Karolina Stasiak, Andrzej Martyna

VSoft SA

Uporządkować chaos informacyjny. Wpływ semantycznych technologii na wyszukiwanie informacji[1]

1. Dostęp do danych

Wydobywanie danych ze wciąż powiększających się zbiorów jest niebanalnym wyzwaniem. Zderzamy się z informacjami o różnorodnej strukturze, pochodzącymi z rozmaitych źródeł i w efekcie coraz trudniejsze staje się dotarcie do tych istotnych. W dobie funkcjonowania przedsiębiorstw opartych na wiedzy i społeczeństwa informacyjnego sprawny dostęp do poszukiwanej wiedzy jest kluczowy.

Dostarczenie najbardziej trafnych dokumentów w odpowiedzi na zapytanie użytkownika jest zadaniem systemów wydobywania informacji[2] (*Information Retrieval* – IR). Użytkownik powinien mieć łatwy dostęp do informacji, której potrzebuje, używając systemu IR z odpowiednią reprezentacją i organizacją danych. Na ogół użytkownik ujawnia swoje wymaganie informacyjne w formie zapytania, zazwyczaj jako ciąg słów kluczowych, następnie system przedstawia elementy, które uzna za trafne[3].

Współcześnie użytkownik komputera korzysta dwutorowo z dostępnych dla niego informacji – są to informacje internetowe oraz bazodanowe (zazwyczaj wewnętrzne przedsiębiorstw). W obu źródłach eksplozja informacji jest trudna do opanowania i wiąże się z przedstawionymi w tabeli 1 problemami.

[1] Badania opisane w niniejszym artykule zostały przeprowadzone w ramach projektu no. POIG.01.04.00–12–075/11 – „Opracowanie innowacyjnej platformy narzędziowo-utrzymaniowej" współfinansowanego przez Europejski Fundusz Rozwoju Regionalnego, Program Operacyjny Innowacyjna Gospodarka 2007–2013.

[2] P. Tiwari, *Extraction of user specified web knowledge using Spatial Data Mining*, „International Journal of Scientific and Research Publications" 2012, vol. 2, issue 2.

[3] C. Macdonald, *The Voting Model for People Search*, PhD thesis, University of Glasgow, 2009.

Tabela 1. Źródła informacji

Dostęp do informacji	
Internet	Bazy danych
– strony internetowe w Web 2.0, prócz kilku znaczników, są jedynie ciągiem znaków, niezrozumiałym dla maszyny; – bez zaawansowanych systemów sztucznej inteligencji wyszukiwarki nie są w stanie wychwycić semantyki zamieszczonej treści	– dane są ustrukturyzowane, wydobywanie wiedzy wymaga umiejętności tworzenia zapytań do bazy danych; – użytkownicy bez znajomości baz danych korzystają z gotowych, niemodyfikowalnych widoków danych

Źródło: opracowanie własne.

2. Internet

W Web 2.0 to internauci są twórcami treści. Dzięki coraz powszechniejszemu dostępowi i globalnemu zasięgowi Internetu użytkowników jest coraz więcej, a w konsekwencji treści przybywa lawinowo – do 2013 r. ruch przepływający rocznie przez Internet wyniósł 667 eksabajtów[4].

Interakcje z dzisiejszymi wyszukiwarkami webowymi można scharakteryzować jako *one size fits all*. To znaczy, że wszystkie zapytania podawane przez różnych użytkowników są traktowane podobnie – jako proste słowa kluczowe, a celem jest wydobycie stron webowych dopasowanych do tych kluczy. Mimo że użytkownik ma konkretne wymagania informacyjne, to liczba zwróconych wyników dla poszczególnych kluczy jest olbrzymia[5].

Współczesne narzędzia wyszukiwania w sieci dobrze radzą sobie z odpowiadaniem na najbardziej powszechne zapytania, jednak nie są bardzo skuteczne w rozpoznawaniu unikalnych celów poszukiwań różnych użytkowników. Główne problemy wynikające ze współczesnej struktury Internetu – Web 2.0 można przedstawić z punktu widzenia poszczególnych interesariuszy: internautów, wyszukiwarek i właścicieli witryn.

[4] *Data, Data Everywhere. A special report on managing information*, „The Economist" 27.02.2010.

[5] U. Rohin, V. Ambati, *Improving Re-ranking of Search Results Using Collaborative Filtering*, AIRS 2006, s. 205–216.

2.1. Internauci

Internauta nie otrzymuje syntetycznych rezultatów poszukiwań. Dostaje tylko listę stron, na których może ewentualnie szukać odpowiedzi, podczas gdy potrzebuje zestawu konkretnych danych, wydobytych z tychże źródeł, np. „lista filmów niemieckich, które powstały na podstawie powieści francuskich". Zapytania typu lista obiektów o zadanych cechach mogą być niewykonywalne w realnym dla człowieka czasie.

Wyszukiwarki nie odgadują kontekstu zapytania internauty. Dadzą te same odpowiedzi dla klucza „kredyt" zarówno dla poszukujących ofert kredytowych, jak i dla poszukujących publikacji naukowych powiązanych z tym kluczem. Niestety, dodanie kolejnego klucza nie zawsze gwarantuje zawężenie wyników. I tak, dla klucza „kredyt" najpopularniejsza wyszukiwarka Google wyszukuje 4 mln wyników, a po doprecyzowaniu „kredyt oferta" wyników jest już 13 mln. Żaden użytkownik nie jest w stanie przetworzyć tylu wyników – gdyby poświęcić 1 sekundę na odwiedzenie każdego dokumentu, zajęłoby to 150 dni!

2.2. Wyszukiwarki, porównywarki

Strony internetowe w technologii Web 2.0 oprócz zawierania kilku znaczników są jedynie ciągami znaków, niezrozumiałymi dla maszyn i trudno przetwarzalnymi przez wyszukiwarki. Z tego powodu powstaje wiele porównywarek poświęconych konkretnym branżom. Jednak i one są ograniczone, bo muszą posługiwać się sztywno określonymi kryteriami wyszukiwania, w którym to podejściu brakuje elastyczności.

2.3. Właściciele witryn

Właściciele witryn, głównie firmy przedstawiające swoje produkty i oferty, pragną przekazać rzetelne informacje swoim klientom, jednak ze względu na ograniczenia pozycjonowania opartego na słowach kluczowych przedstawianie swoich ofert jest bardzo wąskie.

Z powodu ograniczonych kryteriów przedstawiania ofert w porównywarkach firmy nie mają możliwości dotarcia tym kanałem z ofertami specjalistycznymi. Dodatkowo, oferty w porównywarkach są często nieaktualne. Dane nie zawsze są aktualizowane automatycznie, a wykrycie przez klienta nieaktualnej oferty może powodować utratę zaufania do firmy oferującej produkt.

3. Technologie semantyczne

W następstwie pogłębiającego się chaosu w sieci nieuchronna staje się próba uporządkowania tej sfery i ułatwienia maszynom „rozumienia" (rysunek 1).

Rysunek 1. Powstanie koncepcji Web 3.0

Źródło: opracowanie własne.

W starciu z rzeczywistością tradycyjne metody przetwarzania danych nie są wystarczające. Aby dostarczona informacja była wartościowa i najlepiej dostosowana do kontekstu, musi być pod względem znaczeniowym wieloaspektowo i poprawnie opisana, a narzędzia muszą analizować ją w sposób inteligentny. Takie możliwości zarządzania informacją zapewniają technologie semantyczne, które dają maszynom zdolność rozumienia oraz przetwarzania informacji. Technologie te są wyposażone w mechanizmy wnioskowania oparte na logice deskryptywnej pozwalającej m.in. na automatyczną kategoryzację, analizę tekstu, przetwarzanie języka naturalnego czy analizę sentymentu.

4. Web 3.0 – trendy rozwojowe

Na podstawie technologii semantycznych powstała koncepcja Web 3.0, zwana inaczej „semantyczną siecią". Koncepcja ta nie jest nowa. T. Berners-Lee, twórca Internetu i dyrektor W3C, sformułował ten termin ponad dekadę

temu, wierząc, że przemiana ówczesnej sieci nieustrukturyzowanych informacji w „sieć danych" będzie sednem Web 3.0. Sieć semantyczna jest rodzajem grafu połączonych informacji (*linked data*) w taki sposób, aby informacja była wygodna do przetwarzania przez maszyny. Bazując na standardach WC3, języki OWL i RDF służą do strukturalnego i znaczeniowego opisu obiektów w zasobach sieciowych, wykorzystując rozbudowane słowniki semantyczne.

Technologie semantyczne rozwijają się dynamicznie. Jeszcze 10 lat temu organizowano rocznie mniej niż 5 konferencji poświęconych tej tematyce, teraz jest ich już powyżej 30.

Wikipedia stworzyła ontologiczną bazę wiedzy (DBpedię), ekstrahując informacje z udostępnionych artykułów. DBpedia gromadzi miliony danych o miejscach, ludziach, wydarzeniach i relacjach pomiędzy tymi pojęciami. Dzięki tej formie możemy wyszukać np. wszystkie misje kosmiczne, które w załodze miały mniej niż 3 osoby. Również Facebook zapisuje dane w postaci RDF, aby umożliwić wyszukiwanie znajomych np. według ich zainteresowań czy preferencji kulinarnych.

Zastosowanie Web 3.0 można zaobserwować w najpopularniejszej wyszukiwarce Google, która (od 2013 r. w Polsce) w panelu bocznym dostarcza dane dotyczące obiektu, który wychwyciła jako kontekst zapytania. Szybko i w jednym miejscu użytkownik dostaje ustrukturyzowane dane na temat obiektu oraz linki do obiektów relatywnych. Na przykład po wpisaniu „NATO", z prawej strony zobaczymy dane typu: opis, rok założenia, założycieli. Google pracuje również nad porównywarkami usług finansowych. Wprowadził porównywarkę kart kredytowych oraz ubezpieczeń na rynku brytyjskim. Mając na uwadze popularność Google, po wdrożeniu podobnych porównywarek w Polsce można spodziewać się upadku aktualnie działających porównywarek finansowych.

Ustrukturyzowanie informacji nie rozwiązuje problemu dotarcia przez użytkownika do istotnych dla niego danych. Istnieją bazy mające schemat, jakimi na ogół są firmowe bazy danych o klientach, pracownikach, finansach, produktach itd. Okazuje się, że mimo przejrzystych struktur użytkownicy nadal mają problem z szybkim wydobyciem informacji i żądanych danych.

5. Przedsiębiorstwa

Jakość szukania w przedsiębiorstwie (*enterprise search*) ma bezpośredni wpływ na wydajność przedsiębiorstwa. *Enterprise search* na ogół odnosi się do szukania treści w obrębie organizacji. Mimo osiągniętych w ostatnich latach postępów

technologicznych badania pokazują, że pracownicy wciąż poświęcają ogromną ilość czasu na szukanie informacji[6]. S. Feldman i C. Sherman następująco oszacowali koszt przedsiębiorstwa poniesiony przez nieznajdowanie informacji: przedsiębiorstwo zatrudniające 1000 pracowników umysłowych traci od 2,5 mln do 3,5 mln USD rocznie, szukając nieistniejącej informacji, nie znajdując istniejącej informacji lub odtwarzając informacje, które nie mogą być znalezione[7].

Największymi bolączkami są brak wiedzy na temat zawartości dostępnych baz oraz ograniczenia widoków danych tylko do tych stworzonych przez projektanta. Predefiniowane widoki danych stworzone przez projektanta nie zaspokajają wszystkich potrzeb użytkowników, a większość prób stworzenia indywidualnego zestawienia danych powoduje poświęcenie ogromnej ilości czasu na wykonanie zadania, a w efekcie zmniejszenie wydajności przedsiębiorstwa. Dlatego jednym z największych wyzwań, jakim muszą sprostać współczesne systemy informacyjne, jest wydobywanie użytecznych informacji, które odpowiadają wymaganiom informacyjnym użytkownika.

Często informacje są przechowywane w wielu rozproszonych bazach, przez to tworzenie zestawów danych pochodzących z oddzielnych repozytoriów jest uciążliwe i wymaga ręcznej pracy. Dane w firmach powinny być jednolite szczególnie w branżach takich, jak produkcja, medycyna, finanse, bankowość itd. Z powodu różnic pomiędzy firmami i jednostkami organizacyjnymi struktura danych nie jest współdzielona, co może powodować powtarzanie pracy związanej z rozwojem systemów informacyjnych w tych firmach oraz utrudniać dostęp do danych. Wraz z rozwojem baz wiedzy i semantycznego Internetu wiele przedsiębiorstw zaadaptowało ontologię jako ich podstawę koncepcyjną do zarządzania danymi firmowymi[8].

6. Wydobywanie danych

Implementacja technologii semantycznych ma pozytywny wpływ zarówno na przedsiębiorstwa, jak i na użytkowników Internetu. Wybór odpowiedniego

---

[6] D. Hawking, *Challenges in enterprise search*, Proceedings of the 15th Australasian database conference, Dunedin, New Zealand 2004, s. 15–24.

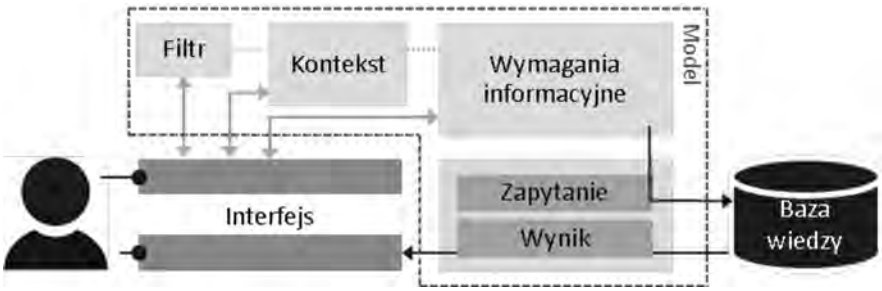
[7] S. Feldman, C. Sherman, *The high cost of not finding information*, Technical Report 29127, IDC, 2003.

[8] Ch. Xie, L. Jiang, H. Cai, *Instance-Driven Ontology Evolution Mechanism towards Enterprise Data Management*, In Proceedings of ICEBE, 2011, s. 24–30.

modelu danych jest jedną kwestią, kolejnym krokiem jest rozwinięcie systemu wydobywania danych dla istniejącego modelu.

W ramach projektu Streamliner opracowano system Vincit, który służy do elastycznego przeszukiwania dowolnej ontologicznej bazy wiedzy. Dzięki tej generyczności może działać zarówno w ontologicznych bazach danych przedsiębiorstw, jak i w zasobach sieciowych opisanych w koncepcji Web 3.0. Vincit jest systemem, który umożliwia dotarcie do wszystkich danych dostępnych w bazie wiedzy, rozpoczynając od dowolnego pojęcia, oraz dalsze eksplorowanie wiedzy w sposób niewymagający kompetencji eksperckich w dziedzinie formułowania zapytań. Użytkownik nadaje kontekst wyszukiwaniu i precyzuje zapytanie w dowolnym kierunku, formułując warunki dotyczące poszukiwanych obiektów oraz danych, jakie chce uzyskać w ich kontekście (rysunek 2). W odróżnieniu od tradycyjnych, jednokrokowych, systemów wyszukiwania, w systemie Vincit można wyróżnić trzy etapy: 1) wybór początkowego pojęcia do eksploracji z listy dostępnych pojęć lub za pomocą fragmentu nazwy, 2) określenie zakresu danych przez filtrowanie według dostępnych atrybutów oraz dołączanie powiązań z innymi pojęciami, 3) wzbogacanie wynikowego zestawu danych o wybrane atrybuty.

Etap 1 odpowiada naturalnej potrzebie użytkownika upewnienia się, czy w bazie wiedzy są potrzebne dla niego pojęcia oraz jak one są tam zaklasyfikowane. Etapy 2 i 3 służą doprecyzowaniu szukanego zakresu danych.



Rysunek 2. Procedura wydobywania danych w systemie Vincit

Źródło: opracowanie własne.

Przykład wyniku zapytania do DBpedii (grafowej bazy danych Wikipedii) przedstawia rysunek 3. Zapytanie brzmi: „Misje kosmiczne z silnikiem startowym (ang. *booster*), który ma więcej niż jeden stopień. Wielkość załogi wynosi mniej lub równo 3. Dla przefiltrowanych misji chcę zobaczyć informację, jaki ma silnik startowy, członków załogi, wielkość załogi, datę lądowania i poprzednią misję. Dla każdego członka załogi chcę zobaczyć jego czas w kosmosie".

space mission	W kategorii	booster	crew member		crew size	landing date	previous mission
			astronaut	time in space (m)			
Apollo 11	event space mission	Saturn V	Neil Armstrong	11520.0 0.5	3	1969-07-24	Apollo 10
			Buzz Aldrin	17280.0 52.0			
			Michael Collins (astronaut)	15964.0			
Soyuz 1	event space mission	Soyuz (rocket)	Vladimir Komarov	3064.0	1	1967-04-24	Voskhod 2
Soyuz TMA-13	event space mission	Soyuz-FG	Yury Lonchakov	289119.0	3	2009-04-08	Soyuz TMA-12
			Michael Fincke	549551.0			
Soyuz TMA-14	event space mission	Soyuz-FG	Michael Barratt (astronaut)	304546.0	3	2009-10-11	Soyuz TMA-13

Rysunek 3. Przykład zestawu danych uzyskanych w systemie Vincit

Źródło: opracowanie własne.

Przykłady innych złożonych zapytań możliwych do przetworzenia w kilku krokach przez Vincita:

- Spis firm informatycznych z miejscowości, których kod pocztowy zaczyna się od 97, z nazwiskami właścicieli.
- Lista utworów Michaela Jacksona, które wykonuje z innym muzykiem. Dla każdego utworu wyświetl datę wydania oraz wykonawców.
- Lista pracowników, którzy posiadają certyfikat z ITIL i biorą udział w przynajmniej trzech projektach dla największego klienta oraz pochodzą z miasta mającego więcej niż 500 tys. mieszkańców, z ich danymi kontaktowymi.

7. Podsumowanie i kierunki dalszych badań

Ilość informacji z każdym rokiem wzrasta wykładniczo, dotarcie do tych istotnych jest coraz trudniejsze, a czasem niemożliwe do wykonania w realnym dla człowieka czasie. Wyszukiwarki niewątpliwie ułatwiają nam korzystanie z zasobów WWW. Jednak pomimo rozwoju teorii wyszukiwania, metod przetwarzania języka naturalnego, procedur analizy stron internetowych oraz zastosowania wydajniejszych algorytmów sprawność wyszukiwarek nie jest jeszcze zadowalająca. Czas wyszukania jest zdumiewająco krótki, jednak liczba wyników jest nierealna do przetworzenia, a dodanie kolejnych słów kluczowych nie gwarantuje zawężenia wyników.

Straty związane z nieznajdowaniem informacji ponoszą przedsiębiorstwa, gdy pracownicy szukają nieistniejących informacji, nie znajdują istniejących informacji lub gdy z powodu utrudnionej wymiany wiedzy pomiędzy działami nieświadomie duplikują już istniejące dane. Dlatego potrzebne są nowe metody zarządzania informacją, które łączy wspólna potrzeba: szybkie wydobycie dostępnych i potrzebnych użytkownikowi danych. Zarówno dla przedsiębiorstw, jak i dla zasobów sieciowych pomocne są technologie semantyczne, ontologie i grafowe bazy danych. Technologie semantyczne nie zastąpią tradycyjnych metod przedstawiania treści, ale stanowią osobną warstwę, odblokowując ogromne możliwości eksplorowania danych.

Bibliografia

Data, Data Everywhere. A special report on managing information, „The Economist" 27.02.2010.

Feldman S., Sherman C., *The high cost of not finding information*, Technical Report 29127, IDC, 2003.

Hawking D., *Challenges in enterprise search*, Proceedings of the 15th Australasian database conference, 2004, Dunedin, New Zealand, s. 15–24.

Macdonald C., *The Voting Model for People Search*, PhD thesis, University of Glasgow, 2009.

Rohini U., Ambati V., *Improving Re-ranking of Search Results Using Collaborative Filtering*, Information Retrieval Technology, Third Asia Information Retrieval Symposium, 2006, s. 205–216.

Tiwari P., *Extraction of user specified web knowledge using Spatial Data Mining*, „International Journal of Scientific and Research Publications" 2012, vol. 2, issue 2.

Xie C., Jiang L., Cai H., *Instance-Driven Ontology Evolution Mechanism towards Enterprise Data Management*, In Proceedings of ICEBE, 2011, s. 24–30.

* * *

To capture the information chaos. Impact of the semantic technologies on the information retrieval

Summary: The amount of information grows exponentially every year. Storing, accessing and retrieving the information becomes a real challenge in enterprises as well as in the World Wide Web. Enterprises bear the huge costs when they fail to find information. The increasing number of web datasets results in difficult retrieval of information from the Internet. Some queries, such as "German movies which are based on a French novel", are unable to be performed by a human in real time. That inconvenience originates from the Web 2.0 structure, the content of which is hard to process for machines. The response to that problem are semantic technologies, and their implementation is beneficial for databases and web users. The paper describes information retrieval difficulties, semantic technologies and the development trends of the Semantic Web. The Vincit system is presented, a flexible solution that allows for complex queries creation for any ontological base of knowledge. It works both in enterprise environment and in web resources as well.

Keywords: semantic technologies, Web 3.0, data retrieval, search system