

# Monitoring semantycznej pajęczyny

tekst: Cezary Biernacki

Doceniając znaczenie monitorowania informacji, które pojawiają się w internecie, stajemy przed problemem efektywnej realizacji tego zadania. Rozproszenie źródeł informacji i wzrost znaczenia wkładu każdego użytkownika Web uniemożliwia ręczne analizowanie przychodzących danych.

**M**onitoring prasy zawsze był ważnym narzędziem pracy działów PR i marketingu. Boom Web 2.0 – masowo tworzonych treści przez użytkowników w internecie – otworzył nowe możliwości i wyzwania dla specjalistów od PR, komunikacji i marketingu i również dla samego monitoringu. Obecnie trudno przecenić moc oddziaływania sieci. Wystarczy przypomnieć, że afera z Moniką Lewinsky rozpoczęła się, gdy pewien blog poinformował o tym, że renomowane czasopismo wycofało się z publikacji artykułu na ten temat. Internauci to szybko przechwycili i już po kilku dniach informacja stała się czołową wiadomością w prasie i telewizji. Wszyscy pamiętamy, jak skończyło się to dla reputacji prezydenta Stanów Zjednoczonych.

Ale już sama próba ręcznego sprawdzenia, ile jest stron zawierających nazwy popularnych marek, pokazuje skalę trudności. Oto ilości stron znalezionych przez Google, przy próbie wyszukania kilku przykładowych marek tylko dla polskich stron: Nokia – 1 610 tys., empik – 1060 tys., Zumi – 507 tys., Fortis – 403 tys., Pampers – 363 tys., Marlboro – 92 tys.

Niemale to liczby, a nie uwzględniają zawartości grup dyskusyjnych oraz całego tzw. szarego weba, tzn. miejsc, do których wyszukiwarki internetowe nie potrafią dotrzeć.

Jeszcze do niedawna nie istniały narzędzia, które sprostająby tego typu zadaniom, a obecnie wchodzi one do komercyjnego użytku. Jednym z motorów ich rozwoju była walka z terroryzmem. W pierwszej kolejności tam zaistniała konieczność przeglądania milionów e-maili, setek tysięcy stron internetowych, po to by

znaleźć wskazówki na temat potencjalnych ataków, rozpracować organizacje terrorystyczne, znaleźć miejsca rekrutacji i zbadać nastroje społeczności sympatyzujących z terrorystami. Dało to potencjał do przyspieszenia prac nad tzw. przetwarzaniem języka naturalnego, czyli sposobom pozwalającym komputerom (przynajmniej częściowo) zrozumieć to, co jest wyrażone w zwykłym tekście pisany lub mówionym.

Z bardziej cywilnych obszarów – rozwój narzędzi monitoringu stał się możliwy dzięki postępowi w zakresie gromadzenia wiedzy i odpowiedniego zarządzania nią. Narzędzia tego typu powstały w ramach prac nad inicjatywą Semantic Web. Zamierzeniem Semantic Web, czyli semantycznej pajęczyny, jest zorganizowanie danych udostępnionych w webie tak, aby programy same mogły je wiązać ze sobą i wnioskować bez udziału człowieka. Ostatnim kluczowym elementem układanki są doświadczenia z analizą zależności w dużych zbiorach danych, które to analizy są wykonywane często w tzw. korporacyjnych hurtowniach danych.

Jak działa automatyzacja monitoringu? Kolejne kroki to:

- ▶ Zbieranie i indeksowanie danych z internetu (lub innych dostępnych cyfrowych źródeł informacji)
- ▶ Preselekcja interesujących danych
- ▶ Analiza znaczeniowa wyselekcjonowanych danych
- ▶ Agregacja danych
- ▶ Wyznaczenie trendów i powiązań między danymi
- ▶ Prezentacja i raportowanie dla końcowych użytkowników

Efektywność tego procesu jest kluczowym elementem, dlatego między kolejnymi fazami przekazywane są zwrotnie informacje umożliwiające ciągłe uczenie się systemu tak, aby lepiej mógł wykonywać swoje zadania przy kolejnych analizach.

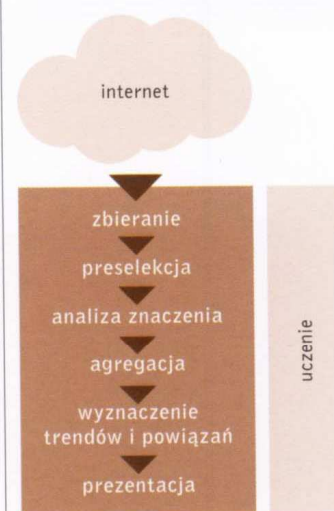
Zbieranie danych w najprostszej postaci polega na kolekcjonowaniu kolejnych stron, które można znaleźć poruszając się po linkach na tych stronach. Wykonują to programy zwane web-crawleami. Znalezione dokumenty są indeksowane według zawartych na nich słowach. Podobnie działają wyszukiwarki internetowe, więc nie ma tu nic szczególnego, poza samą skalą przedsięwzięcia. Trzeba ściągnąć co najmniej setki milionów stron, angażując do tego

dziesiątki komputerów połączonych w specjalne klastry dysponujące bardzo pojemnymi łączami do kluczowych punktów internetu.

Ponieważ ściągniętych dokumentów jest zwykle zbyt dużo, by było możliwe ich efektywne przetworzenie, następnym etapem jest ich preselekcja. Polega ona na zautomatyzowanym wyszukaniu tylko tych treści, które mogą potencjalnie dotyczyć monitorowanych zagadnień. Wykorzystuje się tu przygotowane wcześniej listy słów kluczowych – nazw marek, terminów fachowych, określeń powiązanych itd. Działa to podobnie jak użytkownik Google, który wpisuje swoje zapytania, a następnie przegląda kolejne znalezione strony, by stwierdzić, czy zawierają interesujące go zagadnienia. Podstawowa różnica w tym przypadku kryje się w liczbie zapytań i znalezionych stron – zwykle automaty zadają tysiące zapytań i przekazują do dalszej obróbki kilkaset tysięcy znalezionych stron. Na tym etapie następuje też odrzucenie stron zawierających tekst w językach, których system nie obsługuje (najczęściej monitoring prowadzony jest tylko dla jednego wybranego języka).

Analiza znaczeniowa to jeden z najtrudniejszych etapów procesu. Głównym jego celem jest wydobycie znaczenia informacji zawartych na każdej stronie wyznaczonej na etapie preselekcji. Wciąż nie ma jednego, doskonałego rozwiązania, więc stosuje się równocześnie dziesiątki różnych taktyk. Przede wszystkim znalezione strony są klasyfikowane według treści, formatu czy budowy i w zależności od wyznaczonej klasy dobierany jest dalszy proces obróbki. Wynika to z tego, że trochę innym językiem posługują się autorzy komentarzy na blogach, a innym autorzy artykułów z prasy profesjonalnej czy publikacji naukowych, a dostosowanie do konkretnego profilu tekstu ma dużą wagę dla osiągnięcia zamierzonych rezultatów w dalszej obróbce. System maksymalnie stara się też wykorzystać elementy strukturalne, to znaczy pewne wzorce w budowie stron ułatwiające zrozumienie ich zawartości. Są to na przykład tagi, popularne w Web 2.0 etykiety związane z artykułami i wiadomościami albo niewidoczne dla zwykłego użytkownika, zaszyte w treści informacje w postaci tzw. mikroformatów, czy punktowe oceny dostępne na wielu stronach przeznaczonych do recenzowania produktów przez

## Automatyzacja monitoringu



» użytkowników. Generalnie jednak w większości przypadków znaczna część najciekawszych dla monitoringu informacji zawarta jest w treści napisanej językiem naturalnym. Także tutaj jest dostępna szeroka gama narzędzi, które można podzielić na dwie kategorie – regułowe i statystyczne. W skrócie można powiedzieć, że te pierwsze wymagają opisanego procesu analizy tekstu w postaci ścisłej sekwencji operacji wykonywanych na tekście, natomiast metody statystyczne wykorzystują mechanizmy nauki maszynowej, gdzie oprogramowanie samodzielnie odkrywa zasady przetwarzania, korzystając z ogromnej liczby uprzednio przygotowanych przykładów i zdefiniowanych reguł.

Niezależnie od metody, analiza znaczeniowa w przypadku monitoringu brandów ma wychwycić dwa interesujące nas fakty: po pierwsze, czy i gdzie są wzmianki dotyczące monitorowanych marek, produktów, po drugie – jaka jest wyrażana opinia. Zadanie wychwytywania nazw z pozoru wygląda na proste, ale w praktyce wymaga pokonania wielu skomplikowanych problemów. Wiele bowiem nazw produktów lub marek to potoczne słowa (np. Era, Plus) i przetwarzany tekst może oczywiście zawierać takie słowo, ale w zupełnie innym kontekście. Niejednoznaczności tego rodzaju zmuszają do opracowania zaawansowanych mechanizmów dostosowujących algorytmy przetwarzania tekstu do konkretnych przypadków. W zamian można osiągnąć poprawność ich działania na poziomie sięgającym 95 proc., co generalnie jest wystarczające w tego typu zastosowaniach. Trudniejszym zadaniem jest zrozumienie, co tak właściwie napisano o konkretnej marce, czy jest to np. opinia pozytywna czy negatywna. Oczywiście taktyk i metod rozwiązania tego zagadnienia jest wiele. Przykładowo jedna ze stosowanych metod zamiast doszukiwać się pełnego znaczenia, sprawdza jedynie, czy w okolicy nazwy produktu występują słowa oceniające, nacechowane emocjonalnie lub takie np. jak: „tani”, „drogi”, „kiepski”, „dobry”, „lubię”, „polecam”. Oczywiście przy zastosowaniu prostych metod wielokrotnie może się zdarzyć nieprawidłowe zrozumienie sensu. Na przykład w zdaniu „Mam produkt A i B, ale tylko ten pierwszy jest dobry” może spowodować powiązanie produktu „B” ze słowem „dobry”. Jednak w globalnym podsumowaniu

pomyłki na korzyść powinny się zniwelować z pomyłkami na niekorzyść.

I tak dochodzimy do etapu agregacji. W etapie tym podlicza się, jak często nazwy produktów, usług i marek monitorowanych występują razem z określonymi ocenami i w jakich kontekstach (np. jak często produkt porównywany jest pozytywnie lub negatywnie do konkurencji). Dodając wymiar czasowy, można śledzić, jak opinie internautów zmieniają się np. pod wpływem kampanii marketingowych.

Ostateczne wyniki monitoringu prezentowane są w postaci regularnie wykonywanych raportów wizualizujących obecność poszczególnych marek na rynku i związane z nimi opinie. Dla użytkowników potrzebujących bardziej zaawansowanych raportów i zestawień mogą zostać udostępnione narzędzia analityczne, które pozwalają na porównanie dowolnie wybranego wymiaru i umożliwiające „wglębnienie się” w dane, dochodząc aż do konkretnych stron stanowiących ich źródło.

Aby system monitoringu nadążał za zmianami w zwyczajach językowych lub pojawianiem się nowych nazw, w każdy etap procesu muszą być wbudowane mechanizmy umożliwiające jego samodzielne lub nadzorowane uczenie się. Może to polegać np. na uzupełnianiu list zapytań do preselekcji lub modyfikowaniu reguł sterujących analizy znaczeniowej. Odpowiednio elastyczny system nie dość, że zachowa swoją sprawność, to może nawet ulepszać swoje działanie bez kosztownych prac programistycznych.

Przedstawione powyżej rozwiązania działają już eksperymentalnie. Można się spodziewać, że w ciągu najbliższych dwóch, trzech lat staną się codziennym narzędziem w działach PR, komunikacji i marketingu lub w firmach monitorujących opinie konsumentów na temat produktów, ale również np. polityków. Podobne technologiczne mechanizmy, choć w obszarze monitoringu danych osobowych i reputacji osobistej, oferuje już dziś np. brytyjska firma Garlik, [www.garlik.com](http://www.garlik.com). ■

Cezary Biernacki, Senior Project Manager, Software Mind.  
cezary.biernacki@softwaremind.pl