Przedmowa

Dopóki nie skorzystałem z Internetu, nie wiedziałem, że na świecie jest tylu idiotów.
Stanisław Lem (1994)

Na progu drugiej dekady XXI wieku chyba nikt już nie jest w stanie wyobrazić sobie życia bez Internetu, którego użytkowanie staje się powoli standardem i powszednieje – tak jak oglądanie telewizji czy posługiwanie się telefonem komórkowym. Najbardziej charakterystyczną cechą tego medium jest jego *egalitarność*, gdyż wszyscy jego użytkownicy mają praktycznie dostęp do wszystkiego. W Sieci szukamy odpowiedzi na nurtujące nas pytania, reklamujemy usługi i produkty naszej firmy, przeprowadzamy transakcje, publikujemy wyniki swojej pracy, prowadzimy polemiki, komentujemy wydarzenia, nawiązujemy kontakty, organizujemy grupy ludzi do podejmowania różnych działań. To wielka zaleta, ale zarazem też wielka wada tego medium. Z jednej strony zawarte tam treści tworzą ważny i nieustannie powiększający się zasób cywilizacyjny ludzkości, z drugiej strony stanowią modelowy przykład anarchii. Wszelkie próby jej opanowania są z góry skazane na niepowodzenie – nie tylko dlatego, że Sieć to pewien standard demokratyczny, rozumiany jako swobodny dostęp do wolnej informacji która, jak pokazuje historia, wcześniej czy później zawsze potrafi pokonać opresję. Zdecydowanie bardziej decydującym czynnikiem jest tutaj dynamika przyrostu zasobów informacyjnych. W roku 1990 liczbę stron internetowych szacowano na ok. 220 milionów, obecnie w roku 2011 tę liczbę ocenia się na 50 miliardów. Wziąwszy pod uwagę, że w zasadzie nie ma jednej uznanej metodologii szacowania wielkości Sieci (np. jak liczyć archiwalne kopie stron czy, strony generowane dynamicznie – Google, Bing, Yahoo i Ask podają rozbieżne dane), należałoby przyjąć, że jest to zasób praktycznie nieskończony.

Pojawia się zatem pytanie: skoro publikowanych w nim treści nie da się ani ograniczyć, ani zorganizować, to jak docierać do stron, które zawierają interesującą nas informację? Nawiązując do motta Stanisława Lema zacytowanego na wstępie, który obok innych wielkich pisarzy s-f snuł wizje globalnego systemu informacyjnego (choćby w „Obłoku Magellana" z 1955 roku), treści „idiotyczne" dla jednych odbiorców mogą mieć wartość dla innych. Użytkownicy Sieci mogą szukać określonych informacji nie tylko w celu prowadzenia badań, analizy i przewidywania trendów, śledzenia wydarzeń, ale także monitorowania zagrożeń, np. gdyby publikowane treści miały stanowić zagrożenie dla innych użytkowników Sieci.

Przedstawiamy Czytelnikowi książkę, która dokumentuje wyniki ponad dwuletniej współpracy zespołów lingwistów, programistów i naukowców z firmy Fido Intelligence Sp. z o.o. z Pomorskiego Parku Naukowo-Technologicznego w Gdyni i Katedry Inżynierii Wiedzy, Wydziału Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej, prowadzonej w ramach projektu rozwojowego "ISPAD - Inteligentny system pozyskiwania i analizy danych z serwisów społecznościowych w Internecie", finansowanego ze środków na naukę w latach 2009-2011 przez MNiSW i NCBR. Prototyp systemu i wspierające go narzędzia zostały wdrożone w środowisku testowym firmy Fido, rozwijającej zaawansowane technologie automatycznego rozumienia i przetwarzania treści w języku naturalnym. Istotą opracowanego rozwiązania jest jego elastyczność, pozwalająca identyfikować treści wyrażane przez poszczególne grupy użytkowników Internetu, posługujących się charakterystyczną dla nich odmianą języka polskiego. Jest to bardzo ważne dla skutecznego wyszukiwania treści, bowiem w Sieci dominują teksty wyrażane w żywym i zmieniającym się języku potocznym. Ze względu na stosowaną terminologię, podejmowane tematy, komentowane wydarzenia czy charakterystyczne zwroty, spotkamy tam język specyficzny dla hobbystów albo profesjonalistów z rozmaitych dziedzin (np. muzyków rockowych), celebrytów albo przedstawicieli różnych subkultur czy sekt, osób prywatnych albo instytucji, itd. Budowę wypowiedzeń w każdym z tych języków można opisać za pomocą szablonów językowych, które w połączeniu z

odpowiednimi słownikami i regułami uczenia maszynowego pozwalają systemowi ISPAD lokalizować interesujące strony i automatycznie analizować ich treść. Wyposażenie systemu w szablony językowe i reguły uczenia maszynowego wymaga oczywiście uprzedniego przeprowadzenia odpowiednich badań, które w przypadku pierwszej wersji prototypu systemu ISPAD dotyczyły języka wyrażającego treści potencjalnie niebezpieczne, propagujące przemoc, rasizm i nawołujące do przestępstwa. Po odpowiednim skonfigurowaniu systemu poprzez wprowadzenie opracowanych szablonów do jego bazy wiedzy, ISPAD analizuje automatycznie Sieć, ucząc się rozpoznawać treści z opracowanego zakresu coraz szybciej i skuteczniej.

W kolejnych rozdziałach książki przedstawiamy najważniejsze zagadnienia i sposoby ich rozwiązania, niezbędne do budowania systemów tej klasy, co ISPAD. Rozdział 1. wprowadza pojęcie gramatyki języków formalnych i możliwości ich automatycznej analizy, ze szczególnym uwzględnieniem ograniczeń teoretycznych, znanych w lingwistyce matematycznej. Rozdział 2. koncentruje się na kluczowym dla analizy języka pojęciu automatu. W szczególności opisuje podstawowe typy automatów, metody ich budowy oraz przekształcania, niezbędne m.in. do tworzenia słowników. Rozdział 3. podejmuje zagadnienie analizy języka naturalnego jako takiego, i prezentuje warstwowy model jego przetwarzania, wykraczający poza możliwości analizatorów języków formalnych opisanych w rozdziałach poprzedzających. Rozdział 4. koncentruje się na konkretnych szczegółach analizy tekstów w języku polskim, istotnych przy budowaniu szablonów językowych bazy wiedzy systemu ISPAD i stanowi zwieńczenie rozważań prowadzonych w wymienionych wcześniej rozdziałach. Rozdział 5. prezentuje problematykę języków zapytań stosowanych w wyszukiwarkach internetowych. Ich cechą charakterystyczną jest brak pogłębionej analizy treści stron, niemniej stanowią one ważny element wspierający działania praktycznie każdego użytkownika dowolnej przeglądarki internetowej. Rozdział 6. uzupełnia wcześniejsze rozważania dotyczące analizy treści tekstowych o podstawowe zagadnienia analizy treści graficznych. Choć opisany w Rozdziale 9. prototyp systemu ISPAD nie analizuje jeszcze treści w postaci graficznej (np. napisów prezentowanych jako obrazy, czy symboli graficznych), taka funkcjonalność jest przewidziana w jego dalszym rozwoju. Dwa kolejne rozdziały 7. i 8. przedstawiają najważniejsze modele i algorytmy, niezbędne przy wyposażaniu systemu w zdolność uczenia się. I tak, wpierw wprowadzane są pojęcia użyteczne przy reprezentowaniu cech rozpoznanej i wydobytej ze strony internetowej treści, a następnie przedstawione są podstawowe mechanizmy uczenia maszynowego pozwalające przyspieszyć działanie systemu i podnieść skuteczność rozpoznawania treści opisanych szablonami językowymi. Rozdział 9. opisuje w formie studium przypadku konkretny przykład – system ISPAD, jego architekturę, komponenty i niezbędne środowisko narzędziowe, wraz z wynikami badań pierwszego prototypu ISPAD, jakie zostały przeprowadzone podczas jego eksploatacji, bezpośrednio po wdrożeniu w środowisku docelowym.

Bogdan Wiszniewski
Gdańsk, lipiec 2011