STUDIA INFORMATICA 2002
Volume 23 Number 2B (49)

Henryk KRAWCZYK, Sebastian KRUK
Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki

INTELIGENTNE WYSZUKIWANIE INFORMACJI W INTERNECIE

Streszczenie. W pracy przedstawiono krótką charakterystykę architektury i technologii wykorzystywanych przy wyszukiwaniu informacji w Internecie. Zaproponowano koncepcje inteligentnej wyszukiwarki spełniającej różnorodne wymagania użytkownika. Przedyskutowano jej wykorzystanie w telemedycynie, w celu wsparcia inteligentnych funkcji rekomendacji.

INTELLIGENT INFORMATION SEARCH IN INTERNET

Summary. In the paper short technology and architecture characteristics used by search engines are presented. A new concept of intelligent search engine for various users requirements is proposed. Its utilisation in telemedicine to support intelligent recommendation functions is discussed.

1. Wprowadzenie do poszukiwań w Internecie

Współczesna postać Internetu powoduje, że coraz bardziej powszechna staje się opinia, iż "czego nie ma w Internecie, tego nie ma w ogóle". Wraz ze wzrostem Internetu efektywność odnalezienia poszukiwanej informacji maleje. Dzięki swojej otwartej architekturze Internet stał się medium, które umożliwia praktycznie każdemu wyrażanie swoich opinii, publikację swoich osiągnięć, prezentowanie swojej firmy, itp. Niestety, takie liberalne podejście do publikacji spowodowało, że ilość informacji umieszczanej w Internecie rośnie gwałtownie z dnia na dzień. Powoduje to, że odnalezienie danej informacji staje się praktycznie niemożliwe bez pomocy serwisów wyszukiwawczych. Coraz wymyślniejsze technologie stosowane w powstających wyszukiwarkach internetowych nie są w stanie dotrzymać kroku, tak iż bardzo często nasze poszukiwania kończą się na przeglądaniu ogromnej liczby stron w celu znalezienia żądanej informacji.

1.1. Klasyfikacja serwisów wyszukiwawczych

Kiedy stajemy przed koniecznością znalezienia pewnej informacji w Internecie, do naszej dyspozycji mamy całą gamę najróżniejszych serwisów wyszukiwawczych, grup dyskusyjnych, katalogów. O tym, jakiej metody w danym momencie użyjemy, decyduje często przypadek. Z czasem wyrabiamy sobie pewne schematy tak, iż poszukując informacji danego typu korzystamy z jednej wyszukiwarki, a innego typu z np.: grup dyskusyjnych.

Ogólnie nasze poszukiwania w Internecie możemy oprzeć na narzędziach należących do następujących kategorii:

- **Katalog.** Przechowuje informacje o stronach w postaci hierarchicznego drzewa tematycznego. Aby odszukać daną informację musimy zdawać sobie sprawę, do jakiej kategorii ona należy. Zaletą takiego rozwiązania jest często fakt tworzenia takiego serwisu „ręcznie" – dzięki czemu dana strona jest umieszczana we właściwym miejscu katalogu zgodnie z intencją jej twórcy (często ten proces jest automatyzowany dzięki wykorzystaniu informacji umieszczonej w tagu <META>). Niestety, katalogi zawierają informacje o dużo mniejszej liczbie stron niż wyszukiwarki.
- **Wyszukiwarka.** Dzięki zastosowaniu wydajnej bazy danych i „armii" agentów przeglądających sieć, dobrze zaprojektowana wyszukiwarka jest dość dobrze skalowalna – i może nadążać za rozrastającym się Internetem. Niestety, wprowadzanie stron do bazy danych wyszukiwarki, z uwagi na swoje zautomatyzowanie, powoduje, że dana strona jest najczęściej katalogowana z bardzo małą precyzją, a co za tym idzie – na nasze pytanie wysłane do wyszukiwarki otrzymujemy często bardzo dużo wyników, których weryfikacja może zająć bardzo dużo czasu, a z których większość i tak nie odpowiada naszym kryteriom. Dobrym rozwiązaniem są tzw. *wyszukiwarki tematyczne*, które indeksują strony należące jedynie do danej dziedziny wiedzy. Niestety, jeżeli nie wiemy, jak sklasyfikować nasze zapytanie, konieczne staje się sprawdzenie wyników co najmniej z kilku wyszukiwarek. Z pomocą przyjść mogą wówczas metawyszukiwarki.
- **Meta-Wyszukiwarki.** Najczęściej jedyną informacją, jaką przechowują, są informacje, które wyszukiwarki należy obdarzyć większym zaufaniem w przypadku danego zapytania i jak ewentualnie przetransformować zapytanie, tak aby było zrozumiane przez inną wyszukiwarkę. Wyniki uzyskane od innych wyszukiwarek są następnie przetwarzane i prezentowane użytkownikowi. Dzięki temu nie ma konieczności przeglądania kolejno wyników kilku wyszukiwarek, lecz wystarczy zadać zapytanie tylko jednej metawyszukiwarce.

1.2. W kierunku inteligentnego poszukiwania

Analiza powyższego podziału serwisów wyszukiwawczych prowadzi do wniosku, że wzorcowy serwisem byłby taki, który łączyłby najlepsze cechy wszystkich trzech wymienionych:

- strukturę drzewiastą i logiczny sposób odnajdywania informacji, jaki udostępnia katalog;
- liczba zaindeksowanych stron zbliżoną do liczby stron udostępnianej przez wyszukiwarki;
- zawężenie liczby proponowanych dokumentów do niezbędnego minimum – co jest niewątpliwie cechą metawyszukiwarek

Powyższe trzy kryteria można prawie jednoznacznie przełożyć na (odpowiednio) trzy parametry określające *efektywność* danego serwisu wyszukiwawczego:

Precyzja Określa ile z wydobytych dokumentów pasuje do zapytania w stosunku do liczby wszystkich wydobytych dokumentów

$$Precyzja = \frac{z_p}{z_p + z_n} \cdot 100\%$$

Przywołanie Określa liczbę znalezionych dokumentów w stosunku do wszystkich istniejących dokumentów pasujących do zapytania

$$Przywołanie = \frac{z_p}{z_p + n_p} \cdot 100\%$$

Odrzut Określa liczbę dokumentów odrzuconych w stosunku do wszystkich nie pasujących do zapytania

$$Odrzut = \frac{z_n}{z_n + n_n} \cdot 100\%$$

gdzie: z_p – liczba dokumentów znalezionych pasujących do zapytania, z_n – liczba dokumentów znalezionych i nie pasujących do zapytania; n_p – liczba dokumentów nie znalezionych, pasujących do zapytania; n_n – liczba dokumentów nie znalezionych, nie pasujących do zapytania.

Serwis wyszukiwawczy o dużej efektywności ma wysoką *precyzję* i *przywołanie* oraz niski *odrzut*. Jeżeli przełożymy to na cechy zaproponowanej wyszukiwarki łączącej pozytywne cechy wszystkich trzech kategorii, to osiągnięcie wysokiej efektywności uzależnione zostaje od kilku czynników:

- wysokiej wydajności w indeksowaniu zawartości Internetu – cecha wyszukiwarek,
- umiejętności podziału indeksowanych dokumentów na kategorie tematyczne – cecha katalogów,

- precyzyjnego wyboru dokumentów, które najlepiej pasują do podanego zapytania – cecha metawyszukiwarek.

W dalszej części artykułu przedstawimy techniki, jakie wykorzystują typowe wyszukiwarki w celu poprawy efektywności. Dotychczasowe prace nad wyszukiwarkami pokazały jednak, że stosowane techniki nie są dostatecznie skalowalne – a co za tym idzie, mało efektywne wraz ze wzrostem liczby dokumentów publikowanych w Sieci.

W celu poprawnego zawężania wyników poszukiwań, konieczne jest zastosowanie metod Sztucznej Inteligencji. Poprawnie skonstruowane algorytmy bazujące nie tylko na konkretnym zapytaniu, ale również na poprzednich zapytaniach danego użytkownika, jego profilu zawodowym i tym podobnych informacjach będą w stanie efektywnie zawęzić zestaw dokumentów podanych jako wynik poszukiwań [20]. Okazuje się jednak, że wykorzystanie Sztucznej Inteligencji w samych wyszukiwarkach może być niewystarczające. Proponowanym rozwiązaniem jest tzw. Nowy Internet – czyli SemanticWeb (Sieć Semantyczna).

1.3. Podejście semantyczne

SemanticWeb stanowi kolejne stadium rozwoju Internetu, będąc jednocześnie zbiorem technologii umożliwiających maszynom zrozumienie zawartości Sieci, tym samym czyniąc ją jeszcze bardziej użyteczną dla ludzi [18]. Sieć Semantyczna ma stanowić nie tylko rozszerzenie Internetu do postaci, dzięki której proces poszukiwania informacji w Sieci mógłby zostać maksymalnie zautomatyzowany, przy jednoczesnej poprawie jego efektywności. Otwarta specyfikacja SemanticWeb wyznacza jedynie ścieżki rozwoju Sieci, owocuje to pojawianiem się coraz to ciekawszych propozycji wykorzystania struktury SemanticWeb. W tym artykule ograniczę się jednak do propozycji, które związane są bezpośrednio lub pośrednio z wyszukiwaniem informacji w Internecie.

Podstawowym pojęciem Sieci Semantycznej jest *Ontologia*. Ontologie składają się z taksonomii definiujących klasy obiektów i relacje pomiędzy nimi oraz zbiory reguł wnioskowania. Dzięki zastosowaniu reguł wnioskowania możliwe jest tworzenie nowych połączeń pomiędzy klasami obiektów. Na przykład jeżeli wiemy, że z domem jest związany jego adres, zaś częścią adresu jest nazwa miasta, to możemy zastosować regułę „Jeżeli z adresem jest związany dom i w skład adresu wchodzi nazwa miasta, to z danym domem związana jest jednoznacznie nazwa miasta".

Zdefiniowanie ontologii dla większości klas obiektów otaczającego nas świata, zaowocuje znaczną poprawą rozumienia przez maszyny m.in. zapytań kierowanych do przeglądarki. Aby możliwe było wykorzystywanie wiedzy zgromadzonej w ontologiach, konieczne jest zastosowanie technik Sztucznej Inteligencji. Umożliwi to nie tylko

wydobywanie żądanych zależności pomiędzy klasami obiektów opisanych przez ontologie, ale również umożliwi przeprowadzanie dowodów na poprawność wcześniejszego „rozumowania". Dzięki temu będzie możliwe np. zaprezentowanie użytkownikowi toku przeprowadzonego rozumowania i jego umotywowanie.

Wykorzystanie technologii SemanticWeb przez wyszukiwarkę, która miałaby posiadać cechy katalogów, wyszukiwarek i metawyszukiwarek, zaowocuje nie tylko zawężeniem liczby prezentowanych dokumentów w odpowiedzi na zapytanie, ale również umożliwi poprawne, automatyczne klasyfikowanie zawartości dokumentów w postaci drzewiastej (jak występuje to w katalogach, w których jednak klasyfikowanie odbywa się ręcznie).

2. Architektura wyszukiwarki

Klasyczna wyszukiwarka internetowa składa się z 3 zasadniczych części:

1) baza danych – centralna część wyszukiwarki, w niej przechowywane są informacje o zaindeksowanych stronach,
2) interfejs z użytkownikiem Internetu – odpowiedzialna za pobranie zapytania od użytkownika i zaprezentowanie wyniku wyszukiwania w bazie danych – jest to tzw. część on-line,
3) moduł wyszukujący – odpowiedzialny za przeglądanie zawartości Internetu i indeksowanie w bazie danych – jest to tzw. część on-line.

Schemat typowej wyszukiwarki przedstawia rys. 1. Aby odpowiedzieć na zapytanie klienta, wyszukiwarka musi posiadać wiedzę na temat zasobów Internetu zgromadzoną lokalnie – przeglądanie Sieci za każdym razem, gdy otrzymamy zapytanie jest obecnie niemożliwe, z uwagi na ogromne rozmiary Sieci. Większość wyszukiwarek używa zamiast tego techniki indeksowania, tak aby każdy wyszukany i przetworzony (parsowanie) w Internecie dokument opisać w formie skrótowej.

Najczęściej do takiego skrótu dokumentu używa się opisu w postaci statystyki znajdujących się w nim słów i fraz, sygnaturowanie dokumentów za pomocą funkcji haszujących (i porównywanie z wynikiem działania takiej funkcji na zapytaniu). Najpopularniejszą jednak metodą jest tzw. odwrotny indeks. Metoda ta polega na przechowywaniu w bazie słów i fraz wraz z odnośnikami do dokumentów, w których się one znajdują.

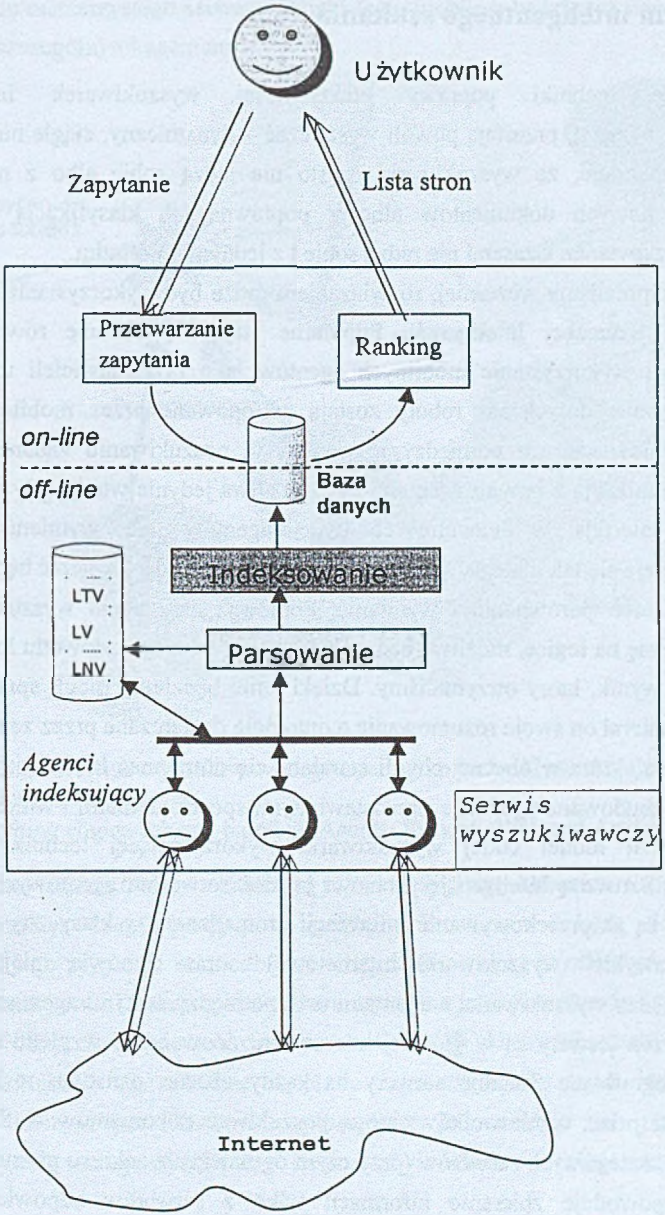
Każdy odczytany dokument jest parsowany również z innego powodu. Wydobywane są z niego adresy innych dokumentów w Sieci. Adresy te są umieszczane w podręcznej bazie danych, z której pobierają je agenci (roboty) wyszukujące. Każdy z dokumentów może być oznaczony jako „do odwiedzenia" (LTV – „link to visit") lub „nie odwiedzać" (LNV – „link

not to visit"). Adresy dokumentów przetworzonych przez agentów wyszukujących oznaczane są jako „odwiedzone" (LV – „link visited").

Poprzednia część odbywa się, oczywiście, w trybie off-line – czyli niezależnie od tego, czy istnieją klienci korzystający z serwera, czy też nie. Właściwa odpowiedź na zapytanie klienta polega na wyszukaniu w lokalnej bazie danych stron (w przypadku korzystania z odwrotnego indeksu operacja taka przebiega bardzo szybko), które odpowiadają zapytaniu. Ponieważ liczba pasujących stron może być bardzo duża, dodatkowo przeprowadza się ocenę trafności – czyli sortowanie wyniku według stopnia, w jakim poszczególne dokumenty spełniają warunki postawione w zapytaniu.

Ponieważ oparcie działania wyszukiwarki jedynie na takiej architekturze byłoby wyjątkowo nieefektywne, zważywszy na obecny rozmiar Internetu, współczesne wyszukiwarki stosują najróżniejsze techniki poprawy efektywności swojej pracy, są to m.in.:

- *klasteryzacja dokumentów* – czyli grupowanie dokumentów o podobnej zawartości, które pasują do tych samych zapytań,
- *stosowanie rankingu wyszukiwarek, określanie dziedziny wiedzy na podstawie zapytania, wielokrotne odpytywanie baz danych innych wyszukiwarek* – techniki stosowane przez metawyszukiwarki, w celu uściślania wyników zapytania do minimalnej grupy wyszukiwarek,
- *web content mining* – wydobywanie wiedzy z zawartości (treści i formatowania) strony,
- *web structure mining* – wydobywanie wiedzy ze struktury powiązań pomiędzy dokumentami w Sieci – określanie stron będących *autorytetami* oraz *węzłów* zawierających odniesienia do autorytetów – algorytm „losowego Surfera"[1],
- *web usage mining* – wydobywanie wiedzy z informacji na temat stopnia wykorzystania poszczególnych obszarów Sieci,
- *word sens disambiguation* – ujednoznacznianie słów tworzących zapytanie, dzięki zastosowaniu serwisów typu WordNet – podane zapytanie jest analizowane i zmieniane z wykorzystaniem np.: tezaurusa, tak aby odnaleźć maksymalną liczbę dokumentów dotyczących danego zapytania,
- *wydobywanie informacji na podstawie formatowania tekstu* – frazom zostają przypisane wagi w zależności od rodzaju formatu, w jakim mają zostać wyświetlone przez przeglądarkę [7, 11], lub następuje próba odtworzenia ontologii na podstawie formatowania HTML (ontologie w SemanticWeb są zapisywane w XML lub RDF).



Rys. 1. Architektura typowej wyszukiwarki
Fig. 1. Typical search engine architecture

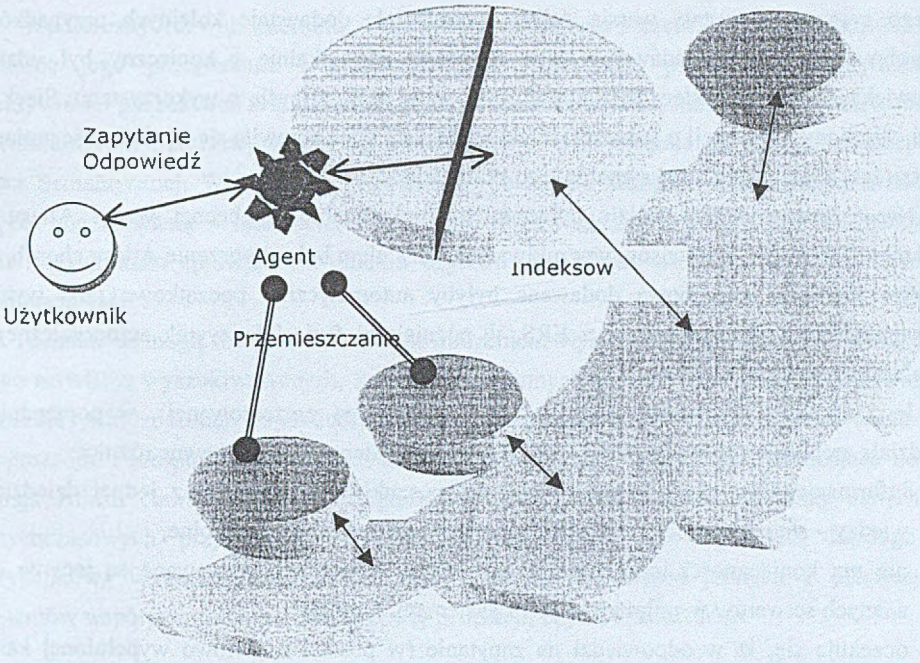
3. Algorytm inteligentnego szukania

Prezentowane techniki poprawy efektywności wyszukiwarek internetowych, wykorzystywane obecnie, przestają powoli wystarczać – dynamiczny, ciągle nie zwalniający rozwój Sieci powoduje, że wyszukiwarki często nie radzą sobie albo z nadążaniem z indeksowaniem nowych dokumentów albo z poprawną ich klasyfikacją i precyzyjną odpowiedzią na zapytania. Czasami nie radzą sobie i z jednym i z drugim.

Jak już wspomnieliśmy wcześniej, rozwiązaniem może być wykorzystanie technik Sieci Semantycznej i Sztucznej Inteligencji. Popularne stają się obecnie również techniki agenturalne, czyli wykorzystanie mobilnych agentów jako przedstawicieli użytkowników Internetu. Stosowane dotychczas roboty zostają zastępowane przez mobilnych agentów mogących przemieszczać się pomiędzy serwerami w poszukiwaniu żądanej informacji. Dotychczas komunikacja z innymi agentami była możliwa jedynie wtedy, gdy oba programy miały ten sam interfejs. W SemanticWeb dwóch agentów może wymieniać informacje między sobą. Dzieje się tak dlatego, że informacje, które posiadają, zawierać będą semantykę, która pomoże w ich porównaniu i wymianie. Ponieważ nasz agent wyszukujący będzie działał opierając się na logice, możliwe będzie otrzymanie od niego „dowodu logicznego" na taki, a nie inny wynik, który otrzymaliśmy. Dzięki temu będziemy mogli sprawdzić, w jak dużym stopniu opierał on swoje rozumowanie o ontologie dostarczane przez zaufane serwery.

Wyszukiwarka, która w obecnej chwili starałaby się dotrzymać kroku rozwojowi Sieci, powinna zostać zbudowana zgodnie z przedstawionymi spostrzeżeniami i wnioskami. Na ich podstawie powstał model takiej wyszukiwarki wykorzystującej techniki agenturalne, SemanticWeb i Sztuczną Inteligencję. Stanowi ją sieć serwerów agentowych. Serwery te odpowiedzialne są za przechowywanie informacji gromadzonej w klasyczny sposób (są to jednocześnie „zwykłe" wyszukiwarki internetowe) oraz stanowią miejsce wymiany informacji pomiędzy wyszukiwarką a agentami oraz pomiędzy samymi agentami. Informacje indeksowane przez serwery są w głównej mierze zróżnicowane ze względu na tematykę i „położenie" geograficzne (lokalne serwery na każdy obszar: państwo, region). Mobilni agenci, wysyłani przez użytkowników, mogą poszukiwać dokumentów w Sieci zbierając informacje z poszczególnych serwerów (przy czym ograniczenie zakresu poszukiwań do np.: domen „.pl" spowoduje zbieranie informacji tylko z serwerów odpowiedzialnych za indeksowanie stron z tych domen). Na podstawie informacji o profilu użytkownika, którego reprezentują, i zastosowaniu Sztucznej Inteligencji możliwe będzie nie tylko sprawdzenie istnienia wyszukanych stron przez agentów, ale również ich weryfikacja pod kątem zwartości oczekiwanej przez użytkownika. Historię takich poszukiwań każdy agent przechowywać

będzie w bazie macierzystego serwera. Dzięki temu możliwa będzie też wymiana informacji pomiędzy poszczególnymi agentami.



Rys. 2. Schemat serwisu wyszukiwawczego opartego o techniki agenturalne, SemanticWeb i Sztuczną Inteligencję
Fig. 2. Searching engine scheme based on Agents, SemanticWeb and Artificial Intelligence

Nieodłączną częścią takiego systemu są również programy klienckie. Programy takie będą stanowiły interfejs komunikacyjny z agentem. Przekazywać będą do agenta profil użytkownika, jego zapytania oraz odbierać od agentów wyniki poszukiwań. Ponieważ agent może na bieżąco powiadamiać użytkownika o wynikach jego poszukiwań (np.: wysyłając mu e-maila lub SMSa), nie będzie konieczna praca on-line. Wykorzystując bazę wiedzy i metody Sztucznej Inteligencji agent będzie dążył do zaprezentowania stron, które rzeczywiście odpowiadają naszym poszukiwaniom, prezentując jednocześnie tok rozumowania, jaki doprowadził go do takiego wyniku. Rys. 2 przedstawia schemat prezentowanej wyszukiwarki.

4. System rekomendacji zastosowań medycznych

Praca nad Atlasem chorób gastroenterologicznych, będącym częścią systemu Endoscopy Recomendation System wykonywanego w ramach grantu KBN, uzmysłowiła nam, że

wykorzystywanie informacji o reprezentatywnych chorobach gastroenterologicznych w postaci statycznej bazy wiedzy będzie w zdecydowany sposób ograniczało funkcjonalność całego systemu. Powstała wersja Atlasu umożliwiała dodawanie kolejnych przypadków chorobowych. Jednak dodawanie takie odbywało się lokalnie i konieczny był udział człowieka. Ponieważ system ERS został zbudowany m.in. z myślą o wykorzystaniu Sieci w celu wymiany informacji o jednostkach chorobowych, więc pojawiła się konieczność zmiany Atlasu tak, aby osiągnąć maksymalną jego skalowalność.

Nowa wersja Atlasu będzie połączeniem funkcjonalności obecnej wersji Atlasu z technikami znanymi z serwisów wyszukiwawczych. Celem było stworzenie Atlasu chorób, w którym jednostki chorobowe dodawane byłyby automatycznie początkowo jako wynik zapytania lekarzy korzystających z ERS, w późniejszej fazie jako wynik autonomicznego działania Atlasu.

Jego architektura została oparta w dużej mierze na prezentowanej w poprzednim rozdziale architekturze inteligentnej wyszukiwarki. Występują jednak pewne różnice:

- informacje, które będą przechowywane i poszukiwane pochodzą z jednej dziedziny wiedzy – stąd też ontologie je opisujące mogą być dużo bardziej spójne,
- nie ma konieczności indeksowania całej Sieci – zapytania kierowane są jedynie do znanych serwerów współpracujących z systemami typu ERS,
- oczekuje się, iż w odpowiedzi na zapytanie (w postaci częściowo wypełnionej karty badania) lekarz otrzyma spójny dokument sugerujący dalsze wypełnianie karty badania; zatwierdzenie takiego dokumentu spowoduje dodanie do bazy wiedzy Atlasu,
- części odpowiedzialna za tworzenie Atlasu w trybie off-line bazować będzie głównie na indeksowaniu nowych przypadku pojawiających się w bazach wiedzy systemów rekomendacyjnych,
- zastosowanie Sztucznej Inteligencji ma wzbogacić wynik zwracany przez Atlas o dodatkowe informacje w postaci dowodu na prezentowany wynik,
- użytkownikiem takiego systemu będzie lekarz lub sam system rekomendacyjny,
- korzystanie z mobilnych agentów umożliwi zbieranie informacji z innych serwerów systemów rekomendacyjnych, które będą się pojawiać; można przyjąć, że na kilka z takich serwerów będzie przypadał jeden serwer Atlasu,
- wykorzystanie technik SemanticWeb zapewni możliwość rozwoju Atlasu, poprawnego przekazywania informacji pomiędzy agentami i serwerami oraz współpracy z innymi systemami.

5. Wnioski końcowe

Współczesny rozwój Internetu wymaga odpowiednich narzędzi, które umożliwiałyby sprawne jego przeglądanie i przeszukiwanie. Prowadzone prace nad poprawianiem efektywności poszukiwań w Sieci skupiają się obecnie na wykorzystaniu do tego celu technik agenturalnych i Sztucznej Inteligencji. Dużym wsparciem okazuje się również koncepcja Sieci Semantycznej. Wszystko to prowadzi do powstania narzędzi wykonujących żmudny proces poszukiwania w Internecie, obdarzonych dużo większą niż dotychczas inteligencją i działających dużo sprawniej w imieniu użytkownika, dzięki wiedzy o jego profilu np.: zainteresowań.

Techniki te mogą również zostać z powodzeniem wykorzystane podczas tworzenia nie tylko serwisów wyszukiwawczych, lecz również i innych systemów, w których użytkownik oczekuje jednoznacznej odpowiedzi na zadane pytanie, a informacje na ten temat są rozproszone i niespójne. Przykładem takiego systemu może być właśnie powstająca nowa wersja Atlasu chorób gastroenterologicznych, z uwagi na rozproszenie bazy danych o dotychczasowych przypadkach chorobowych oraz dużej indywidualności każdego z przypadków. Wykorzystane techniki plasują ten system wśród jednych z pierwszych systemów współpracujących z Siecią nowej generacji, czyli SemanticWeb.

LITERATURA

1. Węgrzyn S.: Podstawy automatyki. PWN, Warszawa 1978.
2. Sugiura A., Etzioni O.: Query Routing for Web Search Engines – Architecture and Experiments.
3. Hammer J., Garcia-Molina H.: Extracting Semistructured Information form the Web.
4. Lawrence S., Giles C. L.: Searching World Wide Web.
5. Beigi M., Benitez A. B., Shih-Fu Chang: MetaSEEk Engine for Images.
6. Benitez A. B., Beigi M., Shih-Fu Chang: Quering the World Wide Web.
7. Bogan J., Freitg D., Joachims T.: A Machine Learning Architecture for Optimizing Web Search Engine.
8. Howe A. E., Dreilinger D.: SavvySearch: A Meta Search Engine that Learns which Search Engine to Query.
9. Craven M.: Learning to Extract Symbolic Knowledge from World Wide Web.
10. Moldovan D. I., Mihalcea R.: Using WordNet and Lexical Operatos to Improve Internet Search.
11. Salton G.: Developments in automatic text retrieval – Science, 1991.

12. Etizioni O.: The world-wide web: Quagmire or gold mine? – Communication of ACM, 1996.
13. Kleinberg J. M.: Authorative sources in a hyperlinked environment, 1998.
14. Głęboki Internet – Świat Nauki.
15. Hender J., Lassila O., Berners-Lee T.: The Semantic Web – Scientific American (May '01).
16. Frauenfelder M.: A Smarter Web.
17. Allen J.: Making a Semantic Web.
18. Dumbill E.: The Semantic Web: A Primer, 2000.
19. Długocki J., Fuszara M.: Sieć Semantyczna.
20. Nick Z. Z., Themis P.: Web Search Using a genetic Algorithm.
21. Semantic Markup for Web Services DAML-S: http://www.daml.org/services/daml-s/2001/10/daml-s.htm.
22. Agenci i SemanticWeb: http://www.computer.org/intelligent/ex2001/x2toc.htm.
23. Strona projektu SemanticWeb: http://www.SemanticWeb.org.
24. Strona projektu Annotea: http://www.w3.org/2001/Annotea/.
25. Strona Firmy Ontoprise: http://www.ontoprise.de/home.htm.
26. Strona projektu DAML: http://www.daml.org.
27. SemanticWeb version of WordNet: http://mozart.isi.edu:8003/sensus2/.
28. Ontologia związana z medycyną – projekt HealthCyberMap: http://healthcyber–map.SemanticWeb.org/.
29. Narzędzie do budowania ontologii i generowania plików RDF: http://protege–.SemanticWeb.org/.

Recenzent: Dr inż. Dariusz R. Augustyn

Wpłynęło do Redakcji 29 marca 2002 r.

Abstract

The paper present the idea of intelligent search for information in Internet. Basic efficiency criteria such as *precision* (the ration of the number of accepted documents to all found documents), *recall* (the ration of the number of found documents to all documents corresponding to the query) and *rejection* (the ratio of the number of unaccepted documents

from all found ones to the number of all existing documents not corresponding to the query) are described in the context of using new technologies (SemanticWeb, Artificial Intelligence). Basing on this a new architecture of search engine is proposed and compared to typical architecture of such tool (see Fig. 1 and Fig. 2). The most important difference between them is the method of indexing documents in particular using semantic properties to limit the number of searched documents and point out the document which is really needed. The implementation of Semantic Web search engines is based on agents environment. Agents are using Artificial Intelligence algorithms to understand ontologies describing web documents. Such a solution is proposed to create new version of Electronic Atlas that is the part of Endoscopy Recommendation System. The Atlas will support collaboration of physicians working in remote medical centers.