STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Marcin SKULIMOWSKI
Uniwersytet Łódzki, Wydział Fizyki i Informatyki Stosowanej

ANALIZATOR INTERNETU SEMANTYCZNEGO

Streszczenie. W artykule przedstawiony jest system SWAN (Semantic Web Analyzer) wyszukujący w Internecie informacje zapisane w formatach RDF i OWL, stworzonych na potrzeby Internetu Semantycznego. Informacje znalezione przez system SWAN zapisywane są w relacyjnej bazie danych, co umożliwia ich dalsze przetwarzanie i wykorzystanie.

Słowa kluczowe: Internet Semantyczny, wydobywanie danych

SEMANTIC WEB ANALYZER

Summary. In this article the SWAN system (Semantic Web Analyzer) is presented. This system is designed for searching the Web information stored in RDF and OWL formats developed for the Semantic Web. The data found by SWAN are stored in a relational database, allowing for their further processing and using.

Keywords: Semantic Web, data mining

1. Wprowadzenie

Zawartość obecnie istniejącego Internetu, w szczególności stron WWW, jest przeznaczona dla ludzi i trudno przetwarzalna przez maszyny. Konsekwencją tego są problemy z przetwarzaniem informacji zgromadzonych w Internecie. Z jednej strony ogromna ilość informacji wymusza wykorzystanie maszyn (np. wyszukiwarek), z drugiej jednak możliwości maszyn są raczej ograniczone, przede wszystkim dlatego, że maszyny nie rozumieją przetwarzanej informacji. Przykładem są wyszukiwarki, działające na podstawie porównywania ciągów znaków, które bardzo często zwracają informacje niezwiązane z zapytaniem, a pomijają te istotne. Rozwiązanie zarysowanego powyżej problemu pojawiło się wraz z ideą Internetu Semantycznego [1, 2]. Informacja w Internecie Semantycznym będzie wzbogacona o dane

(nazwijmy je *semantycznymi*) zapisane w językach RDF (ang. *Resource Description Framework*) i OWL (ang. *Ontology Web Language*), umożliwiających ich automatyczne przetwarzanie przez maszyny (agentów). Dane semantyczne (zapisane w języku RDF) mogą zawierać metadane na temat dokumentu (np. strony WWW), a także informacje bezpośrednio związane z treścią dokumentu. Możliwe są przy tym dwa podejścia: dane semantyczne będą umieszczone w dokumencie, którego dotyczą lub będą znajdowały się w oddzielnym dokumencie [3, 4]. Kluczową rolę w Internecie Semantycznym odgrywają ontologie (zapisywane w języku OWL), definiujące formalnie słownictwo wykorzystywane do tworzenia danych semantycznych. Zastosowanie ontologii sprawi, że maszyny będą „rozumiały" przetwarzane informacje, a co ważne będą potrafiły wnioskować, opierając się na tych informacjach. To z kolei przyczyni się do usprawnienia wyszukiwania w Internecie [3]. Maszyny będą np. poszukiwały stron, które odnoszą się do precyzyjnie określonego konceptu w ontologii, a nie stron zawierających pewne, często niejednoznaczne, słowo [5]. Dzięki temu, wyniki poszukiwań będą lepiej odpowiadały zapytaniom.

W obecnej chwili Internet Semantyczny jest dopiero na etapie powstawania. Istnieją już strony internetowe zawierające semantyczne adnotacje lub linki do związanych z nimi dokumentów semantycznych, zapisanych w języku RDF. Na wielu stronach można też znaleźć ontologie związane z różnymi dziedzinami wiedzy (zobacz np. [6]). Niestety obecnie istniejące wyszukiwarki nie są w stanie analizować danych semantycznych, dlatego tworzone są specjalne systemy wyszukujące i/lub przetwarzające informacje, zapisane w Internecie Semantycznym [7-12]. W szczególności systemy takie mogą np. wyszukiwać ontologie, dane w języku RDF czy np. zasoby wybranego typu. Informacje udostępnione przez takie systemy mogą być następnie wykorzystywane przez maszyny (agentów), ale także przez ludzi.

Celem niniejszego artykułu jest zaprezentowanie systemu SWAN (ang. *Semantic Web Analyzer*) przeznaczonego do wyszukiwania i analizy informacji zapisanych w dokumentach semantycznych (Internecie Semantycznym). W artykule omówiona zostanie architektura systemu, jego zastosowania oraz perspektywy rozwoju.

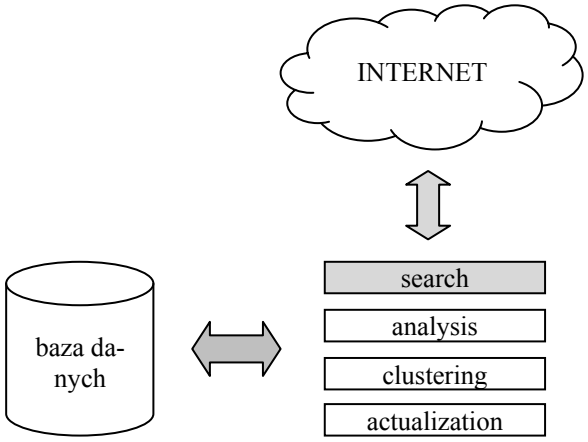
2. Architektura systemu

System SWAN składa się z czterech głównych modułów: *search*, *analysis*, *clustering* i *actualization* oraz bazy danych, stanowiącej podstawę działania całego systemu (rys. 1).

- *Search* – moduł odpowiedzialny za wyszukiwanie dokumentów RDF i OWL oraz danych semantycznych umieszczonych bezpośrednio w kodzie strony WWW. Wyszukiwanie jest realizowane automatycznie z wykorzystaniem robota hybrydowego, tzn. operującego na tradycyjnych dokumentach HTML, połączonych linkami oraz na dokumentach RDF po-

łączonych relacją `rdfs:seeAlso` (tzw. *RDF hyper-linking*) [13, 14]. W działaniu robota wykorzystane są dwie cechy charakterystyczne Internetu [15]:

1. Strony internetowe zawierają linki do stron o podobnej tematyce. Jest to tzw. *lokalność tematyczna* Internetu (ang. *topical locality*).
2. Słowa zawarte w opisach odsyłaczy (ang. *anchor text*) oraz w ich pobliżu związane są z tematyką stron, do których prowadzą odsyłacze.



Rys. 1. Architektura SWAN
Fig. 1. SWAN architecture

Robot wyszukuje i pobiera dane semantyczne umieszczone na stronach internetowych zarówno w przypadku, gdy są one umieszczone w kodzie strony, jak i wtedy, gdy są oddzielnymi dokumentami (z linkami do nich umieszczonymi na stronie). W tym drugim przypadku robot wspiera następujące sposoby dołączania danych semantycznych (zobacz [16]):

1. `<a href="file.rdf">...</a>`
2. `<a rel="meta" type="application/rdf+xml" href="file.rdf">...</a>`
3.
```
<head>
<link rel="meta" type="application/rdf+xml" href="file.rdf"/>
</head>
```
4.
```
<head>
<link rel="file.foaf" href="http://xmlns.com/foaf/0.1/"/>
</head>
```

Robot wyszukuje i pobiera także dane umieszczone w kodzie strony internetowej za pomocą bloku `<script>`:

```
<script type="application/rdf+xml">...</script>
```

oraz dane umieszczone w komentarzach.

Ponadto, robot przetwarza dane RDF dołączone do dokumentów XHTML za pomocą technologii *microformats* [17].

W celu zwiększenia możliwości wyszukiwania danych semantycznych, moduł *Search* wykorzystuje *system robotów równoległych* działających jednocześnie na wielu kompute-

rach, opierając się na jednym centralnym serwerze [18]. Znalezione dane semantyczne są pobierane i przechowywane na serwerze.

- *Analysis* – moduł odpowiedzialny za analizę znalezionych danych semantycznych. Analiza polega na pobraniu z nich *zasobów* i *stwierdzeń języka RDF* oraz na umieszczeniu ich w bazie danych. Każdy *zasób* posiada identyfikator IRI (ang. *Internationalized Resource Identifier*)[1]. Stwierdzenie w języku RDF składa się z *podmiotu*, *orzeczenia* oraz *obiektu*, przy czym podmiot i orzeczenie muszą być zasobami, natomiast obiekt może być zasobem lub literałem [19]. Ponadto, podmiot i obiekt mogą być zasobem nieposiadającym globalnego identyfikatora IRI, a jedynie identyfikator lokalny (tzw. *blank node*). Stwierdzenia RDF można zapisywać w bazie danych na wiele sposobów (zobacz np. [20, 21]). W przypadku systemu SWAN do zapisu stwierdzeń wykorzystywane są trzy tabele związane z trzema rodzajami elementów, wchodzących w skład stwierdzeń:

  resource – tabela zawierająca zasoby,

  literal – tabela zawierająca literały,

  bnode – tabela zawierająca tzw. puste węzły (zasoby bez identyfikatora IRI)

  oraz tabela *triple* zawierająca stwierdzenia. Dodatkowo w tabelach *triple_where* i *resource_where* umieszczone są informacje o stronie internetowej, na której znaleziono dane stwierdzenie i zasób.

- *Clustering* – moduł, którego zadanie polega na klasyfikowaniu zasobów. Klasyfikacja polega na dzieleniu zasobów w klastry oraz na obliczeniu podobieństwa między klastrami. W tym celu brane są pod uwagę *właściwości bezpośrednie* (ang. *immediate properties*) zasobów, tzn. dwa zasoby należą do tego samego klastra, jeżeli posiadają dokładnie te same właściwości (zobacz np. [22]). Każdy klaster jest zatem definiowany przez zbiór właściwości posiadanych przez należące do niego zasoby. Zastosowanie takiej klasyfikacji pozwoliło uzyskać ponad 3.5 tys. klastrów z ponad 800 tys. zasobów[2]. Wykres zależności liczby klastrów od liczby definiujących je właściwości jest przedstawiony na rys. 3. Podobieństwo między klastrami c_i i c_j obliczane jest za pomocą *współczynnika Jaccarda* [23]:

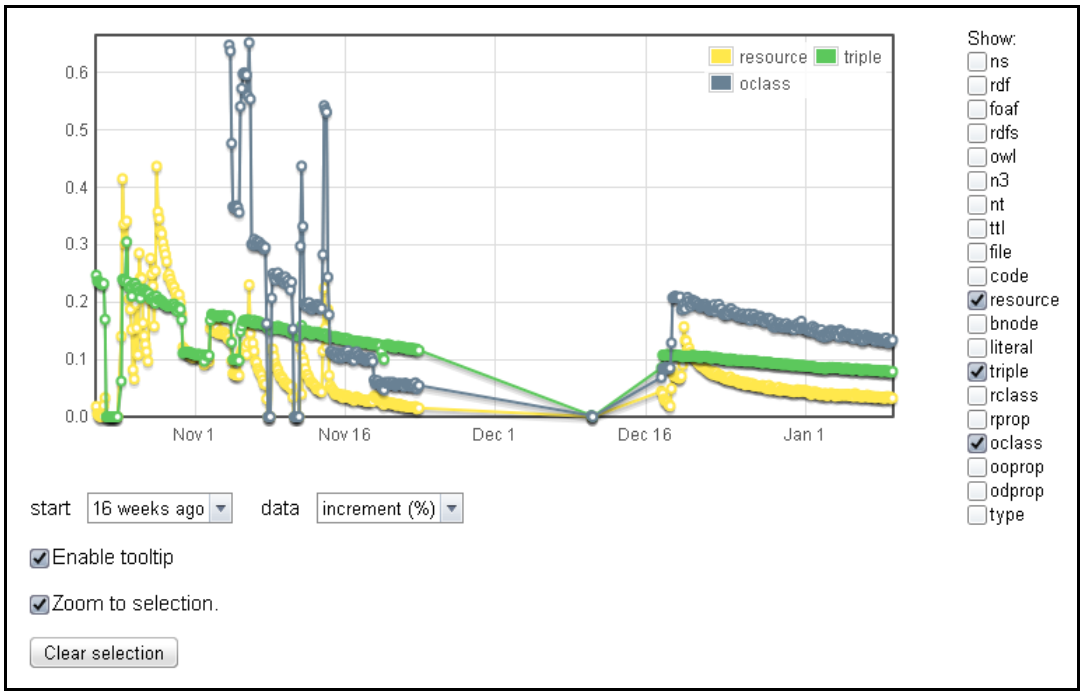
$$sim(c_i, c_j) = \frac{\left| p_i \cap p_j \right|}{\left| p_i \cup p_j \right|},$$

  gdzie p_i jest zbiorem właściwości klastra c_i.

  W obecnej chwili (styczeń 2011) dane zgromadzone w systemie obejmują ponad 9 mln. stwierdzeń w języku RDF oraz ponad 900 tys. zasobów. Rysunek 2 przedstawia wykres

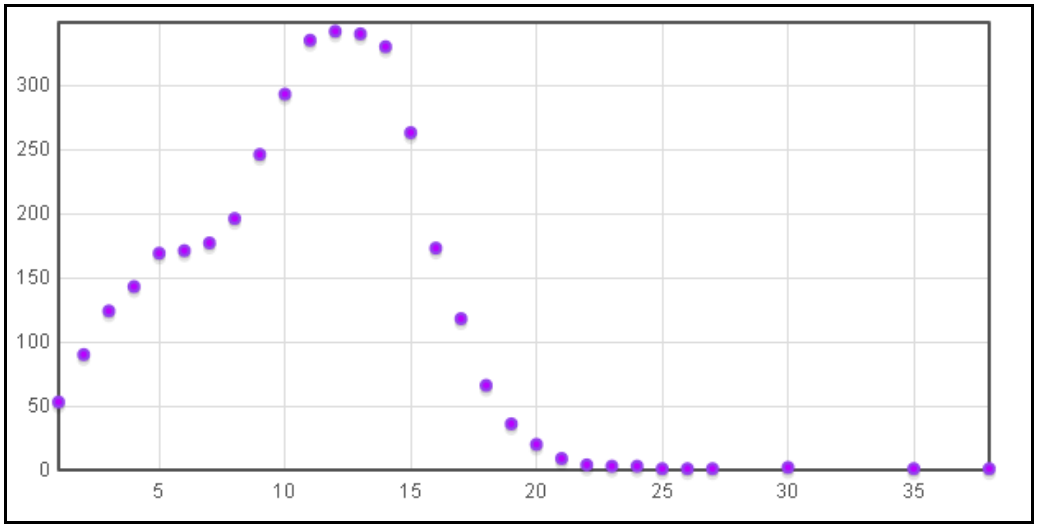
[1] Szczególnym przypadkiem identyfikatora IRI jest URL (ang. *Uniform Resource Locator*).
[2] Dane ze stycznia 2011 r.

miesięcznej aktywności systemu. Najbardziej liczne typy zasobów zgromadzonych w systemie przedstawione są na rys. 4.



Rys. 2. Aktywność systemu SWAN
Fig. 2. SWAN system activity

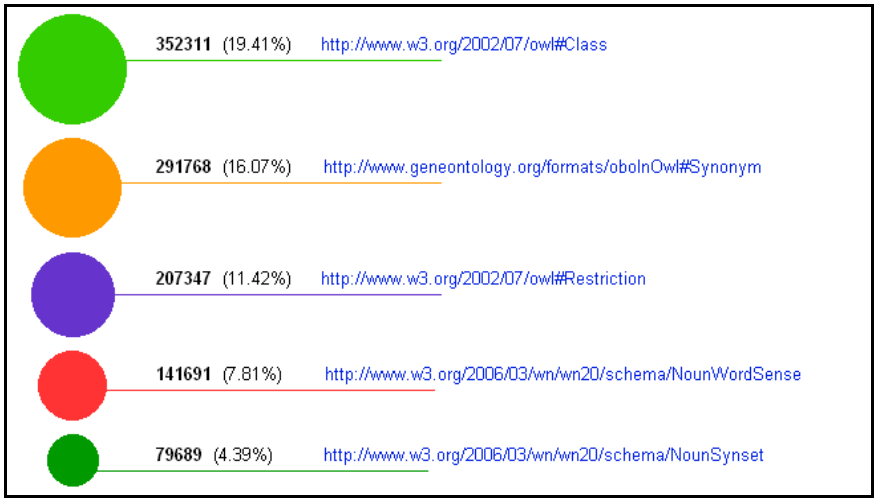
- *Actualization* – moduł odpowiedzialny za aktualizację danych. Aktualizacja polega na sprawdzeniu czy znalezione wcześniej dane semantyczne są nadal aktualne. Jeżeli nie, wówczas moduł automatycznie dokonuje aktualizacji.



Rys. 3. Zależność liczby klastrów (Y) od liczby właściwości (X)
Fig. 3. Dependence of the number of clusters (Y) on the number of properties (X)

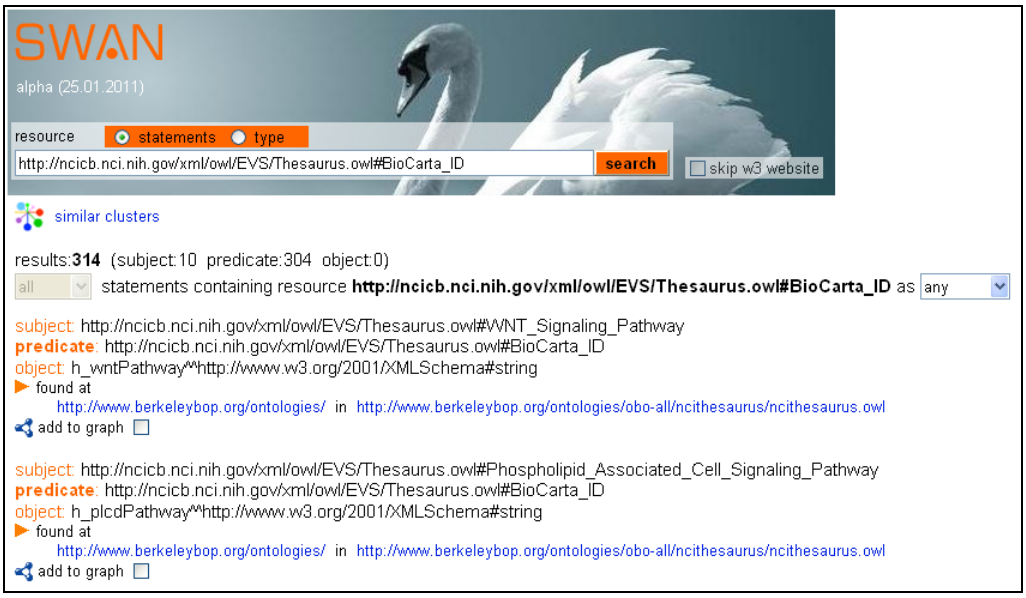
System SWAN stworzony jest z wykorzystaniem technologii Java, a także PHP i MySQL. Interfejs systemu stworzony jest na podstawie technologii JavaScript, AJAX oraz biblioteki FLOT [24]. Ponadto, w systemie wykorzystano:

- *Jena* – środowisko Javy przeznaczone do tworzenia aplikacji w Internecie Semantycznym [25],
- *RDF API for PHP* – biblioteka PHP do przetwarzania dokumentów RDF [26],
- *Graphviz* – środowisko tworzenia grafów [27].



Rys. 4. Typy zasobów
Fig. 4. Types of resources

3. Zastosowania

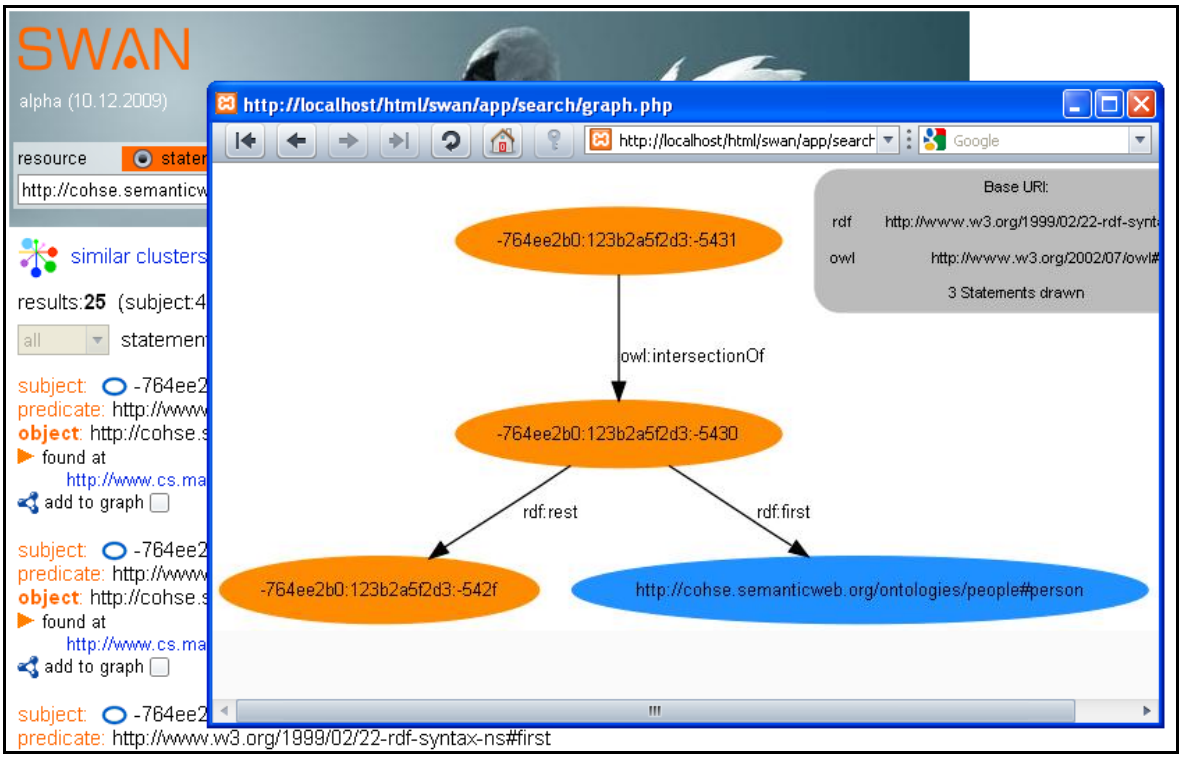


Rys. 5. Interfejs wyszukiwarki zasobów
Fig. 5. Interface of the resources search engine

Dane semantyczne (dokumenty RDF oraz ontologie zapisane w języku OWL), znalezione i przeanalizowane przez system SWAN, mogą być wykorzystana na wiele sposobów. Aktualnie zrealizowane są następujące możliwości:

- *Wyszukiwarka zasobów*

System SWAN oferuje możliwość wyszukiwania stwierdzeń zawierających zasób o określonym identyfikatorze URI. Po wprowadzeniu w pole formularza (rys. 5) dowolnego ciągu znaków, system sugeruje identyfikatory URI zasobów istniejących w bazie, zawierających ten ciąg znaków. Po wybraniu zasobu, system zwraca stwierdzenia zawierające wprowadzony zasób. Na rys. 5 przedstawione są wyniki wyszukiwania stwierdzeń w wersji przeznaczonej dla odbiorcy ludzkiego. Rezultaty wyszukiwania mogą mieć także postać dokumentu RDF, zawierającego znalezione stwierdzenia. Dla wybranych przez użytkownika stwierdzeń system umożliwia wygenerowanie związanego z nimi grafu (rys. 6).



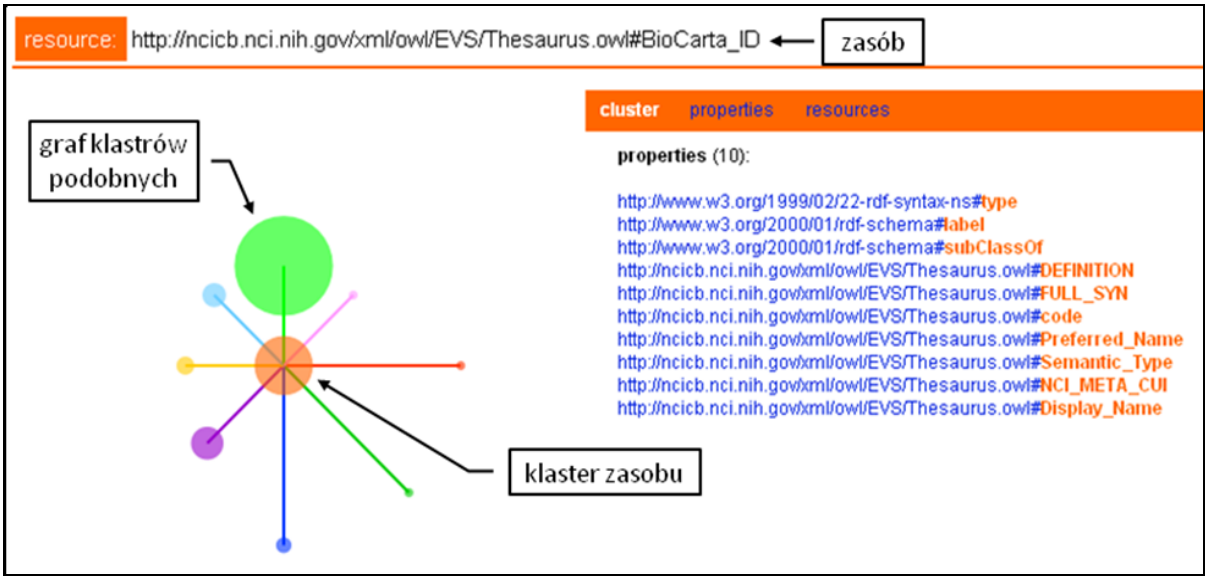
Rys. 6. Graf RDF odpowiadający wybranym zasobom
Fig. 6. RDF graph corresponding to selected triples

System SWAN umożliwia także wyszukiwanie zasobów określonego typu (np. tego samego typu, co wybrany zasób). Ponadto, dla wybranego zasobu, możliwe jest wyszukiwanie stron WWW, zawierających zasoby tego samego typu (rys. 7).

Rys. 7. Wyniki wyszukiwania
Fig. 7. Results of search

- *Podobieństwo zasobów*

Dla wybranego zasobu system dostarcza informacje (m.in. w formie grafu) na temat klastrów, które są najbardziej podobne do klastra zawierającego dany zasób (rys. 8). Wybór (na grafie – rys. 8) dowolnego klastra powoduje wyświetlenie identyfikatorów URI zasobów do niego należących bądź właściwości opisujących dany klaster.

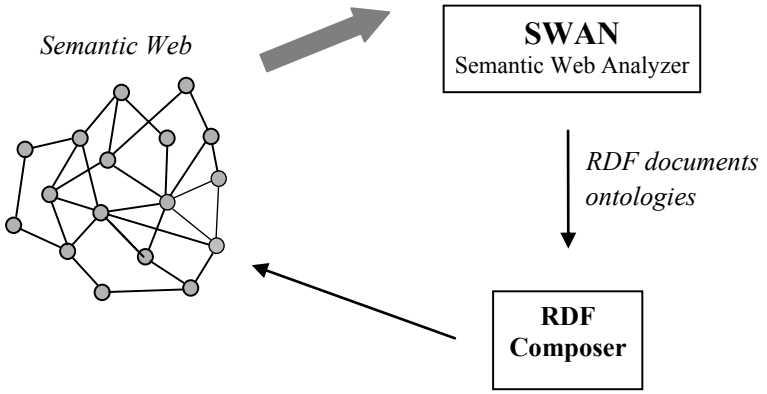


Rys. 8. Podobieństwo klastrów
Fig. 8. Similarity of clusters

4. Perspektywy

System SWAN jest w fazie rozwoju. Aktualnie trwają prace nad następującymi elementami systemu:

- *Search*
  Istnieje wiele sposobów „osadzania" stwierdzeń języka RDF w kodzie strony WWW. Moduł *search* aktualnie jest rozbudowywany o możliwość analizy stwierdzeń umieszczonych w dokumentach XHTML za pomocą technik eRDF i RDFa (zobacz np. [2, 4]).
- *Inference*
  Wnioskowanie może być przeprowadzane na podstawie stwierdzenia języka RDF i ontologii. Istnieje wiele reguł wnioskowania (zobacz np. [28]); w obecnej chwili, w systemie SWAN testowana jest możliwość realizacji wnioskowania na podstawie schematów RDF, za pomocą procedur w bazie danych.
- *Browser*
  Istnieją już narzędzia pozwalające wykorzystywać dane semantyczne w czasie przeglądania stron internetowych (zobacz np. [29, 30]). Dane dostępne w systemie SWAN mogą być wykorzystywane przez przeglądarki internetowe, np. w sugerowaniu stron o podobnej tematyce czy ontologii, związanych z tematyką przeglądanej strony internetowej. W związku z tym, aktualnie trwają prace nad dodatkiem do przeglądarki Firefox, wykorzystującym dane pobierane z systemu SWAN.
- *Integracja z Kompozytorem RDF*
  Kompozytor RDF to aplikacja webowa, wspomagająca tworzenie dokumentów RDF na podstawie istniejących schematów RDF i ontologii [31]. Planowana jest integracja systemu SWAN z kompozytorem RDF (rys. 9). Dzięki temu, w trakcie tworzenia dokumentów RDF możliwe będzie wykorzystanie słownictwa z ontologii, znajdującego się w systemie SWAN.



Rys. 9. Kompozytor RDF i SWAN [31]
Fig. 9. RDF Composer and SWAN [31]

W 2011 roku planowane jest częściowe udostępnienie systemu SWAN w Internecie.

BIBLIOGRAFIA

1. Berners-Lee T., Hendler J., Lassila O.: The Semantic Web. Scientific American, Vol. 284, No. 5, s. 34÷43.
2. Hebeler J., Fisher M., Blace R., Perez-Lopez A.: Semantic Web Programming, Wiley, 2009.
3. Finin T., Mayfield J., Fink C., Joshi A., Cost R. S.: Information Retrieval and the Semantic Web, Proceedings of the 38th International Conference on System Sciences, 2005.
4. Świątkiewicz M.: Dane sieci semantycznej w dzisiejszej sieci WWW. Studia Informatica, Vol. 28, No. 2 (71), Wyd. Pol. Śląskiej, Gliwice 2007, s. 47÷66.
5. Antoniou G., Harmelen F.: Semantic Web Primer, MIT Press, 2004.
6. http://semanticweb.org/wiki/Ontology (2011.01.26).
7. Finn T., et al.: Swoogle: Searching for knowledge on the Semantic Web. Proceedings of AAAI 05 (intelligent systems demo), July 2005.
8. Guha R., McCool R., Miller E.: Semantic search, WWW '03 Proceedings of the 12th international conference on World Wide Web, ACM New York 2003, s. 700÷709.
9. Scheir P., Pammer V., Lindstaedt S. N.: Information Retrieval on the Semantic Web – Does it exists? Hinneburg A. (Ed.): LWA 2007: Lernen – Wissen – Adaption, Halle, September 2007, Workshop Proceedings, Halle-Wittenberg 2007, s. 252÷257.
10. Rocha C., Schwabe D., Aragao M. P.: A hybrid approach for searching in the semantic Web, Proceedings of the 13th international conference on World Wide Web, ACM New York 2004, s. 374÷383.
11. Davies J., Weeks R.: QuizRDF: Search Technology for the Semantic Web, Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04), 2004.
12. Fazzinga B., Lukasiewicz T.: Semantic search on the Web, Semantic Web – Interoperability, Usability, Applicability 1, 2010, s. 1÷7.
13. Dodds L.: Slug: A Semantic Web Crawler, http://slug-semweb-crawler.googlecode.com-/files/slug-a-semantic-web-crawler.pdf.
14. Brickley D.: RDF Hyper-linking, http://www.w3.org/2001/sw/Europe/talks/xml2003/-Over-view.html (2011.01.26).
15. Davison B. D.: Topical locality in the web. Proc. 23rd Annual Intl. ACM SIGIR Conf. on Research and Development In Information Retrieval, Athens 2000, s. 272÷279.
16. Palmer S. B.: RDF in HTML: Approaches, http://infomesh.net/2002/rdfinhtml/ (2011.01.26).
17. Allsopp J.: Microformats: Empowering Your Markup for Web 2.0, Apress, 2007.
18. Nowak Ł.: System robotów równoległych przeszukujących strony WWW, praca magisterska, Wydział Fizyki i Informatyki Stosowanej UŁ, 2010.

19. Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ (2011.01.26).
20. Melnik S.: Storing RDF in a relational database,
21. http://infolab.stanford.edu/~melnik/rdf/-db.html (2011.01.26).
22. Wolf B.: Storing RDF Metadata in a Relational Database, http://www.kbs.uni-hannover.de/Arbeiten/Studienarbeiten/01/Wolf/olr_documentation.pdf (2011.01.26).
23. Grimnes G. A., Edwards P., Preece A.: Instance Based Clustering of Semantic Web Resources, Lecture Notes in Computer Science, 2008, Vol. 5021/2008, s. 303÷317.
24. Van Rijsbergen C. J.: Information retrieval, http://www.dcs.gla.ac.uk/Keith/Preface.html (2011.01.26).
25. Flot – Javascript plotting for jQuery, http://code.google.com/p/flot/ (2011.01.26).
26. Jena – A Semantic Web Framework for Java, http://jena.sourceforge.net/ (2011.01.26).
27. RAP – RDF API for PHP, http://www4.wiwiss.fu-berlin.de/bizer/rdfapi/ (2011.01.26).
28. Graphviz – Graph Visualization Software, http://www.graphviz.org/ (2011.01.26).
29. RDF Semantics, W3C Recommendation, http://www.w3.org/TR/rdf-mt/ (2011.01.26).
30. Dzbor M., Mottaa E., Dominguea J.: Magpie: Experiences in supporting Semantic Web browsing, Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 5, Issue 3, September 2007, s. 204÷222.
31. Huynha D., Mazzocchib S., Kargera D.: Piggy Bank: Experience the Semantic Web inside your web browser, Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 5, Issue 1, March 2007, s. 16÷27.
32. Kopczyński P., Skulimowski M.: RDF Composer – an application supporting creation and editing of RDF documents, Annales UMCS Informatica AI XI, 2 (2010), s. 39÷47.

Recenzent: Dr inż. Michał Świderski

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

The Semantic Web is planned as an extension of the current Web in which information will be processable by machines due to an application of a machine understandable markup. Nowadays there already exist web pages containing semantic annotations or links to semantic documents. The problem is, however, that popular search engines are not able to analyze semantic data. They have been developed to process text documents and the semantic content is

invisible to them. So there is an increasing need for systems designed for finding and analyzing knowledge encoded in the semantic data. SWAN (Semantic Web Analyzer), which is described in the paper, is an example of such system. The architecture of the system, its applications and perspectives for development are presented.

Adres

Marcin SKULIMOWSKI: Uniwersytet Łódzki, Wydział Fizyki i Informatyki Stosowanej, ul. Pomorska 149/153, 90-236 Łódź, Polska, mskulim@uni.lodz.pl.