Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej Nr 22

XVI Seminarium
ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE' 2006
Oddział Gdański PTETiS
Referat nr 14

METODY POSZUKIWANIA I SELEKCJI INFORMACJI

Grzegorz KOCHAŃSKI

Politechnika Gdańska
Wydział Elektroniki, Telekomunikacji i Informatyki
Katedra Inżynierii Oprogramowania
e-mail: grzegorz.kochanski@eti.pg.gda.pl

Wzrost liczby nieuporządkowanych publikacji o różnej wiarygodności, znacznie przewyższający zdolność ich przetwarzania sprawił, że znalezienie informacji o wymaganym dopasowaniu jest złożonym problemem. Algorytmy wspomagające poszukiwanie mają na celu przefiltrowanie dużej, nieznanej ilości informacji i zwrócenie kilku reprezentatywnych odpowiedzi o dostatecznej wiarygodności. W artykule dokonano przeglądu rozwiązań dla realizacji tego problemu. Przedstawione zostanły metody siłowe, oparte o bezznaczeniową analizę tekstu, metody selekcji informacji pod względem semantycznym, w oparciu o taksonomię oraz bardziej zaawansowane oparte o automatyczne pozyskiwanie wiedzy i wnioskowanie. Pomimo usilnych starań i różnorodności podejść do problemu wciąż nie znaleziono satysfakcjonujących rozwiązań, które za pomocą danych identyfikujących zapytanie znajdywałyby najbardziej odpowiednie informacje.

1. DEFINICJA PROBLEMU

Dzięki postępowi technologicznemu dostępności Internetu i pamięci masowej magazynowanie oraz rozpowszechnianie informacji nie jest dzisiaj problemem. Internet pełen jest informacji, trzeba ją tylko odszukać. Przy czym odszukanie informacji oznacza wprowadzenie zapytania przez użytkownika, analizę i zrozumienie zapytania przez system, przejrzenie zasobów Internetu, analizę ich i udzielenie użytkownikowi sensownej odpowiedzi. W powyższej procedurze dwukrotnie wykonywany jest proces wnioskowania. David Berlo (1960) [1] opisał rozumienie w kilku predykatach:

- komunikacja nie opiera się o transmisję znaczeń, lecz o transmisję komunikatów,
- znaczenie nie jest w komunikatach, znaczenie jest w nadawcach i odbiorcach komunikatów,
- znaczenia ciągle ewoluują, wraz z doświadczeniem ludzi,
- nie istnieje słowo, które dla dwojga ludzi miałoby dokładnie to samo znaczenie.

Recenzent: Dr inz. Piotr Brudło - Wydział Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej

A zatem automatyczne rozumowanie jest bardzo trudne, jeśli nie niemożliwe. W artykule zostały opisane współczesne metody wyszukiwania, oraz techniki generowania odpowiedzi na zapytanie użytkownika.

2. METODY WYSZUKIWANIA INFORMACJI

Metody wyszukiwania zostały pogrupowane według sposobu definiowania informacji. W rozdziale 2.1 opisane zostały tradycyjne metody wyszukiwania, opierające się na dopasowaniu terminów. Terminem nazywamy tutaj słowo i jego odmiany. Rozdział 2.2) opisuje problem zastosowania statycznej struktury ontologii w ciągle zmieniającym się świecie Internetu. Kolejny rozdział 2.3 przedstawia pierwszą próbę inteligentnego wyszukiwania: grupowanie, które wyszukuje za pomocą silników opartych o metody siłowe oraz grupuje wyniki, pozwalając użytkownikowi odrzucić odpowiedzi nieprzydatne. Wyszukiwanie rozmyte opisane w rozdziale 2.4 jest wzbogacone o analizę już na poziomie zapytania użytkownika. Ideę najbardziej zaawansowanego podejścia, opartego o wiedzę powszechną i wnioskowanie przedstawia rozdział 2.5.

2.1.Metody siłowe

Pierwsza, a zarazem najpopularniejsza grupa metod wyszukiwania, działa w oparciu o dopasowanie terminów (ang. term based). Silnik wyszukiwarki (ang. search engine) taki jak AltaVista [2] czy Yahoo [3], poszukuje dokumentów zawierających terminy z zapytania użytkownika. System wyszukuje najważniejsze słowa i jego odmiany. Dodając kolejne terminy z zapytania użytkownika zawęża obszar wyników. Zaimki i inne często występujące słowa są pomijane przez system, ponieważ nie powodują zmniejszania zbioru.

Zanim system będzie w stanie odpowiedzieć na zapytanie musi zebrać dane. Do przeglądania zasobów Internetu używane są agenty, zwane robotami (ang. robot). Rozpoczynają przeglądanie od wzorcowego zestawu stron, indeksują i oceniają każdą napotkaną stronę. Zindeksowane zostaje każde napotkane słowo, oraz odległości między słowami w tekście. Polega to na tym, że jeśli na stronie X znajduje się zdanie: „Mama ma kota.”, a na stronie Y „Mama Ali ma ładnego kota.” to po wpisaniu w wyszukiwarce słów: „mama kot”, otrzymamy linki do tych dwóch stron, ale na liście wyników strona X będzie wyżej niż strona Y. Następnie agenty przeglądają strony, których linki znalazły się na poprzednich stronach, powtarzając cały proces w głąb. Dynamika Internetu wymusza na robotach ciągłą pracę i ciągłe aktualizowanie ogromnej ilości indeksów. Reguły, wg których robot przegląda strony na serwerze www można definiować publikując plik robots.txt [4]. Ocena trafności strony zależy od tego jakie słowa występują na stronie, oraz w jakich miejscach (np. tytuł, meta-tag, pozycja w tekście, itp.). Wykorzystanie indeksów sprawia, że proces wyszukiwania jest bardzo prosty i efektywny. Znalezienie stron w których występują żądane słowa i posortowanie ich po wcześniej zindeksowanych kryteriach, tj. po ich odległości w tekście rzadko zajmuje więcej niż 1s. Serwisy te oferują również usługi indeksowania danych na komputerze osobistym. Program uruchomiony lokalnie przegląda dokumenty tekstowe, oraz zgromadzoną pocztę i tworzy indeksy za pomocą których możemy później przeszukiwać dokumenty. Filozofia indeksowania polega na traktowaniu zasobów jako płaskiego tekstu, bez analizy treści czy nawet struktury dokumentów.

Powszechnie używana wyszukiwarka Google [5] również działa w oparciu o dopasowanie terminów. W tym przypadku pozycja na liście wyników zależy od cechy popularność. Popularność mierzy się w ilości odnośników do danej strony z innych serwisów. Według Google, o trafności wyników wyszukiwania decyduje ilość odnośników do danej stro-

ny. Silniki wyszukujące nie tyle próbują znaleźć precyzyjne strony odpowiadające zapytaniu użytkownika, co spozycjonować poszukiwaną stroną na odpowiednim miejscu. Nikt nie przejrzy 754 milionów stron po wpisaniu hasła "linux", a jedynie kilka pierwszych z listy. Tabela 1 pokazuje jak ważnym zagadnieniem jest wyszukiwanie informacji. W samych Stanach Zjednoczonych jest to ponad 200 milionów zapytań dziennie. Popularność wyszukiwarki Google oraz brak mechanizmów grupowania/ograniczania liczby wyników wyszukiwania sprawia, że pozycja strony ma ogromne znaczenie, zwłaszcza dla firm komercyjnych.

Tabela 1. Ilość zapytań do serwisów wyszukiwawczych dziennie
(Danny Sullivan, Editor-In-chive, 20 kwiecień 2006)

Searches	Per Day (Millions)	Per Month (Millions)
Google	91	2,733
Yahoo	60	1,792
MSN	28	845
AOL	16	486
Ask	13	378
Others	6	166
Total	213	6,400

2.2.Metody semantyczne

W przeciwieństwie do metod siłowych opartych o pisownię silniki semantyczne podejmują próbę wyszukiwania w oparciu o semantyczne znaczenie zapytania. Sprecyzowanie semantycznego znaczenia zapytania użytkownika odbywa się poprzez szereg interakcji systemu z użytkownikiem, i jest łatwe w realizacji. Jednakże automatyczne budowanie ontologii jak i semantycznych znaczeń słów w dokumentach bez interakcji z użytkownikiem jest trudne i nie daje wymaganych efektów.

Sieć semantyczna organizuje nie tylko pliki multimedialne (tj. strony web, obrazki, pliki dźwiękowe, itd.), ale również ludzi, miejsca, organizacje i zdarzenia. Innymi słowy sieć semantyczna nie ogranicza się do jednego typu relacji (hyperlink) pomiędzy zasobami, lecz stosuje wiele typów powiązań.

Dynamika i brak jednoznaczności w Internecie sprawia, że automatyczna budowa "globalnej" ontologii jest bardzo trudna. Obecnie nie ma w powszechnym użyciu mechanizmów, które automatycznie przeorganizowałyby zasoby Internetu w sieć powiązań semantycznych. Nie oznacza to jednak, że wyszukiwanie semantyczne nie ma sensu. Rozwiązania semantyczne bardzo dobrze sprawdzają się w wąskich zakresach tematycznych, gdzie ontologia budowana jest ręcznie i dotyczy konkretnych, wąskich dziedzin. Definicja zawiera typy obiektów oraz typy relacji między nimi. Wypełnianie sieci semantycznej wymaga od osoby publikującej wprowadzenia metadanych wg powyższej definicji. Narzut wprowadzania dodatkowych danych opisowych wyklucza powszechne użycie metody w Internecie, lecz nie przeszkadza na jej użycie między innymi w firmach, gdzie procesy są dokładnie określone i podlegają ścisłym regułom. Dlatego też, obecnie obserwuje się rozwój komercyjnego rynku silników semantycznych na potrzeby wewnętrznych zastosowań firmowych, korporacji i innych jednostek organizacyjnych wykorzystujących wiedzę jako zasób.

O ile zmienność Internetu nie pozwala zorganizować go w statyczną strukturę, o tyle nie przeszkadza, by w wąskich dziedzinach definiować ontologie tematyczne. Następstwem tego podejścia jest standaryzacja opisu ontologii: definicji obiektów, powiązań [6].

W 1985 roku profesor George A. Miller rozpoczął projekt *WordNet*, w którym słowa języka naturalnego (angielski) są ułożone w leksykalną sieć semantyczną [7]. Obecnie baza *WorldNe*t zawiera ponad 150.000 słów i stanowi podstawę dla wielu systemów semantycznych. Nadzieją na spopularyzowanie semantyki może być możliwość wzbogacania dokumentów o jednoznaczne semantycznie znaczenia słów i zwrotów w wyszukanych stronach przez przeciętnego użytkownika Internetu w popularnych wyszukiwarkach Internetowych.

2.3.Metody grupujące

Silniki grupujące dokonują wyszukiwania w dwóch etapach. W pierwszym etapie strony są wyszukiwane według zasad jak dla metod siłowych. Drugi etap to grupowanie wyników zapytania (ang. post-search process) w rozróżnialne klastry. Bazując na statystykach i technikach wydobywania wiedzy konsolidują "podobne strony" w grupy, które optymistycznie uznają za słuszne. Grupowanie koncepcyjne (ang. conceptual grupping) daje użytkownikowi możliwość zignorowania części wyników niezwiązanych z oczekiwanymi rezultatami, w odróżnieniu od wyszukiwania semantycznego, które ogranicza zbiór wyników w pierwszej kolejności. Przewagą tego rozwiązania w stosunku do silników semantycznych jest brak kosztownej ontologii. Algorytmy statystyczne nie gwarantują, że grupy wyników będą satysfakcjonujące czy nawet zrozumiałe dla użytkownika. Jednym z serwisów stosujących metodę grupowania jest system Vivisimo [8]. W pierwszym etapie wyszukuje strony korzystając z usług ogólnodostępnych serwisów, tj. google, msn, wikipedia, itp., następnie grupuje wyniki jak na rysunku 1 dla hasła „Walt Disney".

Rys. 1. Lista grup dla hasła „Walt Disney" zwrócona przez system Vivisimo

Ann Veling i Peter van der Weerd opisali w [9] technikę usuwania niejednoznaczności poprzez grupowanie. Założyli oni, że słowa współwystępujące (ang. word co-occurence network) mają tendencję to występowania w tym samym znaczeniu. Założenie to nie zawsze jest słuszne i w rezultacie dokumenty zostają nieprawidłowo przydzielone do grup lub powstają niespójne grupy. Mimo to technika ta wciąż się rozwija, Antoni Wolski i Tarik Bouzaziz zaproponowali w [10] metodę definiowania i indeksowania wielkości "blisko siebie" za pomocą logiki rozmytej na poziomie bazy danych.

2.4.Metody rozmyte

Według Lotfi A. Zadeh [11] użycie logiki rozmytej w wyszukiwaniu nie jest opcją, jest potrzebą. Logika rozmyta jest nieprecyzyjna, tak jak i natura Internetu. Wymienia dwa kierunki wykorzystania logiki rozmytej: wyszukiwania i dedukcji. Tim Barnres-Lee twierdzi, że sieć semantyczna musi tolerować logiczne sprzeczności, jak czynią to ludzie. Taka odporność możliwa jest tylko dzięki wykorzystaniu systemów rozmytych.

Do definiowania znaczenia zastosowano rozmyty model pojęcia (ang. Fuzzy Conceptual Model), w którym pojęcie definiowane jest przez serie słów kluczowych i ich wag, zależnie od ich ważności. Dwuznaczności w pojęciach rozwijane są do nieprecyzyjnych, rozmytych zbiorów pojęć. Rozmyte pojęcia powiązane są ze zbiorem słów tworzących kontekst dla danego pojęcia. Następnie, przez szereg interakcji z użytkownikiem, wybierany zostaje kontekst, a tym samym ontologia i dwuznaczność zostaje rozwiązana. W ten sposób bazując na eksperckich formułach lingwistycznych lub wzorcowych stronach Internetowych można przeszukiwać Internet i oceniać trafność napotkanych stron. W rezultacie system dokonując dopasowania koncepcyjnego (wraz z dwuznacznością kontekstową) wyszukuje rezultaty i za pomocą interakcji pozwala mu dostosować strategię wyszukiwania do swoich preferencji.

Inne podejście do reprezentacji znaczenia stosują metody percepcyjne. Metoda percepcyjna opiera swoje założenia o naturę ogólnie rozumianej wiedzy. Wiedza odzwierciedla ograniczone zdolności zmysłów ludzkich i mózgu, między innymi do odtwarzania szczegółowych informacji. Percepcja nie jest precyzyjna, a to powoduje, że reprezentacja pojęć za pomocą spójnej logiki predykatów jest nie jest właściwa (w rozwiązywaniu zapytań użytkownika). Innymi słowy nie ma podstaw teoretycznych, aby stworzyć ortogonalną przestrzeń wektorową pojęć. Rozwinięta przez L. A. Zadeh metodologia wyliczeń opartych na słowach i postrzeganiu (computing with words and perception – CWP) [11] opiera się na założeniu, że postrzeganie można opisać w języku naturalnym. W taki sposób wyliczanie spostrzeżeń zostaje zredukowane do wyznaczenia propozycji wniosków z języka naturalnego, np. jeśli osoba X pracuje w mieście Y, to mieszka w lub blisko miasta. Propozycja wniosku jest precyzyjna jeśli można ją przetransformować do języka precyzyjnego (precisiated natural language – PNL).

2.5. Wnioskowanie

Inne podejście zastosowali twórcy systemu CYC (concept of decuction) [12]. Według nich tylko wiedza w potocznym znaczeniu i wnioskowanie są kluczem do wydobywania informacji. CYC wydobywa znaczenia z informacji przez maszynę, tak jak czynią to ludzie, tzn. z reguł wyciąga wnioski. Gdy nowy wniosek zostanie wydobyty z informacji, może on zostać użyty do wywnioskowania kolejnych, które nie zostały nigdzie jawnie podane. Produkt CYC zawiera ogromną ilość wielokontekstowej wiedzy, bazując na wydajnym silniku wnioskującym. Baza wiedzy zbudowana jest na rdzeniu ponad 1.000.000 wprowadzonych twierdzeń i reguł, zaprojektowanych by posiąść wiedzę ogólną. Na dzisiaj system CYC “wie”, że drzewa rosną zazwyczaj na zewnątrz, czy że ludzie po śmierci przestają kupować rzeczy.

3. WNIOSKI KOŃCOWE

Wzrost dostępnych informacji w Internecie sprawił, że indeksowanie jako technika wydobywania wiedzy przestaje być efektywna. Dodatkowo korzystanie z wyszukiwarki wymaga od użytkownika umiejętności definiowania zapytań. Po pewnym czasie korzystania z wyszukiwarek uczymy się, że poszukując odpowiedzi na pytanie: „Ilu pracowników zatrudnia firma X?”, definiujemy zapytanie jako: „Lista pracowników firma X”. Poza tym użytkownik nie otrzymuje odpowiedzi wprost, a w postaci długiej listy odnośników i we własnym zakresie ocenia ich przydatność. Wydaje się, że wprowadzenie mechanizmu informowania systemu o subiektywnych ocenach użytkowników i uwzględnianie ich przez system podczas procesu pozycjonowania mogłoby poprawić trafność wyszukiwania. Z dru-

giej strony, zauważono dystans jaki dzieli wiedzę od jej słownego opisu. Złożoność automatycznego wnioskowania i analizy języka naturalnego sprawiły, że powstały różne kierunki rozwoju inteligentnego wyszukiwania. W przyszłości słownik ***WordNet*** można wykorzystać do precyzowania znaczeń słów w tekstach, np. przez użytkowników Internetu. Po pewnym czasie system dzięki zgromadzonym danym sam byłby w stanie analizować tekst, odnajdywać kontekst, oceniać wiarygodność. Sam słownik powinien podlegać ewolucji, tak jak język naturalny, który wciąż się zmienia. Wyszukiwanie, selekcja i grupowanie informacji w Internecie to otwarty i ważny współcześnie problem badawczy.

4. BIBLIOGRAFIA

1. http://www.cultsock.ndirect.co.uk/, Mick Underwood, The CCMS Infobase, Communication, Cultural and Media Studies
2. AltaVista, http://www.altavista.com
3. Yahoo, http://www.yahoo.com
4. Koster M., Robots in the Web: threat or treat?, ConneXions, Volume 9, No. 4, April 1995
5. Google, http://www.google.com
6. http://www.w3.org/2004/Talks/0316-semweb-ddc/, Eric Miller, The Semantic Web
7. Miller G. A., Fellbaum C., Miller K. J., Five Papers of WordNet,
8. Vivisimo, http://vivisimo.com
9. Ann Veling, Peter van der Weerd. Conceptual grouping in word cooccurrence network. 16th Join Int. Conf. on Artificial Intelligence (IJCAI'99), PP 694-699, Stockholm, Sweden, 1999
10. Bouaziz T., Wolski A., Fuzzy Triggers: Incorporating Imprecise Reasoning into Active Databases, Proc. IEEE 14th International Conference on Data Engineering. 1998.
11. L. A. Zadeh, From Computing with Numbers to Computing with Words -- From Manipulation of Measurements to Manipulation of Perceptions, IEEE Transactions on Circuits and Systems, 45, 105-119, 1999.
12. CYC, http://www.cyc.com

INFORMATION SEARCHING AND SELECTION METHODS

Finding the well matched information is a complex problem due to the increase number of unorganized, unreliable publications. The search support algorithms are designed to filter out the great amount of data in order to return a few representative, desired reliable answers. The article describes solutions of the issue, including strength methods based on non-meaning text analysis, semantic information selection methods, and more sophisticated techniques based on data mining and reasoning. However, despite of the variety of the searching problem solutions the satisfying mechanism does not exist yet.