

Piotr GAWRYSIAK

[gawrysia@ii.pw.edu.pl](mailto:gawrysia@ii.pw.edu.pl)

Instytut Informatyki, Politechnika Warszawska

ul. Nowowiejska 15/19, 00-665 Warszawa

## **W STRONIE INTELIGENTNYCH SYSTEMÓW WYSZUKIWAWCZYCH W SIECI INTERNET**

W artykule przedstawiono niektóre z problemów związanych z projektowaniem i wykorzystywaniem systemów wyszukiwania informacji w sieci Internet. Zwrócono uwagę na heterogeniczność informacji zawartej w sieci i powodowany przez to wzrost znaczenia metod specyfikowania zapytań w systemach wyszukiwania danych. Rozważono także zasadność stosowania nieklasycznych podejść wyszukiwania informacji w Internecie, polegających na wykorzystaniu informacji nietekstowej i zastosowaniu metod automatycznej eksploracji danych (ang. *data mining*). Postuluje się także traktowanie wyszukiwania informacji w sieci jako procesu, w którym efekty mogą być osiągnięte dzięki współpracy użytkownika, inteligentnych narzędzi przeglądania stron WWW i autonomicznych systemów agencjonalnych.

### **1. WSTĘP**

Internet został stworzony jako wojskowy system komunikacyjny, którego użyteczność została także szybko doceniona i wykorzystana przez społeczność akademicką<sup>1</sup>. W obu tych przypadkach jego użytkownikami byli wysokiej klasy profesjonalistami, dla których efektywne korzystanie ze skomplikowanych rozwiązań sprzętowych i programowych nie stanowiło większego problemu. Najbardziej rozpowszechnionym w latach 80-tych i początku lat 90-tych na wyższych uczelniach systemem operacyjnym był Unix w swoich różnych odmianach. Był to (i pozostaje po dzień ten) bardzo skomplikowany i niezbyt przyjazny użytkownikowi system operacyjny, toteż osoby potrafiące wykorzystywać go w codziennej pracy nie miały zwykle problemów z używaniem narzędzi "internetowych", tym bardziej że wiele rozwiązań zastosowanych w Internecie jest ściśle związanych właśnie z systemem Unix.

Wraz z dynamicznym wzrostem popularności sieci Internet pojawiało się coraz więcej użytkowników nieprofesjonalnych. W chwili obecnej zawodowi programiści, czy też ogólniej osoby których wykształcenie związane jest choćby pośrednio z informatyką<sup>1</sup>, stanowi już mniejszość wśród osób korzystających z usług globalnej sieci. Jednocześnie jednak wiedza i umiejętności techniczne - w szczególności zaś dobra znajomość zasad działania narzędzi takich jak przeglądarki, systemy wyszukiwawcze, czy też serwery HTTP - stają się coraz bardziej przydatne, a w niektórych przypadkach wręcz niezbędne, do efektywnego odnajdywania potrzebnych użytkownikowi informacji w sieci. Główną tego przyczyną wydaje się być zwiększająca się w bardzo szybkim tempie liczba dostawców informacji, co w połączeniu z niezbyt wysoką jakością i aktualnością dostępnych katalogów i narzędzi wyszukiwawczych powoduje, iż proces odnajdywania danych w Internecie bywa czasami dość skomplikowany<sup>1</sup>.

W artykule tym rozważam niektóre aspekty projektowania współczesnych internetowych systemów wyszukiwawczych. Rozdział 2 zawiera rozważania dotyczące procesu wyszukiwania danych przez system wyszukiwawczy, w rozdziale 3 przedstawiono pojęcie *metafory sieci*, rozdział 4 poświęcony jest systemom specyfikacji zapytań, zaś w rozdziale 5 pokrótce omówiono problematykę obróbki wyników wyszukiwania i przedstawiono koncepcję wspomaganego przeglądania stron WWW. Artykuł kończy krótkie podsumowanie i bibliografia.

## 2. PROCES WYSZUKIWANIA INFORMACJI

Obecnie używane systemy wyszukiwania informacji w sieci WWW mają swoje korzenie w systemach przeszukiwania klasycznych baz danych. Dzięki temu są zwykle dość efektywne w indeksowaniu i grupowaniu informacji, nawet pełnotekstowej, jednakże nie są skuteczne w interpretacji potrzeb informacyjnych użytkownika. W wielu przypadkach interfejsy użytkownika w jakie wyposażono te systemy są bardzo niskiej jakości, zaś możliwości specyfikacji kryteriów wyszukiwania jakie są przez nie udostępniane wydają się być bardziej odpowiednie dla prostych baz pełnotekstowych, niż dla bogatego środowiska hipermedialnego w którym działają. Praktycznie wszystkie nowoczesne internetowe systemy wyszukiwawcze pozwalają jedynie na przeszukiwanie tylko tekstu dokumentów WWW, nie pozwalając nawet na określenie typu informacji jakiej potrzebuje użytkownik.

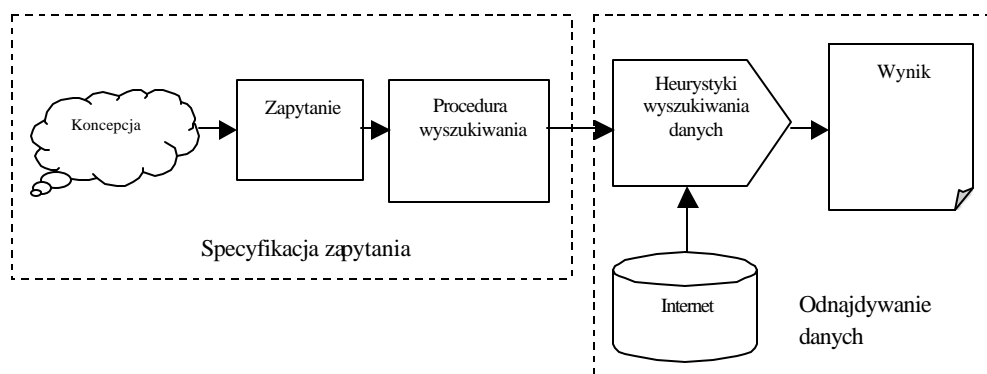
Z drugiej strony nowo powstałe narzędzia, takie jak języki zapytań W3QS[12] i WebML[14] pozwalają na bardzo dobrą kontrolę procesu wyszukiwania informacji i dokładne określenie jego celu. Niestety są to narzędzia stworzone z myślą o użytkownikach posiadających dobrą znajomość systemów baz danych (język SQL), nie są intuicyjne i z tego powodu nie mogą zostać bezpośrednio użyte w interfejsie użytkownika ogólnodostępnego systemu wyszukiwania.

W procesie wyszukiwania informacji, zarówno w klasycznych systemach baz danych, jak i w systemach internetowych, możemy wyróżnić dwie podstawowe fazy : fazę

---

<sup>1</sup> Patrz [8].

specyfikacji zapytania (ang. *query specification phase*) i fazę odnajdywania informacji (ang. *retrieval phase*).



Rys 1. Etapy procesu wyszukiwania informacji

We współczesnych systemach faza odnajdywania informacji przebiega zwykle bardzo efektywnie [2], jedynym wyjściem bywa tutaj prezentacja wyników wyszukiwania. W idealnym systemie wyszukiwania, wygenerowany wynik zawiera dokładnie tę informację, która została opisana w koncepcji użytkownika. Jednakże w rzeczywistości istniejących systemach współczynnik relewancji wyszukiwania rzadko osiąga poziom stu procent. Wynikiem tego jest "zaśmiecenie" wyniku wyszukiwania informacjami nierelevantnymi, nieistotnymi z punktu widzenia użytkownika. Z tego też powodu dalsza obróbka surowych wyników wyszukiwania może często poprawiać użyteczność systemu.

Faza specyfikacji zapytania uważana jest zwykle za najmniej istotny i tym samym skomplikowany element systemu wyszukiwawczego. Warto jednak zauważyć, iż w tej ważnej fazie podejmowana jest decyzja "czego szukać". W większości przypadków wszystkie części systemu wyszukiwawczego, poczynając od parsera analizującego zapytanie, aż po generator wyniku, są niewidoczne dla użytkownika. Użytkownik systemu dostarcza jedynie zapytanie (np. wpisując słowa kluczowe w pole formularza na stronie WWW), po czym otrzymuje wynik, i jeśli jego zapytanie zostało nieprawidłowo zinterpretowane (czego efektem jest błędna procedura wyszukiwania<sup>2</sup>) to wynik całego procesu będzie nie satysfakcjonujący, niezależnie od tego jak bardzo efektywne i skomplikowane będą algorytmy fazy odnajdywania danych. Jest to szczególnie widoczne w środowisku Internetu, gdzie system specyfikacji zapytań stanowi jedyny interfejs pomiędzy systemem wyszukiwawczym, a użytkownikami, którzy na dodatek zwykle nie posiadają doświadczenia w pracy z narzędziami odnajdywania informacji w bazach danych. Większość z nich traktuje pole przeznaczone na specyfikację zapytania jako miejsce do wpisania ogólnej koncepcji tego, co chcieliby w sieci odnaleźć. Jako że ludzie stosują różne wzorce postrzegania otaczającego świata i myśli o Internecie używając wielu różnych metafor, toteż taka koncepcja może mieć wiele form. Dobrymi przykładami będą:

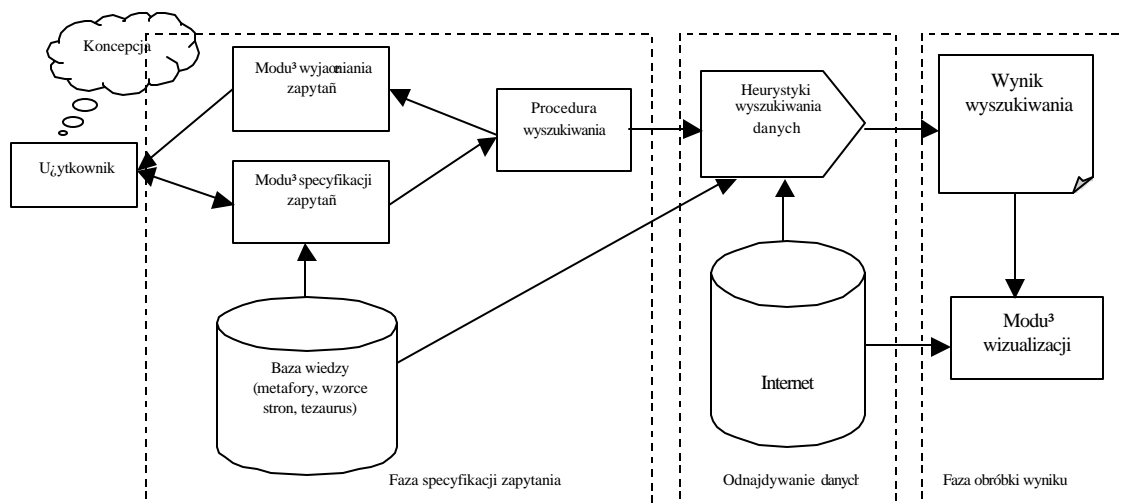
- zadanie systemowi pytania, traktując Internet jako uniwersalną bazę wiedzy

<sup>2</sup> Terminem *procedura wyszukiwania* określam zbiór instrukcji dla serwera wyszukiwawczego zapisanych w języku formalnym, takim jak np. W3QL.

- opisanie za pomocą<sup>1</sup> słów kluczowych charakterystyk poszukiwanej strony WWW
- podanie słów, jakie powinny pojawić się w treści poszukiwanej strony

Wszystkie powyższe zapytania można by najprawdopodobniej przetłumaczyć na dobre procedury wyszukiwania, jednak nie jest to możliwe bez poinformowania systemu o tym która z metafor została użyta. Na to z kolei nie pozwalają żadne istniejące internetowe narzędzia wyszukiwawcze. Uważam zatem, iż faza *specyfikacji zapytania* zasługuje na więcej uwagi niż poświęcano jej dotychczas.

Poniżej prezentuję model hipotetycznego systemu wyszukiwania informacji, który będzie przedmiotem rozważań w kolejnych rozdziałach.



Rys. 2. „Inteligentny” system wyszukiwania informacji

### 3. MODELE SIECI WWW

Większość ludzi postrzega Internet nie jako całkowicie nowe narzędzie, ale raczej jako elektroniczną wersję jednego z do tej pory istniejących systemów archiwizacji i dystrybucji informacji. Nie ma w tym nic dziwnego - w procesie edukacji uczeni byliśmy jak do tej pory jedynie posługiwania się klasycznymi systemami, wykształcenie to z kolei na tyle silnie wpływa na nasze zachowanie, że instynktownie traktujemy je jako *naturalne*<sup>3</sup>. Kiedy korzystamy z sieci, traktujemy ją zwykle jak księgarnię, lub bazę danych, czy też gazetę, wreszcie używamy jednej z wielu innych metafor które są dla nas znajome. Wybór ten staje się szczególnie istotny w momencie w którym poszukujemy informacji w sieci, ponieważ ogranicza rodzaj informacji o wyszukaniu którego możemy w ogóle pomyśleć. W efekcie wybrana przez nas metafora, innymi słowami sposób postrzegania Internetu, określa stosowane przez nas podejście do konstruowania zapytania.

<sup>3</sup> Dobrym przykładem będzie tu traktowanie pisma odręcznego jako *naturalnego* systemu wprowadzania tekstu, podczas gdy jest on znacznie bardziej skomplikowany i trudniejszy do nauczenia czynności niż np. korzystanie z klawiatury.

Ta sytuacja wcale nie musi być niekorzystna. Internet jest w istocie tak heterogenicznym Źródłem danych, że postrzeganie go z wielu różnych perspektyw może pomóc w zrozumieniu tego jakiego rodzaju informacje może on zawierać. Musimy przy tym jednak pamiętać o ograniczeniach konkretnej metafory i dobierać ją w zależności od wyników jakie pragniemy osiągnąć.

Poniżej podaję przykłady najpopularniejszych metafor Internetu, wraz z typowymi zapytaniami w języku naturalnym jakie się z nimi związane:

- **gazeta (newspaper metaphor)**

Przykładowe zapytanie: "Jakie nowe informacje związane z moimi zainteresowaniami pojawiły się wczoraj?"

- **artykuły naukowe (proceedings metaphor)**

Przykładowe zapytanie: "Czy istnieją inne dokumenty weryfikujące informacje zawarte w przeglądany dokument?"

- **baza danych (database metaphor)**

Przykładowe zapytanie: "Odszukaj dokumenty zawierające słowa X i Y, ale nie Z"

- **serwer plików (file repository metaphor)**

Przykładowe zapytanie: "Odszukaj najnowszą wersję oprogramowania X"

- **encyklopedia (encyclopaedia metaphor)**

Przykładowe zapytanie: "Jaka jest definicja terminu X?"

- **baza wiedzy (knowledge base metaphor)**

Przykładowe zapytanie: "Dlaczego ryby nie mówią?"

Nie jest możliwe bezpośrednio użycie powyższych zapytań w istniejących systemach wyszukiwawczych, jako że system który byłby w stanie w pełni zrozumieć semantykę nawet tak prostych zdań nie został jak do tej pory stworzony i najprawdopodobniej nie powstanie w najbliższej przyszłości. Jest jednak możliwe przetłumaczenie tych zapytań do bardziej formalnej postaci, jeśli poinformujemy system wyszukiwawczy jaka konkretna metafora została użyta w każdym przypadku. Warto zauważyć że dostarczyłoby to także systemowi wyszukiwania danych wielu istotnych informacji, związanych z procesem wyszukiwania. Dla przykładu jeśli użyjemy zapytania związanego z metaforą *serwera plików*, to możemy proces wyszukiwawczy ograniczyć jedynie do kilku, wyspecjalizowanych serwisów internetowych (takich jak *download.com*, *shareware.com* itp.), przynajmniej na początkowym etapie. Z kolei osoba postrzegająca sieć WWW przez metaforę *artykułów naukowych* najprawdopodobniej przypisuje dużą wagę do występowania hiperpołączeń pomiędzy stronami WWW, a zatem system wyszukiwawczy może bardziej intensywnie wykorzystywać algorytmy analizy grafowej w poszukiwaniu danego zbioru dokumentów.

Powyższa optymalizacja procesu wyszukiwania, mimo iż możliwa do "ręcznego" przeprowadzenia, nie powinna raczej być dokonywana przez użytkownika systemu wyszukiwania danych. Zakłada ona bowiem pewną znajomość Internetu *per se* - użytkownik musiałby na przykład wiedzieć jakie są najlepsze istniejące repozytoria plików

w Internecie - gdy tymczasem idealny system wyszukiwawczy powinien być równie użyteczny dla osób nie posiadających prawie żadnej wiedzy o strukturze sieci WWW. Istotnym czynnikiem jest tu także szybka zmienność Internetu, jako że zasoby uznawane dzisiaj dobre wcale nie muszą takimi pozostać w przyszłości.

#### 4. SYSTEMY SPECYFIKACJI ZAPYTAŃ

W poprzednim rozdziale argumentowałem, iż dobrze zaprojektowany system wyszukiwania informacji powinien pozwalać na podanie pewnych *metainformacji* o zapytaniu, takich jak na przykład zastosowana metafora sieci WWW. W rzeczywistości system specyfikacji metafor nie byłby zapewne możliwy do zbudowania ze względów praktycznych (niezależnie od tego jak bardzo rozbudowana byłaby baza metafor dostępna dla algorytmów wyszukiwawczych systemu, i tak nie pokryłaby wszystkich, bardzo przeciętnie zróżnicowanych, potrzeb użytkowników). Wydaje się jednak, iż nawet udostępnienie użytkownikom prostszego zestawu narzędzi może być użyteczne.

Postuluję, iż większość procesów wyszukiwania informacji we współczesnej sieci Internet może być zaliczona do jednej z dwóch grup. Pierwsza z nich, *wyszukiwanie stron* (ang. *page searches*) zawiera te procesy, których celem jest odnalezienie pojedynczej, dobrze zdefiniowanej strony WWW (lub też, nieco generalizując, pojedynczego zasobu sieciowego). Przykładami takich procesów będą: poszukiwanie strony domowej kolegi, poszukiwanie oficjalnej strony konkretnej firmy lub też odszukanie publikacji naukowej przy użyciu jej tytułu i nazwisk autorów.

Druga grupa, określana przeze mnie terminem *wyszukiwanie informacji* (ang. *information searches*) obejmuje zapytania nie odnoszące się do konkretnych dokumentów, czy też nawet do struktury sieci WWW. Celem tych zapytań jest odnalezienie informacji, bez względu na jej źródło czy też formę jej prezentacji. Dla przykładu rozważmy użycie zasobów WWW do odnalezienia informacji o planetach systemu słonecznego. Możliwe, wartościowe rezultaty takiego wyszukiwania mogłyby obejmować tak zróżnicowane obiekty jak dokumenty tekstowe zawierające informacje statystyczne o charakterystykach planet, trójwymiarowe modele układu słonecznego stworzone przy wykorzystaniu języka VRML, plakat ze strony NASA ze zdjęciami planet i tak dalej.

Główną różnicą pomiędzy dwoma powyższymi typami procesów wyszukiwania stanowi to, iż w przypadku *wyszukiwania informacji* sieć WWW jest traktowana jako „czarna skrzynka”, czy też nawet autonomiczny system ekspertowy, natomiast w *wyszukiwaniu stron* struktura sieci jest istotna dla użytkownika.

Ponieważ, jak widać, istnieją różne podejścia do problemu ekstrakcji informacji z sieci WWW, toteż powinny istnieć także różne interfejsy użytkownika, czy też różne zestawy narzędzi, które wspomagałyby użytkowników w procesie konstrukcji zapytań i w ocenianiu wyników wyszukiwania. Pierwszy zestaw narzędzi, przeznaczony dla celów wyszukiwania stron, powinien pozwalać na określanie charakterystyk dokumentów internetowych z możliwie największą precyzją. Drugi zestaw, wykorzystywany przy wyszukiwaniu

informacji, powinien wspomagać użytkownika w analizie możliwie bogatego zbioru zróżnicowanych zasobów internetowych. W pierwszym przypadku najistotniejszy staje się system specyfikacji zapytań, w idealnym wypadku wynik wyszukiwania zawiera tu będzie wyłącznie te strony, które zostały opisane przez użytkownika, w związku z czym rola modułu prezentacji i obróbki wyniku wyszukiwania będzie tu niewielka. Dokładnie odwrotna sytuacja występuje natomiast przy wyszukiwaniu informacji.

Współczesne systemy wyszukiwawcze dają dość dobre możliwości konstruowania zapytań do warstwy tekstowej dokumentów, jednak informacja czysto tekstowa nie jest jedyną, jaka decyduje o postrzeganiu dokumentów multimedialnych jakimi jest większość nawet prostych stron WWW. Równie istotny jest *rodzaj strony* - możemy pamiętać że widzieliśmy *stronę domową* naszego znajomego, nie będąc jednocześnie w stanie sobie przypomnieć jej tekstowej zawartości. Podobnie możemy poszukiwać strony "New York Times" znając jedynie nazwę tej gazety i wiedząc że poszukujemy strony z wiadomościami prasowymi. Jak widzimy w wielu przypadkach możliwość specyfikowania typu strony mogłaby pozytywnie wpłynąć na efektywność procesu wyszukiwania informacji.

Niezawodny system klasyfikacji stron WWW nie został jeszcze opracowany, pomimo kilku interesujących osiągnięć w tej dziedzinie [3], [4]. Wydaje się, iż jednym z największych problemów na jakie natrafia się przy próbach tworzenia takich systemów jest problem rozumienia języka naturalnego (ang. *natural language understanding*). Niektórzy badacze [11] postulują nawet, że uniwersalny system klasyfikacyjny dla dokumentów WWW jest niemożliwy do zbudowania z powodu zbyt wielkiej różnorodności potrzeb i preferencji użytkowników. Rezultaty w grupowaniu stron zaprezentowane przez Pitkova, Pirolego i Rao w [13], oraz mój eksperyment [7] pokazują jednak, iż proces klasyfikacji może być skuteczny nawet jeśli nie uwzględni się znaczenia tekstu zawartego w dokumentach.

W eksperymencie tym grupa około 20 osób została poproszona o ręczną klasyfikację wybranych uprzednio stron WWW według systemu klasyfikacji zaproponowanego przez Kazienkę [10]. Strony WWW zostały zmodyfikowane w ten sposób, że:

- tekst został zamieniony przez losowe ciągi znaków, przy czym zachowano rozróżnienie pomiędzy znakami alfabetu łacińskiego i cyframi
  - została zachowana informacja formatująca (wielkość i kolor czcionki, rozmieszczenie tekstu itp.)
  - wszystkie ilustracje zostały usunięte i zastąpione pustymi prostokątami o odpowiednich wymiarach
  - hiperpołączenia zostały zmodyfikowane tak, aby nie można było odczytać adresu URL, typ hiperpołączenia (ftp, email, http itp.) oraz jego głębokość (liczba części /.../)
- pozostały jednak zachowane

Pomimo braku dostępu do tekstowej treści dokumentów, użytkownicy byli w stanie zaklasyfikować niektóre grupy dokumentów do odpowiednich klas z dość wysoką precyzją. Najlepsze rezultaty osiągnięto dla poniższych klas najwyższego poziomu:

**" strony domowe" ~ 70% dokładności**

"strony organizacyjne" ~ 70% dok³adnoœci

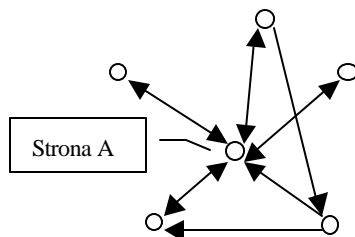
zaœnajgorsze rezultaty dla klasy

"teksty" ~ 40% dok³adnoœci.

Wor³d klas drugiego poziomu najwy¿szym wsp³oczynnikiem dok³adnoœci klasyfikacji charakteryzowa³a siê klasa "wizytówki" - 90% dok³adnoœci klasyfikacji. Dok³adny opis zastosowanego drzewa klasyfikacji zawiera praca [10].

Wydaje siê zatem, i¿ mo¿liwe jest skonstruowanie systemu klasyfikacji, który korzystaæ bêdzie jedynie z informacji nietekstowej o stronach WWW (takiej jak wielkoœæ strony, liczba ilustracji, gêstoœæ hiperpo³czeñ, liczba u¿ytych czcionek itp.). Nawet je¿li skonstruowany system pozwala³by jedynie na bardzo zgrubn¹ kategoryzacjê, to i tak mo¿liwoœæ przyporz¹dkowania danej strony do jednej z ogólnych klas, np. "stron informacyjnych" (ang. *content delivering pages*) i "stron nawigacyjnych" (ang. *navigation pages*) mog³aby mieæ du¿e znaczenie dla wielu u¿ytkowników.

Pojedyncza strona WWW mo¿e byæ tak¿e scharakteryzowana poprzez funkcjê jak¹ pe³ni w wiêkszej grupie stron. Niektóre strony WWW s¹ indeksami do zbiorów stron, inne spe³niaj¹ rolê list adresowych, map serwisów itp. Informacja formatuj¹ca - to jest wygl¹d strony - w wielu przypadkach mo¿e byæ zupe³nie wystarczaj¹ca do identyfikacji takich stron. Mo¿emy tu jednak pos³u¿yæ siê tak¿e grafem hiperpo³czeñ, by próbowaæ okreœliæ funkcjê danej strony. Dla przyk³adu w strukturze przedstawionej na **Rysunku 3**, strona A jest najprawdopodobniej centralnym indeksem grupy typu Webring.



Rys.3. Grupa stron typu Webring

Oczywiœcie trudno zdefiniowaæ rêcznie wystarczaj¹co wielk¹ liczbê podobnych struktur sieciowych, tak aby zaspokoiaæ potrzeby u¿ytkowników. Jednym z rozwi¹zañ tego problemu mo¿e byæ umo¿liwienie zdefiniowania poszukiwanej struktury bezpoœrednio przez u¿ytkownika, w procesie specyfikacji zapytania. W istocie niektóre z jêzyków przeszukiwania sieci WWW wywodz¹ce siê z SQL posiadaj¹ takie mo¿liwoœci. Innym rozwi¹zaniem, zapewne mniej efektywnym, lecz przyjañniejszym dla u¿ytkownika, by³oby adaptowanie technik automatycznego odkrywania wiedzy w bazach danych (ang. *data mining*), takich jak na przyk³ad z³o¿one regu³y asocjacyjne [1], do celów identyfikacji najczêœciej wystêpuj¹cych, interesuj¹cych struktur, w Internecie.

Je¿li zaœcê bêdziemy w stanie okreœliæ choæby zgrubnie, typ strony WWW i funkcjê spe³nian¹ przez ni¹, to mo¿emy wzbogaciæ system budowy zapytania o mo¿liwoœci specyfikacji tych¿e cech.

Inne narzędzia wspomagające proces budowy zapytania mogą także pozytywnie wpływać na efektywność procesu wyszukiwania informacji. Prace prowadzone nad systemami peñnotekstowymi [6] pokazują, iż nawet tak proste narzędzie jak tezaursus dostępny podczas projektowania zapytania, może znacząco przyczynić się do wzrostu wydajności całego systemu.

Godne rozważenia może być także zwiększenie kontroli użytkowników nad funkcją oceniania (ang. *ranking expression*). Większość aktualnie dostępnych internetowych systemów wyszukiwawczych pozwala jedynie na kontrolę nad samym zapytaniem, po czym ocenia zgodność odnalezionych dokumentów z zapytaniem według własnej funkcji oceniania, która niekoniecznie musi odpowiadać potrzebom użytkownika.

Na koniec warto także rozważyć wprowadzenie modułu wyjaśniającego zapytanie (ang. *query explanation module*) do systemu wyszukiwawczego. Moduł taki tłumaczyłby wprowadzone zapytanie na zdanie w języku naturalnym, lub też nawet konstruowałby przykładowe strony WWW dokładnie odpowiadające zapytaniu, aby pokazać jakiego rodzaju dokumentów poszukiwałby system.

## 5. OBRÓBKA WYNIKÓW I WSPOMAGANE PRZEGLĄDANIE STRON

Wyszukiwanie informacji jest zadaniem znacznie bardziej skomplikowanym niż wyszukiwanie stron. Wydaje się, iż efektywne wydobywanie informacji z zasobów sieci WWW powinno być traktowane bardziej jako proces (na który składa się przeglądanie stron, szukanie danych, porównywanie stron itp.) niż tylko pojedyncza sesja wyszukiwania danych. Taki proces mógłby także obejmować inteligentną analizę wyników wygenerowanych przez systemy wyszukiwawcze (razem z prawdopodobnym ulepszeniem zapytania i kolejnymi cyklami wyszukiwania) oraz wspomagane przeglądanie stron WWW (ang. *assisted browsing*).

Techniki odkrywania wiedzy w bazach danych mogą być szczególnie użyteczne w analizie wyników pracy systemu wyszukiwawczego. Jak pokazują w artykule przeglądowym [9] szczególne znaczenie mogą mieć tu techniki analizy topologii hiperpołączeń, które (do pewnego stopnia) pozwalają na sortowanie stron w zależności od ich jakości postrzeganej przez społeczność internetową. Z kolei techniki analizy tekstu (ang. *text mining*) pozwalają na sporządzenie wizualizacji i automatycznych podsumowań, pozwalając na szybkie ocenę treści odszukanych stron.

Proces specyfikacji zapytania jest szczególnie trudny w przypadku wyszukiwania informacji, dlatego też zasadnym wydaje się poszukiwanie takich metod ekstrakcji informacji z Internetu, w których proces ten nie jest w ogóle potrzebny. Wspomagane przeglądanie stron WWW może być jedną z takich ważnych metod.

Wspomagane przeglądanie stron WWW jest procesem, w którym wybór następnej strony WWW jak ogólnie będzie użytkownik korzystający z przeglądarki nie jest dokonywany wyłącznie na podstawie hiperpołączeń dostępnych na danej stronie, ale także na podstawie dodatkowej informacji o stronach sąsiednich w grafie hiperpołączeń i o

topologii tego grafu. Informacja ta jest zbierana i analizowana przez przeglądarkę bez udziału użytkownika. Dobrym przykładem takiego systemu jest zbiór narzędzi oferowanych przez firmę Alexa Inc., które wykorzystują informacje statystyczne o popularności stron WWW, by określać które strony mogą być potencjalnie tematycznie związane z aktualnie oglądanymi stronami. Co ciekawe ostatnie badania w tej dziedzinie [5] wykazują, iż możliwe jest odnajdywanie takich semantycznie powiązanych stron przy użyciu wyłącznie grafu hiperpołączeń.

Zachowanie użytkownika może być także analizowane, w celu automatycznego zbudowania jego listy preferencji czy też profilu, który opisuje jego potrzeby informacyjne. Taki profil stanowi może dobry punkt wyjściowy dla autonomicznych systemów agencjonalnych przeglądających sieć WWW w poszukiwaniu interesujących stron. Z tego podejścia wywodzi się także koncepcja *personalizowanych portali sieciowych*. Obecnie większość użytkowników rozpoczyna przeglądanie zasobów sieci Internet od jakiegoś portalu, takiego jak np. Yahoo, który może być ze sobą uniwersalny system wyszukiwania informacji i tematyczny indeks stron WWW. Te portale, pomimo wysiłków zmierzających do tego by były w pewnym choć stopniu modyfikowalne w zależności od preferencji użytkownika (tzw. profilowanie portali), to i tak są narzędziami bardzo ogólnymi. Mikroportale, tworzone specjalnie dla pojedynczego użytkownika (lub pojedynczego tematu) powinny być znacznie bardziej efektywne, do ich tworzenia może być zastosowane również metody wywodzące się z data mining [9].

Na koniec chciałbym zwrócić uwagę na fakt, iż różnorodność informacji dostępnej w sieci Internet powoduje także różnorodność narzędzi służących do jej wyszukiwania. Obecnie doświadczeni użytkownicy Internetu korzystają przynajmniej z kilku różnych systemów wyszukiwawczych, ich wybór uzależniają od tego gdzie w sieci spodziewają się znaleźć informacje (np. *Deja.com* dla grup dyskusyjnych Usenet'u, *Altavista* dla stron WWW, *WhoWhere* dla adresów poczty elektronicznej itd.). Wydaje się zatem iż system ekspertowy, potrafiący automatycznie wybrać najlepsze indeksy i narzędzia dla konkretnego zapytania, byłby także bardzo użyteczny, szczególnie dla osób, które Internetem posługują się okazjonalnie.

## 6. PODSUMOWANIE

Efektywne odnajdywanie danych w sieci Internet jest złożonym i trudnym problemem, który jednocześnie wydaje się być kluczowym dla użyteczności tego globalnego źródła informacji. Dobrym ilustracją jego złożoności są problemy związane z efektywnym konstruowaniem i interpretacją zapytań w systemach wyszukiwawczych, które starałem się przedstawić w tym artykule. Zaproponowane rozwiązania opierają się na wykorzystaniu dodatkowej, nietekstowej informacji zawartej w strukturze sieci WWW, postuluję także zwrócenie większej uwagi na kwestie projektowania interfejsów użytkownika systemów wyszukiwawczych.

## LITERATURA

- [1]Borges J., Levene M., "Mining Association Rules in Hypertext Databases", AAAI-98 conf. proc., 1998
- [2]Brin S., Page L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine", 1998
- [3]Chakrabarti S., Berg M., Dom B., "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery", IBM Almaden Research Centre, 1999
- [4]Chakrabarti S., Dom B., Indyk P. "Enhanced hypertext categorization using hyperlinks", SIGMOD'98 conference proc., 1998
- [5]Dean J., Henzinger M., "Finding related pages in the World Wide Web", WWW8 conf. proc., 1999
- [6]Frelek M., Gawrysiak P., Rybiński H., "A method of retrieval in flexion-based language text databases", IIS'99 conference proceedings, 1999
- [7]Gawrysiak P., "Web page classification", (in preparation), Warsaw University of Technology, 2000
- [8]Gawrysiak P., "Using Data Mining methodology for text retrieval", DIBS'99 conference proceedings, 1999
- [9]Gawrysiak P., Okoniewski M., "Knowledge discovery in the Internet", submitted to Archiwum Informatyki journal, 1999
- [10]Kazienko P., "Rodzaje stron i odsy³aczy w systemie WWW", Wroclaw University of Technology, 1999
- [11]Macskassy S., Banerjee A., Davidson B., Hirsh H., "Human performance on clustering web pages", KDD'98 conference proc., 1998
- [12]Mihaila G., "WebSQL – An SQL-like Query Language for the World Wide Web", Department of Computer Science, University of Toronto, 1996.
- [13]Pirolli P., Pitkow J., Rao R., "Silk from a Sow's Ear: Extracting Usable Structures from Web", Proc. of the Conference on Human Factors in Computing Systems: Common Ground, pages 118-125, New York, 1996
- [14]Zaiane O., Han J., "WebML: Querying the World Wide Web for resources and knowledge", 1999