ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 261, 2011

Beata Gontar[*], *Joanna Papińska-Kacperek*[**]

SEMANTYCZNE WYSZUKIWARKI INTERNETOWE

1. WPROWADZENIE

Sednem formacji **społeczeństwo informacyjne** jest produktywne wykorzystywanie informacji, zatem jej znalezienie i ocena wartości powinna być istotną umiejętnością wszystkich obywateli. Sieć Internet ułatwiła to, ale okazuje się jednak, że nie do końca. Współczesne wyszukiwarki internetowe nie znajdują efektywnie tego, czego ich użytkownicy rzeczywiście szukają. Stąd powstała koncepcja budowania nowej struktury WWW, czyli sieci semantycznej, która stworzy także nowy model relacji w świecie *online*. Celem autorek niniejszego artykułu jest przedstawienie wyszukiwarek semantycznych, ich budowy oraz próba podsumowania bieżącego etapu ich rozwoju.

2. INFORMACJA I JEJ ZNACZENIE

F. von Hayek jako pierwszy zwrócił uwagę na znaczenie informacji w życiu gospodarczym – traktował ją jako obiekt materialny, czyli towar, już w pracach opublikowanych przed II wojną światową. Dopiero po wojnie zaczęto dostrzegać rolę informacji nie tylko w polityce, ale również w życiu społecznym i gospodarczym. M. Uri Porat w opublikowanej w 1977 r. rozprawie *The Information Economy: Definition and Measurement* prognozował, że gromadzenie i dystrybucja informacji stworzą nową jakość gospodarki, co na pewno będzie miało wpływ na życie społeczne[1]. Obecnie, w epoce społeczeństwa informacyjnego, doceniono jeszcze bardziej znaczenie informacji w naszym życiu. To dzięki niej możemy pracować i kontaktować się wzajemnie. Informacja jest pojęciem definiowanym w wielu dyscyplinach naukowych[2]. W teorii

[*] Dr, Katedra Informatyki, Wydzial Zarządzania Uniwersytetu Łódzkiego.
[**] Dr inż., Katedra Informatyki, Wydzial Zarządzania Uniwersytetu Łódzkiego.

[1] J. Papińska-Kacperek, *Nowa epoka – społeczeństwo informacyjne,* [w:] J. Papińska-Kacperek (red.), *Społeczeństwo informacyjne*, Wydawnictwo Naukowe PWN, Warszawa 2008, s. 13–46.

[2] J. Zawiła-Niedźwiecki, K. Rostek, A. Gąsiorkiewicz, *Informatyka gospodarcza*, C.H. Beck, Warszawa 2010.

systemów i cybernetyce występuje, obok materii i energii, jako jeden z trzech zasadniczych elementów wymiany pomiędzy układami względnie odosobnionymi a otoczeniem. W ujęciu inżynierskim, czyli klasycznej teorii informacji, informacja jest ściśle związana z teoretyczną koncepcją „systemu komunikacyjnego". Podstawy ilościowej teorii informacji przedstawił Claude Shannon w swojej pracy *A Ma-thematical Theory of Cryptography* w 1945 r. Teoria ta opisuje informację za pomocą modelu matematycznego oraz metody jej przetwarzania, np. w celu transmisji i/lub kompresji[3].

Wyróżnić można również informację biznesową, czyli dane, fakty i statystyki potrzebne przedsiębiorstwu do podejmowania decyzji i do budowania wiedzy.

Komunikowanie się jest procesem przekazywania informacji od jednej osoby do drugiej, jednak aby odniosło zamierzony efekt musi być skuteczne, czyli zrozumiałe dla odbiorcy. Na ten aspekt zwraca uwagę tzw. semantyczna teoria informacji, gdzie informacja to zbiór wiadomości o faktach, zdarzeniach, cechach przedmiotów itp. ujęty i podany w takiej formie, że pozwala odbiorcy (człowiekowi lub algorytmowi) ustosunkować się do zaistniałej sytuacji i podjąć odpowiednie działanie. Jest to przedmiot zainteresowania infologii, która zajmuje się wyjaśnianiem znaczenia informacji w aspekcie użytkowym, badaniem jej własności, analizą oczekiwań użytkownika, kierowanych pod adresem informacji oraz poszukiwaniem metod i sposobów ich zaspokojenia. Wiedza to układ opisany wyrażeniem

$$w := < I, C, D >,$$

gdzie: w oznacza wiedzę, I – informację, C – kontekst, D – doświadczenie.

Źródłem wiedzy odbiorcy jest informacja, na której odbiór ma wpływ kontekst sytuacyjny oraz posiadane przez odbiorcę doświadczenie[4]. Można tu zaobserwować tzw. decyzyjność informacji, czyli jej wpływ na podejmowane decyzje i działania.

Powszechnym źródłem informacji w dzisiejszych czasach jest Internet. Zasoby wiedzy zgromadzone w niej, czyli w formie elektronicznej, są ogromne, ale stosowane obecnie metody wyszukiwania nie pozwalają ich w pełni wykorzystać. Dlatego nie zawsze wyniki wyszukiwania informacji w Internecie są zadowalające.

---

[3] M. Pawełczyk, *Informacja a niepewność*, materiały do zajęć [2003] http://marpaw.elisa.pl/wsti/roznosci/pomiar_inform/inform.htm (odczyt 10.12.2011).

[4] B. Stefanowicz, *Informacja*, Wydawnictwo SGH, Warszawa 2004, s. 123.

3. CHARAKTERYSTYKA I STRUKTURA SIECI SEMANTYCZNEJ

Pod koniec XX w. rozpoczęto prace nad projektem T. Bernersa Lee: Semantic Web (sieć semantyczna nazywana też Web 3.0), który ma przyczynić się do utworzenia i rozpowszechnienia standardów opisywania treści w Internecie, w sposób, który umożliwiłby maszynom i programom (np. robotom wyszukiwarek, autonomicznym agentom) przetwarzanie informacji w sposób odpowiedni do ich znaczenia. Czas sieci semantycznej według prognoz N. Spivak miał się zacząć w 2010 r.[5] Czy tak się stało? Istnieją już strony stosujące standardy RDF (*Resource Description Framework*) czy OWL (*Ontology Web Language*) (tab. 1), ale Web 3.0 naprawdę stanie się rzeczywistością, gdy wszystkie strony dostosują się do nowych norm, bowiem wtedy budowane obecnie aplikacje, jak np. wyszukiwarki semantyczne, będą działać tak, jak tego oczekują zwolennicy nowej struktury sieci WWW.

Tabela 1. Liczba plików semantycznych w sieci WWW indeksowanych przez Google

Wyszukiwany łańcuch	Liczba stron (maj 2004)	Liczba stron (czerwiec 2011)	Liczba stron (listopad 2011)
Rdf	5 230 000	141 000 000	148 000 000
Filetype:rdf	246 000	9 800 000	21 600 000
Filetype:owl	1 310	15 400	309 000
Filetype: rdfs	304	628	197 000

Źródło: opracowanie własne na podstawie Ding et al., *Swoogle: A search and metadata engine for the semantic Web*, CIKM 2004.

Wyszukiwarki semantyczne mają sprostać potrzebom internautów lepiej niż stosowane teraz narzędzia, jak popularny obecnie Google. Celem wyszukiwania nie powinno być tylko wyświetlanie linków, ale kompetentna odpowiedź na zapytanie użytkownika (jak robi to Wolfram Alpha). Wyniki powinny być też w czytelniejszy sposób wyświetlane, np. linki mogą być pogrupowane (jak w Hakii), lub skategoryzowane graficznie (jak w KOoLTORCH). Kolejną cechą jest umożliwienie tworzenia zapytań w języku naturalnym, czyli formułowanie pełnych zdań – dziś w Google preferowane są równoważniki, a pewne wyrazy (np. przyimki) nie są w ogóle brane pod uwagę. Kolejne zadanie jest nadal poza zasięgiem tradycyjnych wyszukiwarek – to możliwość szukania odpowiedzi w bazach danych. Google i każda inna wyszukiwarka potrafi znaleźć stronę np. z rozkładem jazdy, ale nie da sobie rady z wypełnieniem formularza i zadaniem pytania o konkretne połączenie. To właśnie ma być elementem nowego modelu przeszukiwania i wykorzystywania zasobów Internetu. Jedna z tworzonych wyszukiwarek Evri promowała hasło *search less, understand more*.

[5] N. Spivak *How the WebOS Evolves?* [2007], http://novaspivack.typepad.com/nova_spivacks_weblog/2007/02/steps_towards_a.html (odczyt 10.12.2011).

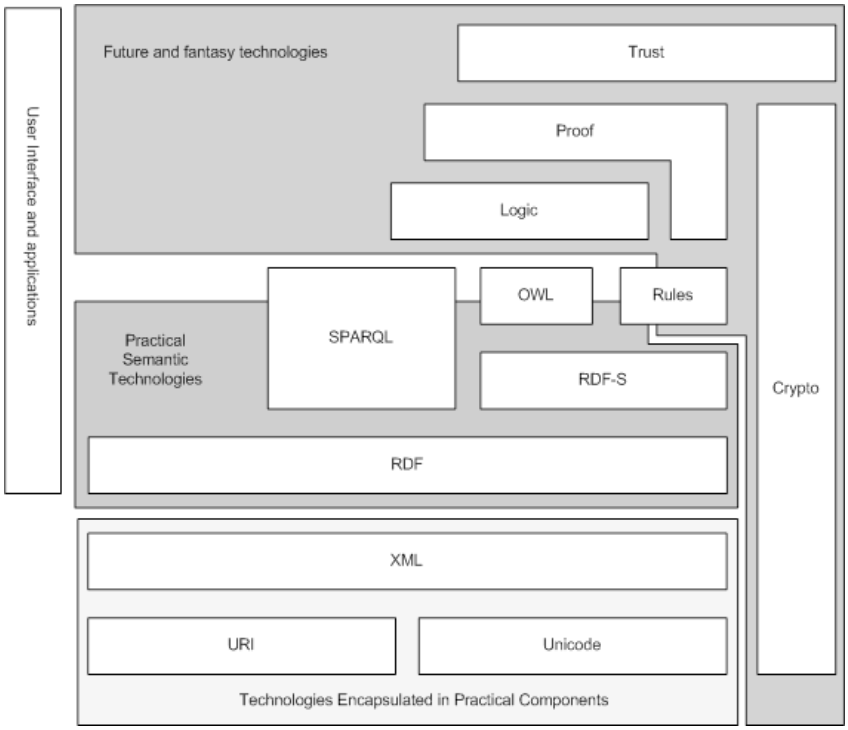
W swoich założeniach sieć semantyczna ma korzystać z istniejącego już protokołu komunikacyjnego, na którym bazuje dzisiejszy Internet. Różnica ma polegać na tym, że przesyłane dane mogą być rozumiane także przez algorytmy aplikacji. Dane przekazywane będą w postaci, w której można powiązać ich znaczenia między sobą, a także w ramach kontekstu, w jakim występują. Informacje przekazywane w sieci wymagają także informacji o nich samych tzw. metadanych, które ułatwiają dostrzeganie powiązań między obiektami. Dzięki temu można łączyć informacje znajdujące się w Internecie w obszarze jednakowych jednostek znaczeniowych (np. strony dotyczące historii sztuki, kuchni włoskiej, wybranej dziedziny nauki), właściwie zinterpretować dane, które są w tej chwili nierozróżnialne ze względu na identyczny zapis tekstowy (np. **zamek** – element do zamykania drzwi; element do łączenia w ustalonym położeniu części ubrania; budowla mieszkalno-obronna), uzyskać nowe informacje, które nie są zawarte w sposób jawny, a otrzymywane są w wyniku wnioskowania, (czyli np. na podstawie zdania **Leszek jest synem Beaty** możemy też dowiedzieć się, że Beata jest kobietą, Leszek mężczyzną, Beata jest mamą, Beata jest mamą Leszka).

Znaczenia zasobów informacyjnych określa się za pomocą ontologii – dziedziny powstałej na gruncie filozofii i łączącej filozofię, matematykę i nauki informacyjne. W celu zapewnienia dokładności opisu wiedzy stosuje się hierarchizację oraz kategoryzację pojęć[6]. Hierarchizacja jest umiejscowieniem pojęcia w strukturze, co umożliwia dziedziczenie cech po pojęciach nadrzędnych. Kategoryzacja jest przypisaniem pojęcia do grupy pojęć podobnych, mających cechy wspólne, wyróżniające klasę pojęć spośród innych. W informatyce ontologia oznacza formalny sposób reprezentacji wiedzy poprzez zdefiniowanie pojęć w pewnej dziedzinie, ich właściwości oraz relacji pomiędzy nimi. Zajmuje się opisywaniem pewnego fragmentu rzeczywistości. W założeniu ontologia powinna możliwie dokładnie określać i reprezentować wiedzę z definiowanej dziedziny i ściśle określać hierarchię jej elementów oraz kryteria ich klasyfikacji za pomocą narzędzi logiki (aksjomatów, definicji, reguł). Praktyczna realizacja ontologii polega na zapisaniu wiedzy w postaci drzewa, gdzie w wierzchołkach zapisuje sie pojęcia, a krawędzie opisują typ relacji. W korzeniu drzewa zapisuje sie pojecie najbardziej ogólne, właściwe dla opisywanej dziedziny wiedzy, a schodząc stopniowo w dół pojęcia o większym poziomie szczegółowości, przy zachowaniu zasady, że wierzchołek nadrzędny zawsze jest uogólnieniem podczepionych do niego wierzchołków podrzędnych. W niektórych realizacjach dopuszcza sie również możliwość posiadania dwóch lub więcej wierzchołków nadrzędnych (na przykład **pióro** może być jednocześnie *narzędziem do pisania* oraz *elementem upierzenia ptaków*), co zamiast

[6] W. Gliński, *Ontologie. Próba uporządkowania terminologicznego chaosu*, [w:] B. Sosińska-Kalata i in. (red.), *Od informacji naukowej do technologii społeczeństwa informacyjnego,* Miscellanea Informatologica Varsoviensia, Wydawnictwo SBP, Warszawa 2005.

drzew wymusza wykorzystanie do zapisu struktury wiedzy acyklicznych grafów skierowanych. Wyszukiwanie informacji w ontologii odbywa się poprzez zbieranie cech z wierzchołków począwszy od pojęcia wyjściowego, poprzez pojęcia bardziej ogólne, aż do wierzchołka drzewa. Ontologie wykorzystują teorie wywodzące się z algebry, teorii zbiorów, sieci semantycznych oraz rachunków logicznych[7].

Sieć semantyczna zbudowana ma być na bazie już istniejących, wykorzystywanych i sprawdzonych standardów internetowych, nadbudowanych przez kilka kolejnych, co ilustruje rys. 1. Każdy standard nakłada się na kolejny. Popularne ich określenie to semantyczny stos (*semantic stack*)[8]. Do elementów semantycznego stosu zaliczamy stosowane już standardy, dedykowane i będące na etapie badań. Stos tworzą zatem Unicode, URI, XML i XML Schema, RDF i RDF Schema, OWL, mechanizmy wnioskowania i mechanizmy certyfikacji i zaufania.



Rysunek 1. Elementy semantycznego stosu

Źródło: T. Segaran, C. Evans, J. Taylor, *Programming the Semantic Web*, O'Reilly Media 2009.

[7] W. Gliński, *Ontologie. Próba uporządkowania…*; idem, *Języki i narzędzia do tworzenia i wyszukiwania ontologii w kontekście semantycznego Weba,* [w:] B. Sosińska-Kalata i in. (red.), *Od informacji naukowej…*

[8] T. Segaran, C. Evans, J. Taylor, *Programming the Semantic Web*, O'Reilly Media 2009.

Unicode jest standardem pozwalającym na przedstawienie w języku maszyn dowolnego znaku pisanego. URI (*Uniform Resource Identifier*) zapewnia unikatowość zasobów internetowych. Wszystkie dane przesyłane w sieci są zasobami internetowymi i wymagają określenia dla nich identyfikatora (łańcucha znaków), który składa się właśnie z zestawu znaków Unicode. Najbardziej popularnym URI jest URN (*Uniform Resource Name*) lub adres URL (*Uniform Resource Locator*) zasobu identyfikowanego przez dany URI.

XML to język pozwalający na zapis danych. Schematy XML wprowadzają ograniczenia dotyczące typu i struktury danych. Zachowanie ich daje gwarancję, że dane w XML są poprawne w sensie syntaktycznym (np. w polu, w którym oczekujemy wartości liczbowej, wartość taka się pojawi).

RDF pozwala na zapis danych w postaci grafu skierowanego. W grafie tym dane zawarte są w wierzchołkach (podmiot i obiekt), a relacje pomiędzy nimi wskazuje predykat. Schematy RDF wprowadzają do grafów takie pojęcia, jak klasy i podklasy, pozwalające na wspólne grupowanie danych mających cechy wspólne. Dowolna dana może znajdować się w wielu klasach. Strukturą każdego wyrażenia RDF jest zbiór trójek (podmiot-relacja-obiekt). OWL jest standardem pozwalającym na definiowanie klas na podstawie własności danych.

Mechanizm wnioskowania (*reasoner*) ma umożliwiać przeprowadzenie wnioskowania na podstawie zdefiniowanych ontologii. Mechanizmy certyfikacji i zaufania pozwolą na określenie praw, na jakich zasoby internetowe mają być przesyłane i mogą być udostępniane.

Jak widać na rys. 1 zarówno mechanizmy wnioskowania, jak i certyfikacji pozostają na razie standardami przyszłości – *future and fantasy technologies*.

Powinno się pamiętać, tworząc nowe dokumenty internetowe, by były widziane przez oba typy wyszukiwarek: standardowe i semantyczne. W przypadku już istniejących, dużych baz danych, nie ma sensu ingerowania w ich strukturę, ale zasadne jest budowanie bazy ontologii. Przykładem może być projekt DBpedii, którego celem jest wydobycie zależności zapisanych w hasłach Wikipedii, czyli stworzenie bazy ontologii. Hasło Wikipedii to w pewnym stopniu uporządkowana struktura, obok podzielonej na sekcje części opisowej, zawiera zdjęcia, informacje o kategoryzacji zdjęć i linki do zewnętrznych źródeł. Struktura ta jest wydobywana i zapisywana w bazie danych, którą można już przeszukiwać, ponieważ jest udostępniona w sieci WWW na licencji *free software*. DBpedia pozwala zadawać zapytania o relacje i właściwości zasobów Wikipedii. Wydobywa odpowiedzi na pytania, które znajdują się w wielu różnych artykułach Wikipedii. Służy do tego język zapytań SPARQL – czyli SQL dla plików RDF. Projekt rozpoczął się na uniwersytetach w Berlinie i Lipsku, we współpracy z OpenLink Software. Jego pierwsze efekty zostały

upublicznione w 2007 r. W styczniu 2011 r. zestaw danych DBpedia zawierał opis ponad 3,5 miliona obiektów, z czego 1 670 000 było sklasyfikowanych w spójnych ontologiach[9].

4. WYSZUKIWARKI SEMANTYCZNE

Tworzenie sieci Web 3.0 nie jest sztuką dla sztuki czy pracą wyłącznie teoretyczną, ale pomysłem na rozwiązanie problemów wszystkich użytkowników Internetu. Zatem nie tylko poprawiana powinna być zawartość dokumentów tworzących WWW, ale powstawać powinny aplikacje wykorzystujące nową strukturę. Ich przykładem są wyszukiwarki semantyczne, czyli narzędzia wyszukiwania informacji w Internecie, korzystające z nowej struktury sieci. Oczekuje się, że to rozwiązanie, dzięki analizie znaczenia, a nie znalezienia zadanej frazy, da lepsze rezultaty niż stosowane dzisiaj np. Google, którego wyniki wyszukania bardzo często nie spełniają oczekiwań użytkowników sieci.

Obecnie w Internecie funkcjonuje kilka rodzajów wyszukiwarek semantycznych. Różnią się formą przyjmowanych zapytań, algorytmem wyszukiwania oraz sposobem wyświetlania wyników. Nie wszystkie były od początku uniwersalne, np. Kosmix zaczynała od wyszukiwania w dziedzinie medycyny, czyli była wertykalna, i po pewnym czasie, w 2008 r. stała się wyszukiwarką horyzontalną, odpowiadającą na pytania z innych dziedzin. Pewne wyszukiwarki nie są nazywane semantycznymi, ale hybrydowymi, ponieważ łączą różne cechy nie zawsze przypisywane tylko semantycznym. I tak, niektóre zaliczane są do grup: *general* (np. Bing), *Natural Language Search Engines* NLSE (np. True Knowledge), *visual search engines* (KOoLTORCH), *automatic answers* (WolframAlpha).

Większość wyszukiwarek semantycznych akceptuje pytania formułowane w języku naturalnym, np. Wolfram Alfa, True Knowledge, Yebol, KtoCo, Hipisek, zatem należą do grupy NLSE, do których zaliczane są też serwisy *crowdsourcing* – Q&A (Questions & Answers), czyli wyszukiwarki społecznościowe jak Yahoo Answers, Answerbag.com, Dig, Aardvark czy Pytamy.pl[10] Te jednak nie starają się analizować zapytań, ale jedynie je magazynują, oceniają, czasem stosują statystyczne metody przydziału pytań. W Polsce powstaje ich szczególna odmiana: semantyczny serwis społecznościowy „Węzełki". Jego użytkownicy mają tworzyć wspólne repozytorium wiedzy, w którym możliwe

[9] J. Papińska-Kacperek, *Przykłady zastosowań serwisów społecznościowych*, „Zeszyty Naukowe Uniwersytetu Szczecińskiego" 656/2011, Studia Informatica 28.

[10] B. Gontar, J. Papińska-Kacperek, *Wyszukiwarki semantyczne*, [w:] M. Pańkowska (red.), *Wiedza i komunikacja w innowacyjnych organizacjach. Komunikacja elektroniczna*, Wydawnictwo UE, Katowice 2011.

będzie semantyczne wyszukiwanie, przeglądanie i współdzielenie. Projekt tworzy Knowledge Hives, jedyna polska firma, która wystąpiła podczas SemTech 2011 – międzynarodowej konferencji poświęconej technikom semantycznym.

Ze względu na źródło szukania odpowiedzi, istnieją dwa typy wyszukiwarek semantycznych: wyszukiwarki analizujące znaczenie indeksowanych dokumentów (Hakia, Bing, Google Squared) oraz wyszukiwarki przeszukujące istniejące zasoby sieci semantycznej (np. Swoogle, Sindice, Falcons, Watson).

Wyszukiwarki analizujące znaczenie przeszukują zawartość stron WWW i starają się zinterpretować (zrozumieć) ich treść poprzez semantyczną i gramatyczną analizę języka naturalnego dokumentów tworzących strony. Przetłumaczenie języka naturalnego na język zrozumiały dla algorytmu jest bardzo trudnym zadaniem, wymagającym zastosowania metod sztucznej inteligencji. Analiza języka naturalnego dokumentów, ale także zapytań do wyszukiwarki, musi obejmować wieloznaczność, specyfikę języka itp.

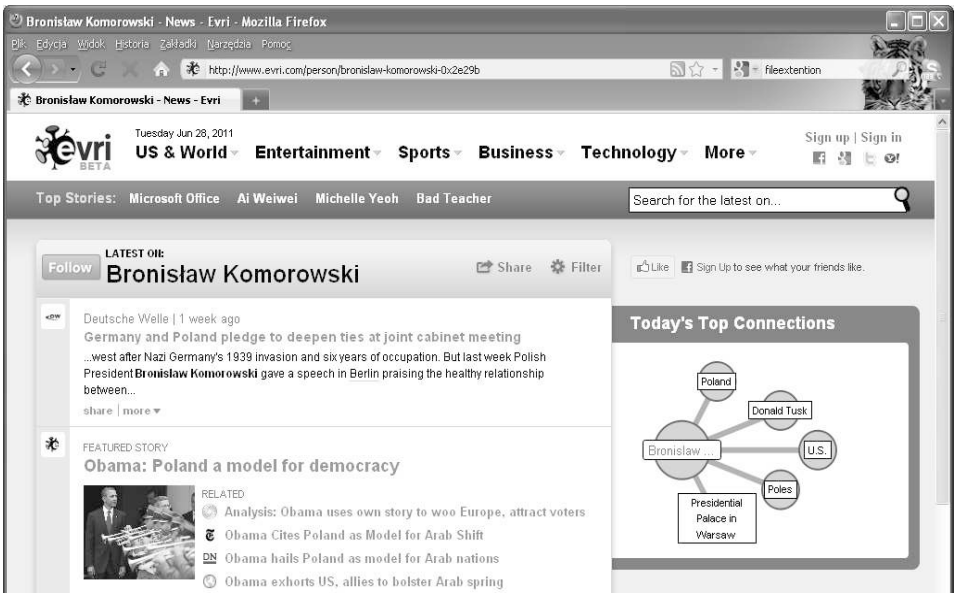
Wyszukiwarki analizujące znaczenie, przeszukując strony WWW tworzą własną bazę ontologii. Uczą się zatem nowych pojęć i relacji między nimi, czyli budują wiedzę, dzięki czemu na kolejne zapytania dostarczane są bardziej relewantne i odpowiednie do zapytań wyniki. KtoCo korzysta ze zbudowanej bazy ontologii, czyli bazy wiedzy zawierającej w momencie uruchomienia serwisu w 2009 r. ponad 800 tysięcy faktów i uwzględniającej powiązania pomiędzy nimi[11]. Kngine obecnie posiada ponad miliard pojęć.

Bardzo często źródłem wiedzy dla wielu tego typu aplikacji jest różnie oceniana, jeśli chodzi o wiarygodność, Wikipedia. Głównie na niej opierają się wyniki Bing, Google Squared i Hakii. Wynika to z otwartości zasobów Wikipedii i Dbpedii, dzięki czemu inne projekty mogą z nich korzystać. Większość wyszukiwarek zdobywa wiedzę z portali, z których jest to łatwe i legalne. Hippisek buduje bazę wiedzy głównie w oparciu o serwisy TVN24 oraz Pudelek, a ostatnio także blogu mBanku, strony hacking.pl i kilku innych. W przyszłości, gdy na takich źródłach aplikacje zaczną działać poprawnie, dołączane będą zapewne bardziej wiarygodne i uznane portale.

Wyszukiwarki przeszukujące zawartość sieci semantycznej przeszukują opisy dokonane przez twórców dokumentów i odwołania do ontologii wskazanych w nagłówkach plików RDF. Przeszukują zatem reprezentację semantyczną dokumentu i nie muszą tłumaczyć sobie jego treści. Pozwolą więc lepiej wyselekcjonować dokumenty zawierające odpowiedź na zapytanie, dzięki czemu na liście wyników użytkownik nie dostanie odnośników do plików, które będą zupełnie bezwartościowe. Jedną z przyczyn niezbyt poprawnego działania wyszukiwarek przeszukujących sieć semantyczną jest ubóstwo jej zawartości, choć jak pokazano w tab. 1, w sieci rośnie liczba dokumentów zgodnych z nową

[11] Zwiastun Web 3.0? Pierwsza polska wyszukiwarka semantyczna, http://webinside.pl/news-5831-zwiastun-web-3-0--pierwsza-polska-wyszukiwarka-semantyczna.html

koncepcją. Najlepiej działającymi są Sindice (12 miliardów ontologii)[12], SWSE (miliard obiektów w maju 2010 r.), Falcon, Watson i Swoogle (10 tys. ontologii w 2007 r.). Wyniki, jakie z nich uzyskamy, są mało czytelne dla ludzi, bo zawierają linki do dokumentów RDF lub OWL i dedykowane są dla algorytmów np. agentów. Projekt Watson[13], nazywany przez jego twórców Semantic Web Gateway, wyróżnia się dużą liczbą aplikacji współpracujących z podstawowym programem. Mają one czytelne interfejsy. Przykładem może być system Scarlet (http://scarlet.open.ac.uk/) wskazujący relacje między podanymi pojęciami i obliczający, w ilu ontologiach razem występują.



Rysunek 2. Zrzut ekranowy wyszukiwarki Evri, 28.06.2011

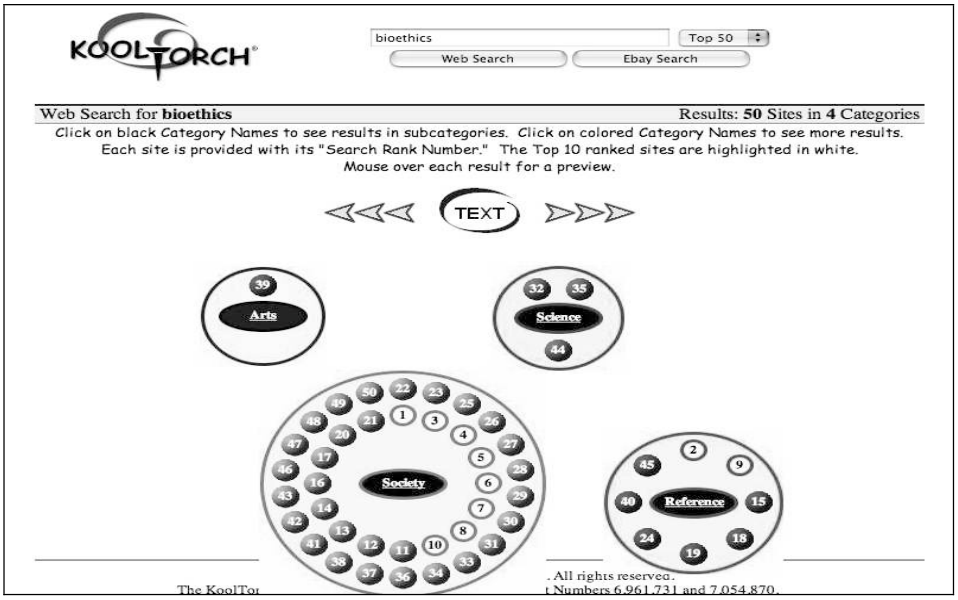
Większość wyszukiwarek semantycznych dokonuje kategoryzacji lub klastryzacji wyników wyszukania, czyli wyświetlają linki podzielone na rożne kategorie. Hakia podaje posegregowane linki w grupach Web, News, Blogs, Credible Sources, Video oraz Images. Kategoryzacji wyników dokonuje także wyszukiwarka Yebol, Kngine, Kosmix. Evri może filtrować wyniki i pokazać

[12] E. Oren et al., *Sindice. com: a document-oriented lookup index for open linked data,* „International Journal of Metadata, Semantics and Ontologies" 2008, vol. 3, issue 1.

[13] M. d'Aquin et al., *What can be done with the Semantic Web? An Overview of Watson-based Applications* [2008], http://people.kmi.open.ac.uk/mathieu/papers/swap.pdf (odczyt 10.12.2011).

tylko wybrane kategorie (Articles, Quotes, Images i Tweets), ponadto rysuje mapę pojęć, np. w postaci drzewa, co widać na rys. 2.

Pewne wyszukiwarki nie wyświetlają w odpowiedzi adresów stron powiązanych z zapytaniem, a podają odpowiedź na zapytanie. Tak działa Wolfram Alpha. Udostępnia tylko konkretne dane, które można zapisać w formacie PDF. Podobnie KtoCo udziela precyzyjnych odpowiedzi na zapytania poprzez wyszukiwanie cytatów ze stron – jednak nie ukrywa ich adresów. Inne wyszukiwarki semantyczne wyświetlają tabele (Google Squared) lub poklastrowane symbole graficzne (KOoLTORCH). Na rys. 3 pokazano przykładowy efekt wyszukania, w którym dopiero po kliknięciu w odpowiednim miejscu pojawiają się linki do źródeł informacji. Google Squared pokazuje wyniki (w tym zdjęcia) w tabeli, którą można wyeksportować do formatu CSV lub arkusza kalkulacyjnego Google. Wskazując jej komórki można zobaczyć źródła zdobytych danych. Można też podawać swoje propozycje lepszych odpowiedzi.



Rysunek 3. Poklastrowane wyniki KoolTorch

Źródło: *Visual & Clustering Search Engines*, http://www.lib.umich.edu/files/visualsearch.pdf

Wszystkie projekty budowania wyszukiwarek semantycznych są potencjalnymi konkurentami tradycyjnych narzędzi, bardzo często ich premiery reklamowane były w taki właśnie sposób, np. Bing Microsoft w maju 2009 r. Każda nowa aplikacja może stanowić zagrożenie, w szczególności dla najpopularniejszej na świecie wyszukiwarki Google. Zapewne dlatego Google już w 2003 r. zaczęło indeksować dokumenty RDF, a w czerwcu 2009 r. wystartowała

semantyczna wyszukiwarka Google Squared. O semantycznych technikach myślą też wyszukiwarki bardziej popularne lokalnie, np. Yandex w Rosji[14] czy Baidu w Chinach.

Tabela 2. Projekty tworzenia wyszukiwarek semantycznych

Nazwa	Rok i kraj powstania	Autor	Finansowanie	Stadium rozwoju projektu
1	2	3	4	5
TextWise	1994, USA	Connie Kenneally	Do 2005 rząd USA, od 2005 TextWise, LLC	rozwój
Hakia	2004, USA	Riza Berkan, Professor Victor Raskin	Prywatne instytucje, anioły biznesu (m.in. R. Krauze)	rozwój
Swoogle	2004, USA	Li Ding, Professor Tim Finin,	DARPA, NSF	skończony, ale strona aktywna
Powerset	2005, USA	Steve Newcomb, Lorenzo Thione, Barney Pell	W 2008 r. kupiona przez Microsoft	kontynuowana jako Bing
Semantifi	2005, USA	Shree Pragada, Vishy Dasari	ExeCue	beta
True Knowledge	sierpień 2005, UK	William Tunstall-Pedoe,	True Knowledge Ltd.,	rozwój
Evri	lipiec 2007, USA	Will Hunsinger	Vulcan Capital	rozwój
Sindice	2007, Irlandia	Renaud Delbru, Giovanni Tummarel-lo, Eyal Oren	Sindice Ltd, na początku UE	rozwój
Watson	2007, UK	Mathieu d'Aquin, Marta Sabou , Enrico Motta	Open University w Milton Keynes. Komisja Europejska	rozwój
KO-oLTORCH	2007, USA	Randy Smith	KOoLTORCH LLC	nieaktywny
Chai Labs	2007, USA	Gokul Rajaram (dawniej Google)	W VII 2010 r. kupiona przez Facebook	wczesna faza rozwoju
Truevert	2008, USA	Arnaud Viviers	OrcaTec	beta
DuckDuckGo	2008, USA	Gabriel Weinberg,	DuckDuckGo, Inc.	rozwój
Yebol	2008, USA	Dr Hongfeng Yin.	anioł	beta
Kngine	2008, USA	Ashrafa i Haythama ElFadeel,.	poszukiwany inwestor	rozwój
Bing	2009, USA	Stefan Weitz	Microsoft	rozwój
Google Squared	czerwiec 2009, USA	Marissa Mayer	Google	rozwój
Wolfram Alpha	maj 2009, USA	Stephen Wolfram	Wolfram Research	rozwój

[14] S. O'Hear, *Russian search engine Yandex gets a semantic injection* [2010], http://eu.techcrunch.com/2010/12/15/russian-search-engine-yandex-gets-a-semantic-injection (odczyt 10.12.2011).

Tabela 2 (cd.)

1	2	3	4	5
Falcon	2009, Chiny	prof. Yuzhong Qu, Wei Hu, Gong Cheng	Websoft Research Group, Nanjing University,	rozwój
KtoCo	2009, Polska	Maciej Stanusch	Stanusch Technologies	beta
Hipisek	2011, Polska	Marcin Walas	POLENG sp. zoo	rozwój

Źródło: Strony WWW projektów, The Free Tech Company Database http://www.crunchbase.com (odczyt 10.12.2011).

Wiele projektów tworzenia wyszukiwarek semantycznych to często prace naukowe, np. Swoogle było przedmiotem rozprawy doktorskiej obronionej na Uniwersytecie Johnsa Hopkinsa w Baltimore, a Hipisek, jest rozwinięciem tematu pracy magisterskiej przygotowanej na Uniwersytecie im. Adama Mickiewicza w Poznaniu. Pierwsze prace badawcze były finansowane przez instytucje naukowo-badawcze, np. Swoogle przez DARPA (*Defense Advanced Research Projects Agency*) i NSF (*National Science Foundation*), a wiele europejskich projektów, jak Watson czy Sindance finansuje lub finansowała Unia Europejska. Z wielu inicjatyw akademickich powstały tzw. spółki odpryskowe (*spinoff*) jak TextWise, która w latach 1994–2005 działała przy inkubatorze Syracuse University i była finansowana przez rząd USA. Wiele projektów to jednak przedsięwzięcia typu StartUp finansowane przez ich twórców, fundusze zalążkowe (*seed capital*) lub *venture capital,* czyli fundusze inwestujące w małe i średnie przedsiębiorstwa wchodzące na rynek, albo wspierane przez tzw. anioły biznesu. Niektórzy inicjatorzy takich przedsięwzięć sami stają się sponsorami – jak Gabriel Weinberg, który po założeniu DuckDuckGo zaczął inwestować w inne startujące projekty. Pewne inicjatywy StartUp, jak np. Kngine, finansowane były na początku przez ich twórców, ale obecnie szukają inwestora. Bywa, że młode przedsiębiorstwa są przejmowane przez gigantów na rynku, np. Powerset w 2008 r. został kupiony przez Microsoft i kontynuowany jako nowy produkt Bing. Innym przykładem jest Chai Labs kupiony w 2010 r. przez Facebook. Omówione powyżej przykłady zebrano w tab. 2. Jak widać, większość z nich powstaje w USA i jest w fazie rozwoju lub *beta*, co oznacza, że narzędzia te funkcjonują w wersji testowej lub jako prototypy projektów. Niestety większość nie działa jeszcze poprawnie, co pokazało badanie przeprowadzone w grudniu 2010 r.[15] Powtórzono je w czerwcu 2011 r. i nadal rozważane wyszukiwarki nie dały lepszych wyników. W tym przypadku *lepsze* oznacza brak lub mniejszą liczbę linków z Wikipedii lub ich dalszą pozycję. W wielu

[15] B. Gontar, J. Papińska-Kacperek, *Wyszukiwarki…*

rankingach najlepiej oceniana jest wyszukiwarka Wolfram Alpha, co potwierdziły oba badania.

Większość przedsięwzięć nie generuje jeszcze zysku, ale inwestorzy liczą, że nie będzie tak zawsze i np. budowane obecnie Hakia czy Evri, kiedy zaistnieje Web 3.0, mogą okazać się bardzo dochodowe. Pewne wyszukiwarki, jak Bing, tworzone i finansowane przez znane firmy, potencjalnie najszybciej mogą zakłócić ustalony rynku wyszukiwarek internetowych porządek. Według badań ComScore[16], w marcu 2011 r. wyszukiwarka Bing miała już 14% udziałów w amerykańskim rynku wyszukiwania, na drugim miejscu był Yahoo – 16%, a na pierwszym Google z odsetkiem 65%. W Polsce, w drugiej połowie czerwca 2011 r. według rankingu gemiusRanking PL – Bing także zajmowała trzecie miejsce, ale z wynikiem 1,20%, Onet 1,32%, a Google na pierwszym miejscu z wynikiem 95%.

5. PODSUMOWANIE

Choć wzrasta liczba dokumentów zgodnych z zaproponowaną nową strukturą sieci semantycznej i powstaje dużo nowych wyszukiwarek semantycznych, to jednak nie są one na razie w stanie zagrozić obecnie używanym popularnym narzędziom. Efekty ich działania mogą obecnie nie satysfakcjonować, ale pamiętać należy, że większość testuje swoje możliwości na źródłach dostępnych dla małych niezamożnych firm, które takie aplikacje tworzą. Z tego powodu większość wyszukiwarek semantycznych korzysta z otwartych zasobów, często z różnie ocenianej Wikipedii. Z roku na rok ich algorytmy budują kolejne ontologie i w ten sposób ich baza wiedzy staje się bogatsza. Sieć ucząca się, jaką staje się właśnie Internet, powinna przekonać do swoich możliwości wszystkich niezadowolonych wynikami wyszukań popularnych obecnie narzędzi. Zatem wydaje się pewne, że wkrótce aplikacje semantyczne mogą zainteresować większą liczbę użytkowników, czego dowodem może być fakt, że silne na rynku Google nie tylko indeksuje dokumenty RDF czy OWL, ale inwestuje w swoje semantyczne aplikacje jak Google Squared.

BIBLIOGRAFIA

Boulton, C. *Google Keeps 65% Search, Bing Tops 14%: comScore* [2011], http://www.eweek.com/c/a/Search-Engines/Google-Keeps-65-Search-Bing-Tops-14-comScore-792394 (odczyt 10.12.2011).

Ding, L. et al., *Swoogle: A Search and Metadata Engine for the Semantic Web*, CIKM 2004.

[16] C. Boulton, *Google Keeps 65% Search, Bing Tops 14%: comScore* [2011], http://www.eweek.com/c/a/Search-Engines/Google-Keeps-65-Search-Bing-Tops-14-comScore-792394 (odczyt 10.12.2011).

d'Aquin M. et al., *What Can be Done with the Semantic Web? An Overview of Watson-based Applications* [2008], http://people.kmi.open.ac.uk/mathieu/papers/swap.pdf (odczyt 10.12.2011).

Gliński, W., *Ontologie. Próba uporządkowania terminologicznego chaosu*, [w:] B. Sosińska-Kalata i in. (red.), *Od informacji naukowej do technologii społeczeństwa informacyjnego,* Miscellanea Informatologica Varsoviensia, Wydawnictwo SBP, Warszawa 2005.

Gliński, W., *Języki i narzędzia do tworzenia i wyszukiwania ontologii w kontekście semantycznego Weba,* [w:] B. Sosińska-Kalata i in. (red.), *Od informacji naukowej do technologii społeczeństwa informacyjnego,* Miscellanea Informatologica Varsoviensia, Wydawnictwo SBP, Warszawa 2005.

Gontar B., Papińska-Kacperek J., *Wyszukiwarki semantyczne*, [w:] M. Pańkowska (red.), *Wiedza i komunikacja w innowacyjnych organizacjach. Komunikacja elektroniczna*, Wydawnictwo UE, Katowice 2011.

O'Hear S., *Russian Search Engine Yandex Gets a Semantic Injection* [2010], http://eu.techcrunch.com/2010/12/15/russian-search-engine-yandex-gets-a-semantic-injection (odczyt 10.12.2011).

Oren E. et al., *Sindice.com: A Document-oriented Lookup Index for Open Linked Data,* „International Journal of Metadata, Semantics and Ontologies" 2008, vol. 3, issue 1.

Papińska-Kacperek J., *Nowa epoka – społeczeństwo informacyjne,* [w:] J. Papińska-Kacperek (red.), *Społeczeństwo informacyjne*, Wydawnictwo Naukowe PWN, Warszawa 2008.

Papińska-Kacperek J., *Przykłady zastosowań serwisów społecznościowych*, „Zeszyty Naukowe Uniwersytetu Szczecińskiego" 656/2011, Studia Informatica 28.

Pawełczyk M., *Informacja a niepewność*, materiały do zajęć [2003], http://marpaw.elisa.pl/wsti/roznosci/pomiar_inform/inform.htm (odczyt 10.12.2011]).

Spivak N. *How the WebOS Evolves?* [2007], http://novaspivack.typepad.com/nova_spivacks_weblog/2007/02/steps_towards_a.html (odczyt 10.12.11).

Segaran T., Evans C., Taylor J., *Programming the Semantic Web*, O'Reilly Media 2009

Stefanowicz B., *Informacja*, Wydawnictwo SGH, Warszawa 2004.

Zawiła-Niedźwiecki J., Rostek K., Gąsiorkiewicz A., *Informatyka gospodarcza*, C.H. Beck, Warszawa 2010.

Strony wyszukiwarek semantycznych

Bing http://www.bing.com
DuckDuckGo http://duckduckgo.com
Evri http://www.evri.com
Falcon http://ws.nju.edu.cn/falcons/objectsearch/index.jsp
Google Squared http://www.google.com/squared
Hakia http://www.hakia.com
Hipisek http://mwalasvm.vm.wmi.amu.edu.pl/~walasiek/hipisek
Kngine, http://www.kngine.com
KtoCo http://www.ktoco.pl
Semantifi http://www.semantifi.com
Sindice http://sindice.com.
Swoogle http://swoogle.umbc.edu,
SWSE http://swse.deri.org
TextWise http://textwise.com
True Knowledge http://www.trueknowledge.com
Truevert http://www.truevert.com
Watson http://kmi-web05.open.ac.uk/WatsonWUI
Wolfram Alpha http://www.wolframalpha.com,
Yebol http://yebol.com

Beata Gontar, *Joanna Papińska-Kacperek*

SEMANTIC SEARCH ENGINE

Paper presents some basic issues on Semantic Web and a semantic search engines market. It is presented growing popularity of Web 3.0 showing many projects, mainly in the U.S., of building and using Semantic web, but also present examples that Web 3.0 still does not work well. Ontology and languages (RDF, OWL) for building ontologies are important part of Web 3.0.The technologies and the tools are ready, the changes on the market – visible, but there is still much work to do with existing documents in the net.