Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej Nr 40

XXIV Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE 2014

Oddział Gdański PTETiS

INTELIGENTNE HYBRYDOWE SYSTEMY WYSZUKIWANIA INFORMACJI

Adam Ł. KACZMAREK

Politechnika Gdańska; Wydział Elektroniki, Telekomunikacji i Informatyki
tel: 58 347 13 78 fax: 58 347 22 22 e-mail: adam.l.kaczmarek@eti.pg.gda.pl

Streszczenie: Istnieje wiele metod wyszukiwania informacji. Artykuł przedstawia możliwości połączenia tych metod i skonstruowania uniwersalnego, hybrydowego systemu wyszukiwania. W artykule zaproponowana została architektura personalnego agenta wyszukiwania (PAW). Posiada on cechy systemu ekspertowego, wyszukiwarki oraz agenta komputerowego. PAW pozwala na pozyskiwanie informacji personalnych tj. numery telefonów osób znajomych. Ponadto umożliwia pobieranie informacji z sieci Internet oraz służy do przeszukiwania zasobów Internetu w celu znalezienia informacji wskazanych przez użytkownika, np. księgarnie internetowe oferujące książki w najniższej cenie. Personalny agent wyszukiwania rozszerza możliwości wyszukiwarek internetowych.

Słowa kluczowe: systemy agentowe, wyszukiwanie informacji

1. WPROWADZENIE

Inteligentne wyszukiwanie informacji (ang. intelligent information retrieval) jest pojęciem rozumianym na kilka sposobów. Pojęcie to odnosi się do: wyszukiwania informacji w sieci Internet (ang. web search) [1], systemów agentowych (ang. agent systems) [2], systemów wspomagających podejmowanie decyzji (ang. Decision Support Systems) [3], sieci semantycznych (ang. semantic web) [4] oraz inteligencji sieciowej (ang. web intelligence) [5]. Zagadnienia te nie są od siebie niezależne. Zakresy obejmowanych przez nie tematów częściowo się ze sobą pokrywają. W szczególności zagadnienie inteligencji sieciowej obejmuje szeroki obszar informatyki, w tym temat systemów agentowych oraz sieci semantycznych.

Różne systemy wydobywania informacji charakteryzują się różnymi zaletami i ograniczeniami. Rozwój tego rodzaju systemów realizowany może być przez poprawę jakości poszczególnych rodzajów wyszukiwania, jak również przez tworzenie systemów hybrydowych wykorzystujących różne technologie. Niniejszy artykuł podejmuje temat rozwoju systemów wyszukiwania poprzez skonstruowanie systemu hybrydowego. Do systemów takich należy przedstawiony w tym artykule personalny agent wyszukiwania (PAW).

2. RODZAJE WYSZUKIWANIA INFORMACJI

2.1. Wyszukiwanie w Internecie

Wyszukiwanie informacji w sieci Internet zwykle utożsamiane jest z wykorzystaniem wyszukiwarek internetowych. W wyniku ich rozwoju opracowanych zostało wiele nowatorskich metod wyszukiwania. Jedne z pierwszych systemów wyszukiwania opierały się na częstości występowania różnego rodzaju słów w tekście [6]. W późniejszym okresie spopularyzowana została metoda oceny adekwatności stron internetowych na podstawie analizy odsyłaczy między stronami [1]. Ta technologia stanowiła podstawę skonstruowania najpopularniejszej wyszukiwarki, jaką jest Google. Obecnie podczas wyszukiwania branych jest pod uwagę wiele różnorodnych cech stron internetowych, na przykład, ich popularność.

Wyszukiwarki skonstruowane są tak, że za pomocą robotów internetowych pobierają zawartość stron internetowych, a następnie analizują i przechowują ich zawartość. W wyniku analizy stron tworzone są indeksy. Indeksy zawierają różne dane, w szczególności stosowane są indeksy, dzięki którym w szybki sposób można określić, jakie strony zawierają pewne wyrazy lub pewną sekwencję wyrazów.

Inteligentne wyszukiwanie informacji rozumiane jest jako znajdowanie danych adekwatnych do oczekiwań użytkownika [7]. Takie postrzeganie inteligentnego wyszukiwania informacji jest równoważne z interpretowaniem tego pojęcia jako wyszukiwanie w odpowiedni sposób. Wyszukiwarki w znacznej mierze realizują ten cel, jednak posiadają pewne ograniczenia.

Wyszukiwarki mają ograniczone możliwości dostępu do tzw. sieci ukrytej (ang. deep web) [8]. Jest to zawartość sieci dostępna między innymi na dynamicznie generowanych stronach internetowych, na przykład na stronach dostępnych poprzez formularze i kwerendy. Ponadto wyszukiwarka może być czasowo niedostępna. Wprawdzie występowanie takich sytuacji jest rzadkie, jednak trzeba mieć świadomość, że może wystąpić. Cechą wyszukiwarek jest również to, że istnieje pewne opóźnienie w uwzględnianiu w wyszukiwarkach zmian występujących na stronach internetowych.

Alternatywą dla typowych wyszukiwarek internetowych są wyszukiwarki typu Peer-to-Peer [9]. Ich działanie polega na tym, że użytkownicy Internetu wzajemnie udostępniają sobie informacje. Wyszukiwarki Peer-to-Peer nie posiadają

scentralizowanych serwerów, na których gromadzone są informacje o stronach internetowych. Wszystkie dane niezbędne do przeprowadzenia wyszukiwania są rozdystrybuowane pomiędzy użytkowników wyszukiwarki.

2.2. Systemy ekspertowe

Innym zagadnieniem, którego dotyczy temat wyszukiwania informacji są systemy wspomagające podejmowanie decyzji DSS (ang. decision support systems) oraz systemy ekspertowe (ang. expert systems) [3]. Mają one bardzo szerokie zastosowanie. Używane są między innymi do celów medycznych, takich jak udostępnianie lekarzom informacji pomocnych przy stawianiu diagnoz. Wyróżnić można wiele rodzajów systemów ekspertowych i systemów DSS w zależności od użytej w nich technologii oraz ich zastosowania. Jednak zawsze systemy takie charakteryzują się tym, że posiadają pewną bazę informacji oraz mechanizmy pozwalające na przetwarzanie tych informacji. Systemy wykorzystujące algorytmy sztucznej inteligencji nazywane są inteligentnymi systemami wspomagającymi podejmowanie decyzji (ang. intelligent decision support systems).

2.3. Systemy agentowe

Innym zagadnieniem odnoszącym się do pojęcia inteligentnego wyszukiwania informacji jest tworzenie systemów agentowych. Pojęcie agenta jest definiowane na różne sposoby. Przyjmując definicję podaną przez Woolridge'a, agent jest to program komputerowy działający w określonym środowisku i będący w stanie wykonywać w tym środowisku autonomiczne działania, aby osiągnąć wyznaczone mu cele [2]. Tworzone są również bardziej złożone agenty nazywane inteligentnymi agentami. Charakteryzują się one między innymi tym, że potrafią dostosować się do zmian w środowisku, komunikować się z innymi agentami oraz osiągać swoje cele na różne sposoby [10]. Wyszukiwanie informacji z wykorzystaniem inteligentnych agentów określane jest jako inteligentne wyszukiwanie informacji [11].

Szczególnie istotne dla zagadnienia inteligentnego wyszukiwania informacji są agenty działające w środowisku Internetu. Zaprogramowane są one tak, aby posiadały zdolność przeglądania zasobów internetowych i porównywania znalezionych informacji. Przykładowo, opracowane zostały systemy agentowe przeznaczone do przeprowadzania zakupów za pośrednictwem Internetu [12]. Agenty pracujące w takim systemie służą do wyszukiwania dostępnych ofert oraz identyfikowania oferty najatrakcyjniejszej. Agenty są tworzone w taki sposób, aby potrafiły filtrować informacje, dzięki czemu są w stanie przetwarzać zasoby internetowe mimo nadmiaru danych znajdujących się w tych zasobach. Agenty służące do kupowania w Internecie pełnią funkcję asystenta człowieka. Dzięki nim człowiek nie musi samodzielnie przeglądać zawartości dostępnych w Internecie sklepów internetowych. Agenty tworzą dla użytkownika listę najbardziej atrakcyjnych ofert. Użytkownik podejmuje ostateczną decyzję, którą z ofert wybrać, jednak agenty przeprowadzają wyszukiwanie i oceniają atrakcyjność produktów na podstawie znanych im kryteriów. Ponadto agenty charakteryzują się tym, że uczą się preferencji użytkownika i dostosowują swoje działanie do jego zwyczajów. W zależności od dokonywanych wyborów użytkownika są w stanie dopasowywać rodzaj znajdowanych produktów i sposób ich oceny.

2.4. Sieć semantyczna

W celu stworzenia sieci semantycznej opracowane zostały zestawy standardów zapisu informacji [4]. Sieć semantyczną stanowią informacje zdefiniowane zgodnie z tymi standardami. Informacje w sieci semantycznej zapisywane są w sposób jednoznaczny. W języku naturalnym występują liczne nieścisłości i niejednoznaczności. Powoduje to problemy podczas prób interpretacji przez systemy komputerowe komunikatów wyrażonych w języku naturalnym. W celu rozwiązania tego problemu opracowane zostały standardy zapisu informacji pozwalające na podawanie tych informacji w strukturach sieci semantycznej w jednoznaczny sposób, co pozwala na ich przetwarzanie za pomocą algorytmów komputerowych. Informacje w sieci semantycznej są wyrażone w taki sposób, że mogą być one wykorzystywane zarówno przez algorytmy komputerowe, jak i przez użytkowników.

Do standardów zastosowanych w sieci semantycznej należy: URI (Uniform Resource Identifier), XML (Extensible Markup Language), RDF (Resource Description Framework) oraz OWL (Web Ontology Language). W sieci semantycznej każdy zasób jest identyfikowany za pomocą URI. Jest to łańcuch znaków odnoszący się do nazwy lub adresu zasobu. W szczególności URI może być adresem strony internetowej. Opracowanych zostało wiele wariantów identyfikatora URI odnoszących się do różnego rodzaju zasobów. Innym standardem, na którym oparta jest sieć semantyczna jest XML. Jest to format zapisu danych w plikach tekstowych. Charakteryzuje się tym, że możliwe jest odczytanie danych zapisanych w postaci pliku XML zarówno przez systemy komputerowe, jak i użytkowników. Kolejnym standardem powszechnie stosowanym w sieci semantycznej jest RDF. Standard ten umożliwia stworzenie opisu zasobów składających się z elementów zawierających nazwę zasobu, jego własność oraz wartość tej własności. Na przykład, w ten sposób można określić, że autorem pewnego zasobu jest określona osoba. Innym bardzo ważnym standardem, na którym oparta jest sieć semantyczna jest OWL. Standard ten pozwala na tworzenie ontologii. Ontologie stanowią opis zależności między różnymi zasobami, obiektami oraz klasami obiektów. Ontologia może dotyczyć zarówno obiektów rzeczywistych jak i istniejących jedynie w postaci elektronicznej. Na podstawie ontologii oraz zależności w nich zdefiniowanych możliwe jest przeprowadzanie procesu wnioskowania.

2.5. Inteligencja sieciowa

Inteligencja sieciowa powstaje w wyniku połączenia wielu różnych technologii [5]. Do technologii tych należą, między innymi, technologie informacyjne, sztucznej inteligencji, wyszukiwania informacji, zarządzania danymi, sieci semantycznej oraz systemów agentowych. Podczas tworzenia inteligencji sieciowej bierze się również pod uwagę psychologiczne aspekty użytkowników, ich zwyczaje i zachowanie. Głównym celem powstania inteligencji sieciowej jest utworzenie struktur danych i zasobów wiedzy mających zastosowanie podczas rozwiązywania problemów przez ludzi będących użytkownikami sieci. Konsekwencją przyjęcia takiego celu jest to, że informacje zapisane w strukturach inteligencji sieciowej powinny być zapisywane w taki sposób,

aby mogły być bez problemów interpretowane i przetwarzane przez użytkowników.

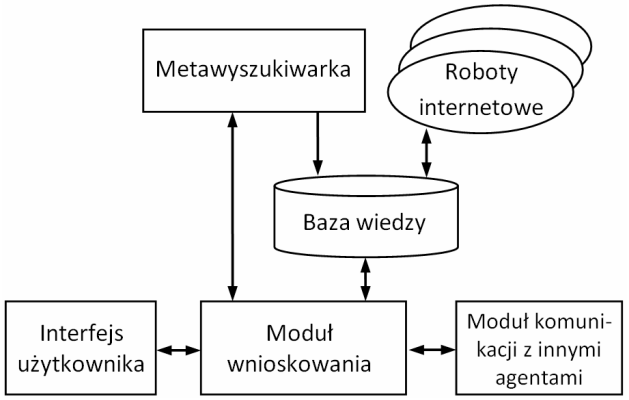
Mechanizmy inteligencji sieciowej w szerokim zakresie opierają się na zastosowaniu technologii systemów agentowych oraz sieci semantycznej. Inteligencja sieciowa obejmuje ponadto technologie pozwalające na pozyskiwanie informacji w wyniku rozpoznawania mowy oraz widzenia komputerowego. W szczególności informacje mogą być pobierane z danych multimedialnych. Dane takie, jak na przykład filmy, zawierają ogromną ilość informacji, która w większości nie jest interpretowana i przetwarzana przez systemy komputerowe. Wprowadzenie możliwości interpretacji danych multimedialnych w inteligencji sieciowej stanowi ogromny potencjał rozwoju systemów informacyjnych.

3. PERSONALNY AGENT WYSZUKIWANIA

Personalny agent wyszukiwania PAW jest systemem hybrydowym posiadającym cechy różnych systemów wyszukiwania.

3.1. Architektura agenta

System PAW składa się z kilku modułów. Jest to interfejs użytkownika, baza wiedzy, moduł wnioskowania, metawyszukiwarka, roboty internetowe i moduł komunikacji. Architektura systemu przedstawiona jest na rysunku 1.



Rys. 1. Architektura personalnego agenta wyszukiwania

Agent posiada elementy systemu ekspertowego, systemu agentowego oraz wyszukiwarki internetowej. W skład agenta wchodzi między innymi baza wiedzy oraz moduł wnioskowania. Są to moduły wywodzące się z konstrukcji systemów agentowych. Agent zawiera również roboty internetowe, które są elementem charakterystycznym dla wyszukiwarek internetowych. Ponadto architektura uwzględnia moduł komunikacji z innymi agentami stosowany w systemach agentowych.

3.2. Moduły wyszukiwania

W bazie wiedzy znajdują się informacje pochodzące z różnych źródeł. Mogą to być zarówno informacje pobrane ze stron internetowych, jak również dane wprowadzone przez użytkownika tj. adresy i dane kontaktowe osób znanych użytkownikowi. Baza wiedzy może przechowywać informacje w różnym formacie. W szczególności mogą one być zawarte w strukturach danych zgodnych z zasadami tworzenia sieci semantycznej. Baza wiedzy może również przechowywać informacje w formatach stosowanych w wyszukiwarkach internetowych oraz systemach ekspertowych. Dane przechowywane w bazie wiedzy powinny ponadto zawierać informacje dotyczącą czasu ich wprowadzenia oraz czasu ostatniej modyfikacji, ponieważ dane te mogą ulegać dezaktualizacji. Na przykład numer telefonu osoby znajomej może zostać przez tą osobę zmieniony. W przypadku konieczności wykorzystywania danych aktualnych niezbędne jest skorzystanie z innych modłów agenta.

Informacje zgromadzone w bazie wiedzy są przetwarzane i prezentowane użytkownikowi za pośrednictwem modułu wnioskowania. Gdy użytkownik poszukuje pewnej informacji, agent w pierwszej kolejności sprawdza dostępność tej informacji w bazie wiedzy. Jeśli jest ona niedostępna wykorzystywane są pozostałe moduły agenta. Agent posiada trzy moduły służące do pobierania informacji. Jest to metawyszukiwarka, roboty internetowe oraz moduł komunikacji z innymi agentami.

Działanie metawyszukiwarki opiera się na tym, że korzysta ona z wyników wyszukiwania innych wyszukiwarek dostępnych w sieci Internet. Metawyszukiwarka jest uruchamiana wtedy, gdy informacja poszukiwana przez użytkownika nie jest dostępna w bazie wiedzy. Metawyszukiwarka może korzystać z różnych serwisów i portali internetowych. Działaniem wyszukiwarki steruje moduł wnioskowania.

W przypadku, gdy wyniki działania metawyszukiwarki nie są adekwatne do oczekiwań użytkownika, personalny agent wyszukiwania posiada możliwość uruchomienia własnych robotów internetowych zdolnych do przeszukiwania sieci Internet. Ich głównym celem jest dostęp do zasobów sieci ukrytej oraz przeszukanie zawartości pewnego zbioru stron internetowych wskazanych przez użytkownika. Użytkownik może wskazać pewien zbiór stron, na przykład zawartość sklepu internetowego lub biblioteki cyfrowej, a następnie agent za pomocą robotów przeszukuje zawartość tych stron w celu zidentyfikowania informacji poszukiwanych przez użytkownika. W przypadku takiego wyszukiwania występuje pewne opóźnienie w podawaniu przez agenta informacji użytkownikowi. Spowodowane jest to koniecznością przetworzenia przez roboty agenta wskazanych im stron internetowych. Jednak dzięki działaniu robotów, użytkownik może uzyskać informacje aktualne i dostosowane do jego oczekiwań.

Kolejną częścią agenta pozwalającą na pobieranie informacji jest moduł komunikacji z innymi agentami. Różni użytkownicy mogą posiadać własne personalne agenty wyszukiwania i może dochodzić do wymiany informacji między tymi agentami. Wymiana informacji odbywać się może w analogiczny sposób jak w przypadku wyszukiwarek Peer-to-Peer. Ponadto agenty różnych użytkowników mogą udostępniać sobie nie tylko informacje ogólnodostępne, które można pozyskać w Internecie. W systemie wieloagentowym istnieje możliwość wskazania grupy agentów zaufanych, którym udostępnia się więcej informacji. Użytkownik mógłby określić, że pewna grupa agentów innych użytkowników należy do grupy agentów zaufanych, przez co agenty te miałyby możliwość dostępu do personalnych danych użytkownika tj. prywatny numer telefonu.

3.3. Interfejs użytkownika

Ważnym modułem personalnego agenta jest interfejs użytkownika. Powinien łączyć w sobie prostotę obsługi z szerokimi możliwościami konfiguracji agenta wyszukiwania. Głównym elementem interfejsu jest pole tekstowe służące do wprowadzania zapytań określających tematy poszukiwanych

przez użytkownika. Użytkownik będzie miał możliwość wskazania czy wyszukiwanie ma być przeprowadzane jedynie na podstawie bazy wiedzy agenta, czy na podstawie metawyszukiwarki czy też w oparciu o roboty lub moduł komunikacji z innymi agentami. Użytkownik może również wskazać, żeby agent przeprowadził wyszukiwanie za pomocą wszystkich dostępnych mu metod.

W konstrukcji interfejsu zakłada się możliwość skorzystania ze słów kluczowych stosowanych w wyszukiwarkach do precyzowania treści zapytań. Na przykład w wyszukiwarce Google do tego rodzaju słów kluczowych należy wyraz *site*. Jeśli użytkownik wpisze to słowo, a następnie po dwukropku poda adres pewnej strony internetowej, to wyszukiwarka Google przeprowadza wyszukiwanie w obrębie tej strony. Personalny agent wyszukiwania również miałby możliwość przetwarzania takich poleceń.

4. WNIOSKI KOŃCOWE

Użytkownicy sieci Internet przeprowadzają wyszukiwanie przede wszystkim za pomocą wyszukiwarek internetowych. Personalny agent wyszukiwania rozszerza ich możliwości. PAW dostarcza funkcjonalność wyszukiwarki, a ponadto udostępnia również inne mechanizmy pozwalające na wydobywanie informacji. Istnieją duże możliwości rozbudowania systemu PAW oraz dostosowania go do potrzeb użytkownika. Korzystanie z personalnych agentów wyszukiwania może stać się równie powszechne, jak używanie wyszukiwarek internetowych.

5. BIBLIOGRAFIA

1. Langville A.N., Meyer C.D.: Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press 2006, ISBN 978-0691122021
2. Wooldridge M.: An Introduction to MultiAgent Systems, John Wiley & Sons 2002, ISBN 978-0471496915
3. Phillips-Wren G., Mora M., Forgionne G.A., Garrido L., Gupta, J.N.D.: A Multicriteria Model for the Evaluation of Intelligent Decision-making Support Systems (i-DMSS), Intelligent Decision-making Support Systems, Springer 2006 s. 3-24, ISBN 978-1-84628-228-7
4. Berners-Lee T., Hendler J., Lassila O.: The Semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American, Nr 284 (5), Nature Publishing Group 2001 s. 34-43, ISSN 0036-8733
5. Curran K., Murphy C., Annesley S.: Web Intelligence in Information Retrieval, Information Technology Journal Nr 3 (2), Asian Network for Scientific Information 2004 s. 196-201, ISSN 1682-6027
6. Salton G., Wong A., Yang C. S.: A Vector Space Model for Automatic Indexing, Communications of the ACM, Nr 18 (11), ACM 1975 s. 613–620, ISSN 0001-0782.
7. Belkin N.J.: Understanding and Supporting Human Information Seeking, Intelligent Information Retrieval: The Case of Astronomy and Related Space Sciences, Nr 182, Springer 1993 s. 9-20, ISBN 978-0-7923-2295-5
8. Hong J.L.: Deep web data extraction, Proceedings of the IEEE International Conference on Systems Man and Cybernetics (SMC), IEEE 2010 s. 3420-3427, ISSN 1062-922X
9. Yang K.-H., Ho J.-M.: Proof: A DHT-Based Peer-to-Peer Search Engine, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, IEEE 2006 s. 702-708, ISBN 0-7695-2747-7
10. Padgham L., Winikoff M.: Developing Intelligent Agent Systems: A Practical Guide, Wiley 2004, ISBN: 978-0-470-86120-2
11. Xiao Y., Xiao M., Zhang F.: Intelligent Information Retrieval Model Based on Multi-Agents, Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing, IEEE 2007 s.5464-5467, ISBN 978-1-4244-1311-9
12. Makris C., Tsakalidis A., Vassiliadis B.: Towards Intelligent Information Retrieval Engines: A Multi-agent Approach, Lecture Notes in Computer Science vol. 1884, Springer Berlin Heidelberg 2000 s. 157-170, ISBN 978-3-540-67977-6

INTELLIGENT HYBRID SYSTEMS FOR INFORMATION RETRIEVAL

Key-words: agent systems, information retrieval

Many different methods have been designed to search for information. This paper presents possibilities of merging these methods in order to acquire a universal, hybrid information search system. In the paper, a novel architecture of a personal search agent is introduced. The agent has features of an expert system, search engine and a computer agent. The personal agent makes it possible to retrieve different kinds of information. This information includes personal data such as telephone numbers and data available in the Internet. Moreover, the agent can process and analyze groups of web pages in order to find specific data indicated by the user. For example, an agent can search for the lowest prices of books in online bookstores. The personal search agent expands capabilities of search engines.