

Traffic Index

Traffic Index, nowatorski algorytm wyszukiwawczy pozwalający ocenić obiektywną wartość serwisów internetowych, to efekt wielomiesięcznej pracy informatyków NetSprinta. Dzięki temu rozwiązaniu wyszukiwarka dostarcza Użytkownikom wartościowe wyniki oraz walczy z nieuczciwym pozycjonowaniem. Wszystko to dzięki zmniejszeniu znaczenia analizy linków na rzecz badania ruchu na stronie.

Na początku 2000 roku zespół warszawskich programistów, kierowanych wówczas przez Tomasza Skalczyńskiego i Kamila Nagrodzkiego rozpoczął pracę nad wyszukiwarką NetSprint. Był to okres, gdy królem wyszukiwarek była jeszcze Altavista.

Tworząc NetSprinta programiści przyjęli, podobnie jak nieco wcześniej zespół Google, że w natłoku milionów stron, jakie pojawiają się w sieci, najważniejszym zadaniem wyszukiwarki będzie nie tylko wyświetlanie stron związanych z treścią zapytania Użytkownika, ale przede wszystkim tych będących obiektywnie wartościowym źródłem informacji.

W wyniku takiego założenia programiści NetSprinta już wówczas stworzyli mechanizm, który określał, jaka jest wartość poszczególnych dokumentów w sieci. Podobnie jak w mechanizmie PageRank stworzonym przez Google, NetSprint oparł się na analizie linków. Badano ile odnośników prowadzi do danej strony www, jak również czy pochodzą one z wartościowych serwisów. Dzięki temu mechanizmowi strony częściej polecane przez inne serwisy były prezentowane wyżej w wynikach wyszukiwania. Precyzyjniej mówiąc, na pierwszej stronie wynikowej znajdowały się te strony, które miały jednocześnie dużo linków prowadzących do siebie z popularnych stron (co świadczyło o ich wysokiej wartości obiektywnej) oraz zawierały treści bezpośrednio związane z zapytaniem Użytkownika (dzięki zaawansowanej analizie językowej).

Badanie linków słabnie

Przez długi czas mechanizm badania linków dobrze spełniał swoje zadanie. Jednak wraz z popularyzacją wiedzy na temat znaczenia linkowania dokumentów coraz częściej okazywało się, że webmasterzy nie zawsze polecają wartościowe dokumenty.

Ponieważ ruch z wyszukiwarek łatwo przekłada się na przychody właścicieli serwisów internetowych, część z właścicieli serwisów zaczęła wykorzystywać tę wiedzę **sztucznie zawyżając pozycję** swojej strony w wynikach wyszukiwania. W efekcie **słuszny w swoim założeniu mechanizm działał coraz słabiej**. Właściciele wyszukiwarek byli zmuszeni rozpocząć walkę z różnego rodzaju nadużyciami

webmasterów i spamerów. W wyniku ich nieuczciwych praktyk powstał cały przemysł mający na celu promocję stron niezastługujących na wysoką pozycję w rankingu.

Dodatkowo, wraz z eksplozją Internetu, powstawało coraz więcej serwisów, których nikt nie linkował. Oznaczało to, że **duża część stron – w tym często wartościowych – pozostawała na marginesie Internetu.**

W tej sytuacji krytyczne stawało się znalezienie nowego mechanizmu, który miał pozwolić NetSprintowi określać wartość obiektywną serwisu. Badania obejmowały szeroką listę czynników wpływających na wartość strony:

- miejsce pochodzenia linku,
- wartość serwisu w opinii internautów i redaktorów katalogów,
- tematykę serwisu z jakiego pochodzi link itd.

W wielu przypadkach zastosowanie dodatkowych kryteriów dawało pozytywne efekty i nowe mechanizmy były implementowane w algorytm NetSprinta. Tym samym ranking badania wartości stron stawał się coraz bardziej złożony. Zespół informatyków NetSprinta cały czas szukał alternatywy, czegoś, co pozwalałoby jednoznacznie określić wartość danego serwisu, a nie byłoby obciążone wadami badania linków.

Najwięcej dobrych stron

„Sytuacja stała się jeszcze bardziej dramatyczna, kiedy podjęliśmy decyzję, że NetSprint będzie indeksował najwięcej wartościowych, polskich stron” – przyznaje Piotr Kozłowski, szef zespołu IT. Wielokrotne zwiększenie liczby przeszukiwanych dokumentów wymagało posiadania bezbłędnego i niepodlegającego manipulacjom algorytmu. Obrazowo rzecz ujmując z 60 lub 100 mln dokumentów dużo trudniej jest wybrać 20 najlepszych wyników niż z 22 mln jakie wcześniej przeszukiwała wyszukiwarka.

W poszukiwaniu wartości obiektywnej

Celem zespołów NetSprinta i WP było znalezienie obiektywnej miary wartości dla każdego polskiego serwisu internetowego. Miara ta miała spełniać następujące kryteria:

- **być kompleksowa** – powinna obejmować cały polski Internet.
- **trudna do zmanipulowania** – właściciel serwisu powinien mieć tylko pośredni wpływ na jego obiektywną wartość, jedynie poprzez poprawianie jego wartości merytorycznej oraz zwiększanie społeczności użytkowników regularnie z niego korzystającej.
- **stałe aktualizowana** – Internet rozwija się w niezwykłym tempie – zastosowana miara powinna to uwzględniać.
- umożliwiać **porównywanie względem siebie** zarówno **najpopularniejszych portali** jak i **małych serwisów hobbistycznych.**

Badanie polskiego Internetu – Megapanel PBI/Gemius

Pod koniec 2004 zakończyły się prace nad unikalnym w skali światowej Badaniem Megapanel PBI/Gemius, łączącego właściwości badania typu user-centric (badanie panelowe) z badaniem typu site-centric (monitorujące ruch na witrynach internetowych).

Celem badania jest poznanie liczby i profilu społeczno-demograficznego użytkowników Internetu oraz sposobu, w jaki internauci korzystają z sieci. Polska sieć długo czekała na niezależną i wiarygodną analizę, która ukaże jej pełny i prawdziwy obraz. **Wyniki badania umożliwiają porównywanie popularności witryn i aplikacji internetowych.** Co również bardzo ważne z punktu widzenia algorytmu Traffic Index badania oparte są na obiektywnej próbie a jego wyniki są praktycznie nie możliwe do zniekształcenia.

Zespół wyszukiwarki z dużym zainteresowaniem obserwował przebieg i wyniki pierwszych badań. Z punktu widzenia NetSprinta szczególnie istotne były badania typu user-centric (skoncentrowane wokół Użytkownika). Obserwacja i zgromadzone dane z zachowań kilkudziesięciu tysięcy Internautów (panelistów) biorących udział w badaniu pozwalają na precyzyjne określenie wartości poszczególnych serwisów (niezależnie od tego czy korzystają one z systemu statystyk Gemiusa, czy też nie).

NetSprint rozpoczął rozmowy z firmą badawczą Gemius na temat uwzględnienia wyników badań w nowotworzonym algorytmie NetSprinta. Finalizację rozmów poprzedziły rozbudowane testy dotyczące możliwości integracji wyników badań Megapanel z silnikiem wyszukiwarki. Po ich pozytywnym zakończeniu firmy podjęły decyzję o rozpoczęciu współpracy i w kwietniu podpisały stosowną Umowę.

Dzięki współpracy z Gemusem NetSprint wzbogacił swoją wiedzę na temat polskiego Internetu – a dokładniej kilkuset tysięcy domen, które zostały kiedykolwiek odwiedzone przez panelistów biorących udział w badaniu. Tym samym Użytkownicy Internetu pomagają ocenić NetSprintowi wartość stron a nowotworzony **algorytm oparł się na obiektywnej i nie zmanipulowanej ocenie Internautów.**

Dodatkowym atutem oparcia się na wynikach badania jest fakt, że są one przeprowadzane co miesiąc. Dzięki temu **NetSprint posiada wciąż uaktualnianą wiedzę** na temat szybko zmieniającego się i rozwijającego Internetu.

Traffic Index - alternatywa dla innych systemów wyszukiwawczych

Informatycy NetSprinta stanęli przed problemem jak można wykorzystać wiedzę otrzymaną dzięki badaniu w mechanizmie wyszukiwawczym. W jaki sposób na podstawie danych o oglądalności serwisu, ilości jego użytkowników, czasie, jaki spędzają na nim czy ich lojalności określić jedną uniwersalną miarę wartości poszczególnych domen?

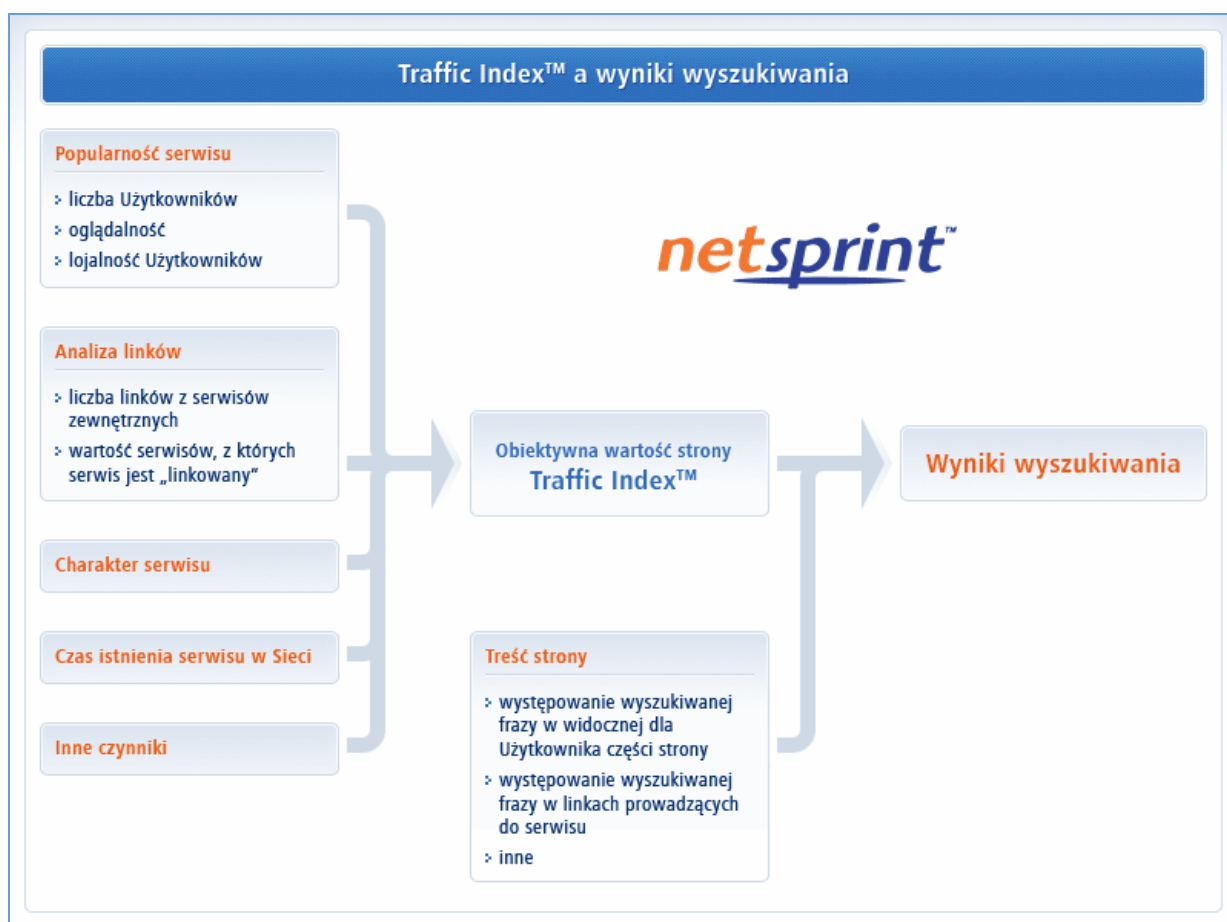
Po wielu testach i miesiącach pracy programiści NetSprinta stworzyli algorytm, który brał pod uwagę wszystkie wymienione czynniki i wykorzystywał całą dostarczaną

wiedzę z badania. Algorytm ten stanowi obecnie bardzo istotną część nowego mechanizmu.

Oczywiście system nadal wykorzystuje dotychczasowe mechanizmy badania wartości stron www (czyli np. badanie linków czy uwzględnianie charakteru poszczególnych serwisów). Ponieważ jednak nowy algorytm jako **pierwszy na świecie uwzględnia w tak dużym stopniu analizę ruchu na poszczególnych serwisach** – został nazwany **Traffic Index**.

„Zależało nam, żeby nowe rozwiązanie stanowiło wyraźną alternatywę dla dotychczasowych mechanizmów wyszukiwawczych dlatego na jego dopracowanie poświęciliśmy dużą część energii i potencjału Firmy. Wykorzystaliśmy też naszą aktywną współpracę z Pionem Technologii Informatycznej Wirtualnej Polski, wdrażającą w swoich mechanizmach wyszukiwawczych nasze rozwiązania. Pracownicy portalu pełnili funkcje doradcze w projekcie, pomagając w podjęciu optymalnych decyzji dotyczących założeń funkcjonowania algorytmu.” – podkreśla prezes NetSprinta Artur Banach.

„Warto też podkreślić demokratyczny charakter nowego algorytmu. Potencjalnie każdy Użytkowników może brać udział w określaniu wartości poszczególnych stron www. Będąc panelistą i korzystając z danego serwisu Użytkownicy głosują na niego wpływając pośrednio na wzrost współczynnika Traffic Index dla danej domeny” – dodaje Piotr Kozłowski.



Stawić czoło problemom

Prace nad nowym algorytmem trwały w sumie ponad 9 miesięcy. Duża część wysiłków koncentrowała się na rozwiązywaniu problemów, jakim muszą stawić czoło nowoczesne wyszukiwarki.

Przed wszystkim zadaniem Zespołu NetSprinta było znalezienie odpowiedniej skali premiowania serwisów o wysokim współczynniku Traffic Index. *„Mieliśmy świadomość, że istnieje bardzo wiele wartościowych serwisów, które nie posiadają jeszcze dużej i lojalnej widowni”* – mówi Piotr Kozłowski – *„Tymczasem to właśnie te serwisy często dostarczają wartościowsze informacje niż znane portale”*.

Kluczem było więc stworzenie odpowiednich mechanizmów, które pozwolą zidentyfikować takie sytuacje i **zapewnić dobrą pozycję serwisom o merytorycznie wysokim poziomie pomimo niewielkiej oglądalności**. W tym celu NetSprint co tydzień przeprowadzał badania jakości swoich wyników wyszukiwania. Pozwalały one NetSprintowi stale optymalizować i udoskonalać działanie algorytmu doprowadzając go do dzisiejszego stanu zaawansowania.

Innym istotnym problemem było traktowanie w rankingu ruchu z innych wyszukiwarek internetowych. W pierwszym założeniu algorytm miał się opierać jedynie na danych z ruchu pochodzącego z odwołań bezpośrednich oraz odnośników z innych serwisów internetowych. Stała za tym obawa, że uwzględnienie ruchu z wyszukiwarek spowoduje, iż duży wpływ na wyniki NetSprinta będą miały inne wyszukiwarki internetowe.

Okazało się jednak, że takie podejście posiadało również istotne wady. Po pierwsze internauci traktują często wyszukiwarki jak przeglądarkę i tam wpisują adres serwisu lub nazwę domeny. Po drugie nie uwzględnianie tego ruchu premiowałoby serwisy o prostych adresach www, które są możliwe do zapamiętania przez innych Użytkowników, nieświadczących jednak o ich zawartości merytorycznej. *„Ostatecznie podjęliśmy decyzję, że ruch z wyszukiwarek otrzyma mniejszą wagę od ruchu wygenerowanego poprzez odwołanie bezpośrednie czy odsyłacze z innych stron WWW, ale będzie jednak uwzględniany w Traffic Index”* – tłumaczy Piotr Kozłowski.

Przed wszystkim zyskają Użytkownicy

Najbardziej na wprowadzeniu Traffic Index skorzystają Użytkownicy. Dzięki uwzględnieniu w algorytmie wiedzy na temat ruchu na poszczególnych serwisach w wynikach wyszukiwania:

- wyżej prezentowane będą serwisy zawierające **wartościowe dla Użytkowników treści**

- witryny o wysokim wskaźniku Traffic Index będą częściej odwiedzane przez spidera NetSprinta. Dzięki temu w wyszukiwarce **przeszukiwane będą aktualne treści** pochodzące z tych serwisów
- w dużym stopniu **eliminowane są strony bezwartościowe**, nie odwiedzane przez Internautów. Dzięki temu w wynikach rzadko znajdują się strony tworzone jedynie w celu ich wysokiego pozycjonowania w wyszukiwarkach (**spam**).

Aby podkreślić znaczenie Traffic Index Zespół NetSprint stworzył zupełnie nową wersję serwisu netsprint.pl. Internauci, korzystający z tej wyszukiwarki, mogą zapoznać się z pozycją najpopularniejszych domen bezpośrednio w wynikach wyszukiwania (tzw. „Popularna strona”). Dokumenty zebrane w ciągu ostatnich 72h dodatkowo oznaczone ikoną aktualna strona.

I to, co prawdopodobnie najważniejsze. Mogą przeszukiwać ponad 60.000.000, a wkrótce dużo więcej polskich stron. Z pomocą Traffic Index powinni łatwo wyszukać te najbardziej wartościowe.