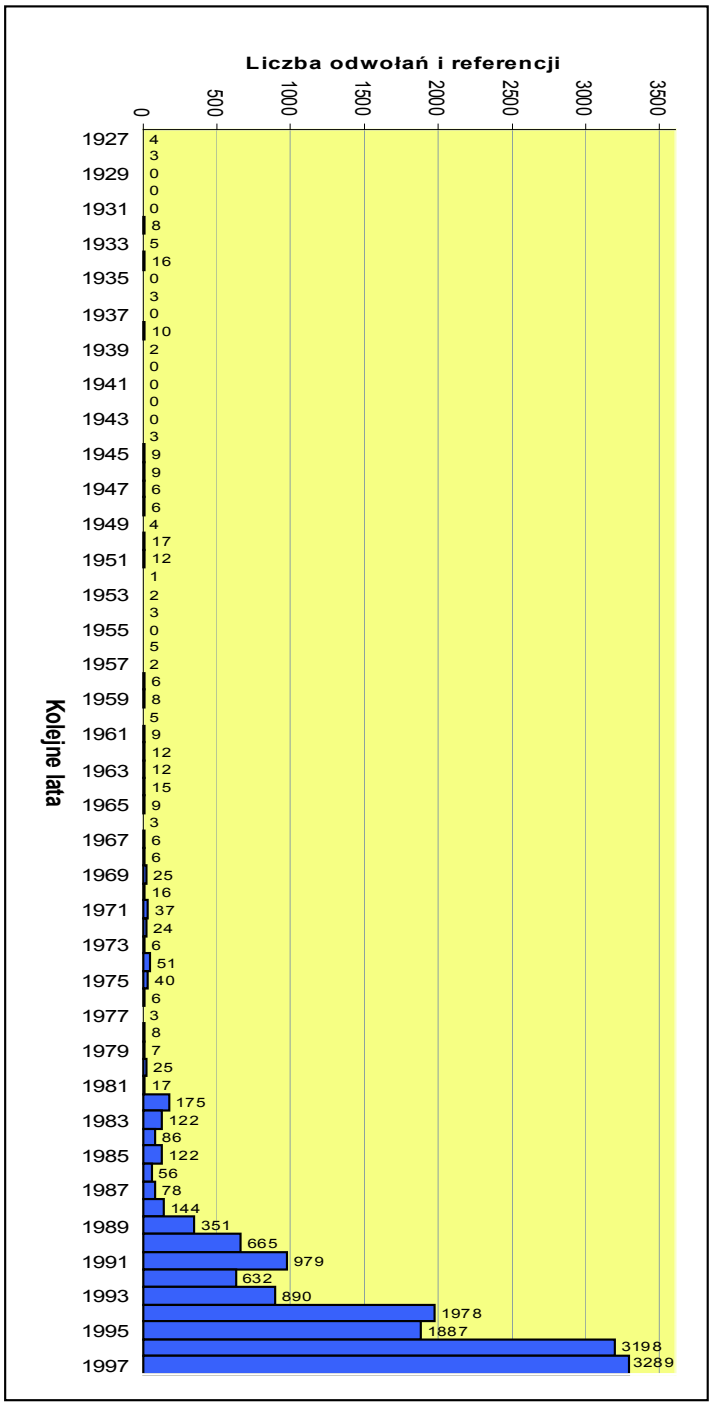
Drugim problemem jest filtrowanie aktów nowelizujących. Najczęściej są one trudno zrozumiałe bez możliwości ich porównania z aktem nowelizowanym, ponieważ zawierają na przykład sformułowania: *usuwa się §12* lub *do §12 dodaje się „lub czasopisma”*. Dlatego zaprojektowaliśmy i zaimplementowaliśmy narzędzia do tworzenia aktów ujednoliconych, powstających z aktów ujednolicanych po uwzględnieniu zmian wynikających z aktów ujednolicających [D. Jakubowska, 1993][419].

Dokumenty prawne mają strukturę syntaktyczną określoną normą prawną[420], która pozwala zapisywać takie dokumenty jako struktury relacyjne (porównaj rysunek 182). Rzadko dokument prawny wykorzystywany jest w całości. Częściej potrzebny jest jego fragment, połączony z innymi fragmentami dokumentów, będących jego podstawą prawną, aktem uchylonym czy aktem zmieniającym. Taka struktura tworzy normę prawną, a jej uzyskanie jest celem wyszukiwania i filtrowania. Ponieważ takie czytanie wynika z konstytucji prawa i jego pragmatyki, możliwe jest projektowanie dokumentów składających się z odpowiednich fragmentów i ich tworzenie

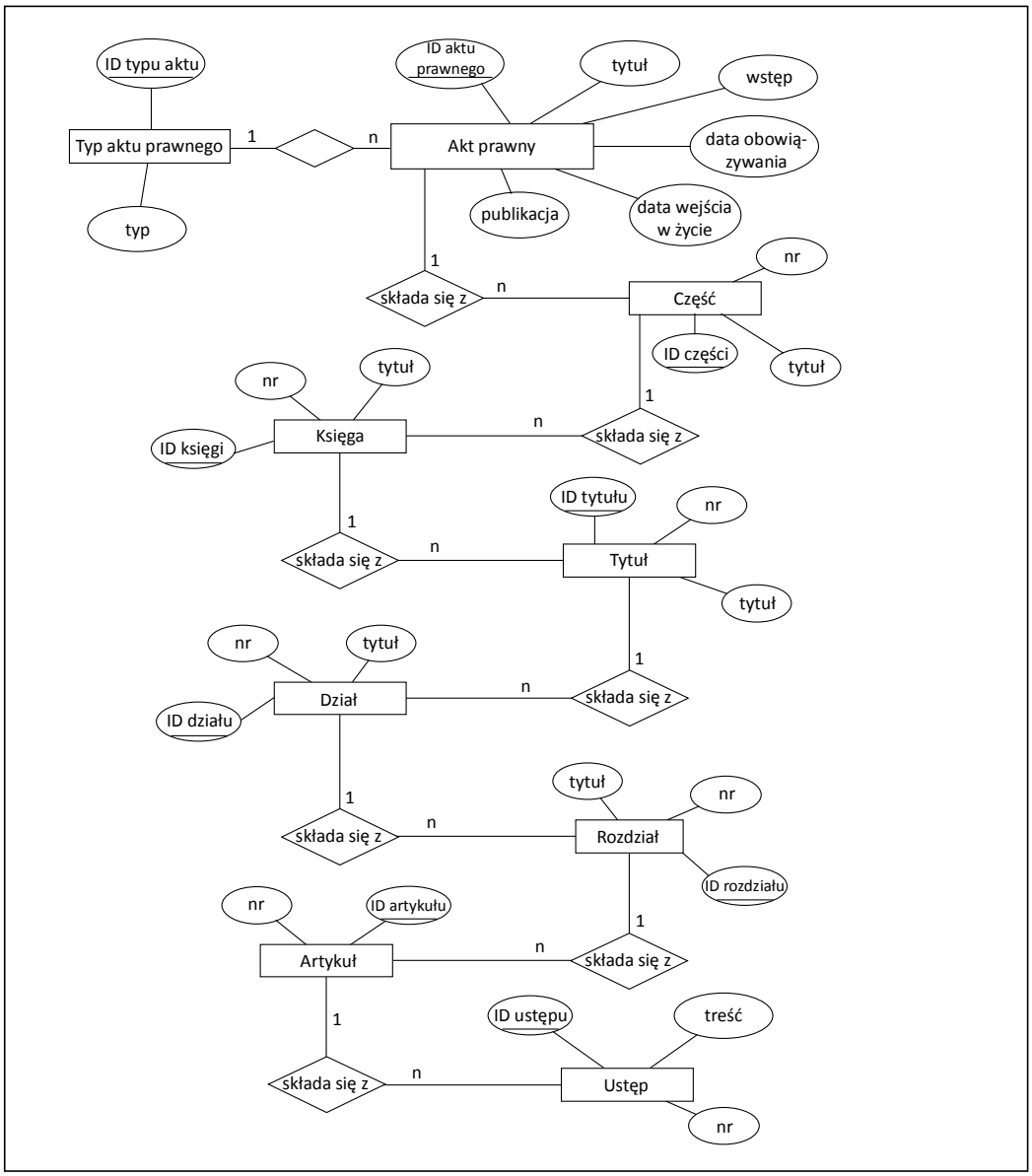
[419] Posługujemy się pojęciem aktów ujednoliconych w odróżnieniu od aktów jednolitych, które są tworzone przez organy władne do wydania samego aktu, na przykład teksty jednolite ustaw ogłasza Sejm.

[420] Uchwała nr 147 Rady Ministrów z dnia 5 listopada 1991 roku w sprawie zasad techniki prawodawczej (M.P. nr 44, poz. 310 z dnia 16 grudnia 1991 roku).

z wykorzystaniem metod ekstrakcji i integracji informacji sterowanych ich strukturą relacyjną (porównaj na stronie 495).



Rysunek 181. Rozkład liczby odwołań i referencji wychodzących z aktów wydanych w 1997 roku do aktów z poprzednich lat wynikający z dokumentów prawnych zgromadzonych w systemie *HyperThemis* [T. Tomaszewski, 1999]



Rysunek 182. Relacyjny model struktury danych dla hipertekstowego systemu informacji prawnej [P. Płuciennik, 1997]

Największego potencjału biznesowego dla filtrów jako elementu funkcjonalności bibliotek elektronicznych dopatrujemy się w rozpowszechnianiu *e-books* w takich bibliotekach jak już wspomniane *Project Gutenberg*, *Google Book Serach* czy *Live Search Books*. Ogromnych możliwości dopatrujemy się na obu końcach rynku księgarskiego: na masowym i wyspecjalizowanym. Dla rynku masowego wyobraźnię pobudzają miliony kupujących kolejne tomy przygód Harry Pottera. Dla rynków wyspecjalizowanych *e-books* mogą okazać się szansą ponownego urynkowienia obrotu książka-

mi, które sprzedawane obecnie w nakładzie kilkuset egzemplarzy, co zmusza do szukania innych niż wydawnicze źródeł finansowania. Barierą rozwoju tych zastosowań nie jest technologia, lecz brak modelu biznesowego, akceptowanego przez wszystkich uczestników rynku: autorów, wydawców, właścicieli praw autorskich, księgarzy oraz czytelników.

Filtry zasilające biblioteki elektroniczne

Filtry informacyjne mogą być wykorzystywane nie tylko jako narzędzia udostępniane użytkownikom bibliotek, ale również jako narzędzia zarządzania relacjami z klientami CRM (*Customer Relationship Management*). Mogą wspomagać zasilanie biblioteki w dokumenty, odpowiadające potrzebom informacyjnym użytkowników jako systemy zarządzania łańcuchem dostaw SCM (*Supply Chain Management*). Wówczas należy je traktować jako systemy rekomendacyjne w tradycyjnym środowisku internetowym, wspomagające racjonalne kupowanie (porównaj na stronie 396). Główne trudności związane z wykorzystaniem filtrów w tym zakresie związane są z budowaniem profili.

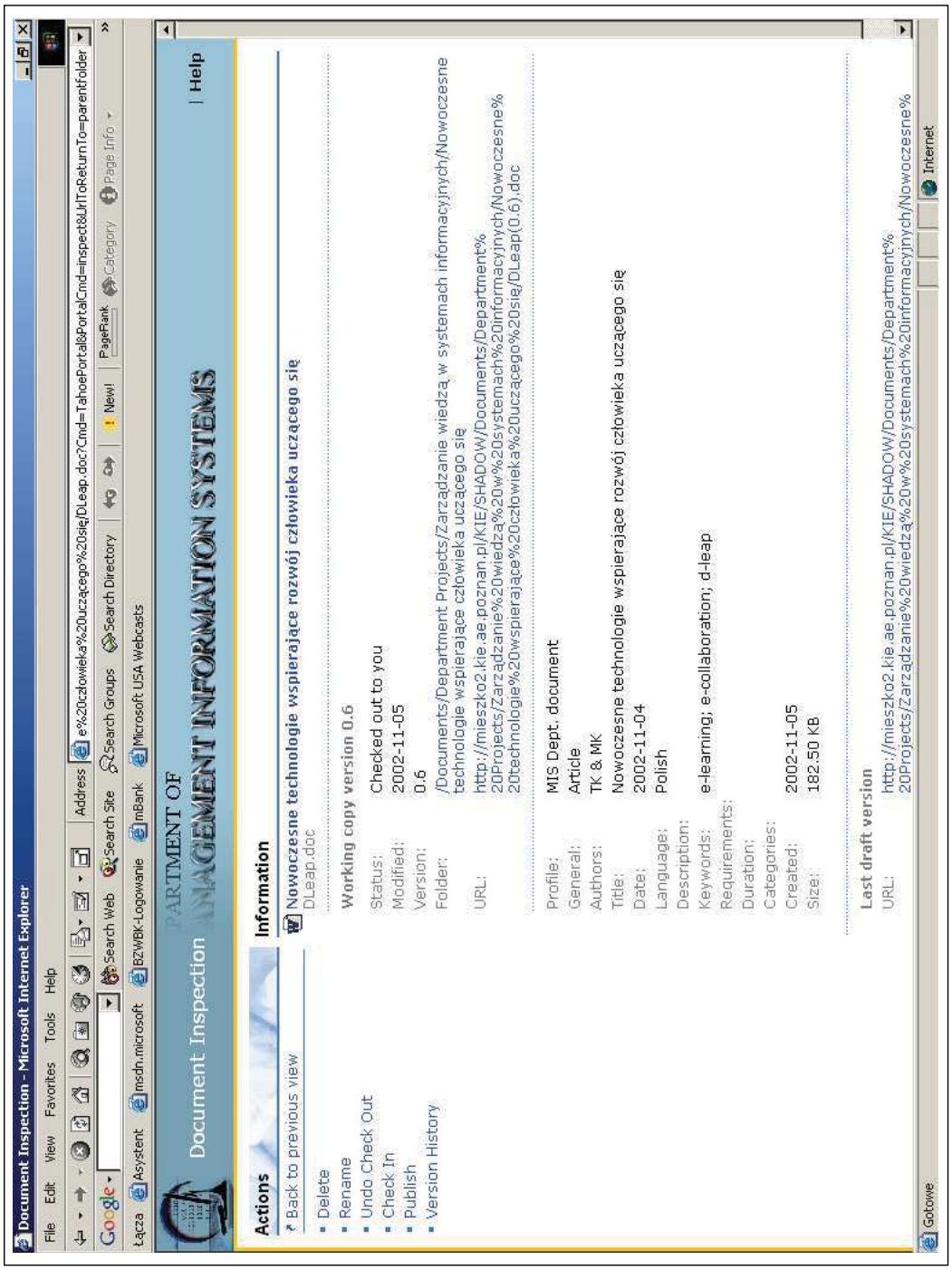
Biblioteki elektroniczne zarządzające wyfiltrowanymi dokumentami

Użytkownik intensywnie korzystający z filtru informacyjnego dostaje zapewne wiele relewantnych dla niego dokumentów. Masowość tej społeczności użytkowników może zapewnić upowszechnienie się elektronicznych książek (*e-books*). Część z nich użytkownik zapewne chce zachować dla późniejszego wykorzystania. Jeżeli jest ich odpowiednio dużo, biblioteka elektroniczna może być narzędziem do efektywnego zarządzania wyfiltrowanymi dokumentami. Oprócz opisanej wyżej funkcjonalności biblioteki użytkownik powinien mieć do dyspozycji narzędzia [G. Buchanan i inni, 2005]:

- wspomagające go przy wyborze dokumentów relewantnych tych, które będą włączane do biblioteki,
- prezentujące metadane dokumentów, ponieważ mogą pochodzić ze źródeł opatrujących dokumenty różnymi metadanymi,
- redagujące metadane dokumentów i metadane biblioteki w celu uzyskania zgodności pomiędzy metadanymi wszystkich dokumentów gromadzonych w bibliotece,
- indeksowania integrowanych dokumentów.

Realizując te postulaty oprócz typowej dla bibliotek klasyfikacji wprowadziliśmy do biblioteki elektronicznej *D-Leap*, opracowane przez nas mapy umiejętności (porównaj na stronie 86) jako dodatkowe kryterium strukturalizacji dokumentów. Nawigacja przez strukturę umiejętności (porównaj rysunek 184) prowadzi do dokumentów związanych z umiejętnościami,

a dotarcie do tych dokumentów może być wykorzystywane w *e-learning* [W. Abramowicz i inni, 2002f].



Rysunek 183. Prezentacja metadanych dokumentu na przykładzie biblioteki elektronicznej *D-Leap*