POLITECHNIKA KRAKOWSKA
IM. TADEUSZA KOŚCIUSZKI
W KRAKOWIE

Wydział Informatyki i Telekomunikacji

Informatyka

Raport z realizacji tematu pt.

INDEKSOWANIE STRON PRZEZ GOOGLE

Michał Cichocki

Adrian Cygan

Szymon Czajkowski

Abstrakt

Niniejszy raport dotyczy zasady działania przeglądarki Google oraz procesów, jakie umożliwiają jej działanie. Na podstawie oficjalnych źródeł, przedstawiany jest sposób w jaki wyszukiwarka wykorzystuje dane do efektywnego i jakościowego indeksowania stron internetowych. Wskazane są kluczowe czynniki jakimi kieruje się Google w celu optymalizowania wyników wyszukiwań i jakie narzędzia są do tego wykorzystywane.

Wstęp

Cel pracy

Przedstawienie procesu w jakim Google indeksuje strony internetowe i sposobów optymalizowania witryn.

Zakres

Praca obejmuje następujące tematy:

1. Wyjaśnienie czym jest indeksowanie i dlaczego jest kluczowe dla widoczności w wynikach wyszukiwania.
2. Omówienie etapów procesu działania wyszukiwarki, takich jak skanowanie stron (crawling), analiza treści i rankingowanie.
3. Analizę czynników wpływających na poprawność indeksacji.
4. Przedstawienie narzędzi wspierających proces indeksowania.
5. Wskazanie sposobów na poprawę skuteczności indeksowania.

Metodyka

W celu opracowania raportu, przeanalizowano oficjalne dokumentacje Google Search Central oraz dane pochodzące z narzędzia Google Search Console. Ponadto wykorzystano artykuły pochodzące od pracowników Google oraz z Uniwersytetu w Stanford.

Część teoretyczna

Czym jest indeksowanie

Indeksowanie stron internetowych polega na analizie witryn pod kątem słów kluczowych, linków czy obrazów i zapisywanie tych danych w bazie danych wyszukiwarki. Wyszukiwarka internetowa nie przeszukuje wszystkich istniejących stron internetowych. Zamiast tego korzysta z indeksu tej bazy danych. Do indeksowania wykorzystywane są specjalnie do tego celu napisane programy, czyli roboty sieciowe.

Jak działa Google: przeszukiwanie, indeksowanie i ranking.

Przeszukiwanie

Celem przeszukiwania jest zebranie jak największej ilości informacji o stronie. Google wykorzystuje oprogramowanie zwane robotami (Googleboty) do budowania indeksu. Roboty wchodzą na publiczne strony internetowe i otwierają linki zawarte na nich, identycznie jak użytkownicy stron. Przechodząc między stronami gromadzą informację na temat ich zawartości. Do indeksacji analizują jedynie pierwsze 15MB plików źródłowych (np. pliki .html, .css czy .js). Googleboty regularnie odwiedzają te same witryny, aby sprawdzić czy zostały one zaktualizowane i poprawić ich indeksacje. Istnieją różne typy robotów. Typowy Googlebot służy do indeksacji stron internetowych na wersje komputerowe i mobilne, Googlebot Image służy do indeksacji obrazów, Googlebot Video służy do indeksacji treści związanej z filmami video.

W procesie przeszukiwania, szczególną rolę odgrywa plik “robots.txt”, Służy on do implementacji protokołu Robots Exclusion Protocol, czyli standardu używanego przez strony internetowe do wskazania crawlerom i botom jakie fragmenty strony są dla nich dostępne

W połączeniu z plikiem “robots.txt” stosuje się także mapy witryny. Są to pliki, w którym znajdują się informacje o stronach, filmach i innych plikach na stronie oraz związkach między tymi elementami, Google odczytuje je, aby skuteczniej indeksować złożone strony. Mapa witryny informuje wyszukiwarkę, które strony i pliki w witrynie są ważne oraz zawiera informacje np. o alternatywnych wersjach językowych strony, dacie aktualizacji strony. Mapa witryny pomaga udostępnić informacje o konkretnych rodzajach treści na stronach np. wpis o filmie może informować o czasie trwania lub kategorii wiekowej, wpis o obrazie o lokalizacji obrazów umieszczonych na stronie, wpis o wiadomościach może zawierać tytuł artykułu, datę publikacji.

Indeksowanie

Google analizuje zgromadzone dane i dodaje je do swojego indeksu – bazy danych zawierającej informacje o stronach internetowych. Google ocenia strukturę strony, treść oraz jej wartość użytkową. Badane są teksty, obrazy, słowa kluczowe, tagi itp. Sprawdzane są dodatkowe elementy, takie jak linki wewnętrzne, zewnętrzne oraz ich kontekst. Strony, które mają błędy (np. brak treści, niedostępność dla robotów), mogą nie trafić do indeksu.

Indeks wyszukiwarki Google zawiera setki miliardów stron internetowych, których rozmiar to ponad 100 milionów gigabajtów z wpisami dla wszystkich słów wyświetlanych na każdej indeksowanej stronie. Systemy indeksowania Google nadążają za nieustannymi zmianami zachodzącymi w Internecie. Uczą się jak często na odwiedzanych stronach zachodzą zmiany oraz odkrywają nowe treści.

Indeks Google zawiera informacje znalezione nie tylko w Internecie, ale pozwala również na przeszukiwanie wśród milionów książek z największych bibliotek, sprawdzanie rozkładów jazdy, danych banku światowego. Niektóre informacje wyświetlane przez wyszukiwarkę są wyszukiwane w tzw. grafie wiedzy.
Graf wiedzy to baza Google, która zawiera miliardy faktów na temat ludzi, miejsc lub rzeczy. Są to głównie informacje znane i oparte na faktach np. wysokość wieży Eiffla. Google korzysta z danych ze źródeł publicznych, ale również pozyskuje licencjonowane dane, takie jak wyniki sportowe, ceny akcji czy prognozy pogody.

Twórcy mogą wpływać na indeksowanie treści korzystając z mapy witryn lub pliku robots.txt. Mogą również wykorzystać zestaw narzędzi Search Console.

Search Console - usługa Google, która pomaga monitorować pozycję witryny w wynikach wyszukiwania Google. Udostępnia narzędzia do:

- sprawdzenia jak Google widzi i indeksuje stronę
- wyświetlania danych o ruchu związanym z witryną (jak często strona pojawia się w wyszukiwaniach, jakie zapytania powodują jej wyświetlanie, jak często użytkownicy klikają te zapytania, by do niej przejść)
- sprawdzania, w jakich innych witrynach znajdują się linki do strony

- raportowania i rozwiązywania problemów z indeksowaniem i wysyłania prośby o indeksowanie nowych treści

Ranking

Kiedy użytkownik wpisuje zapytanie, Google przeszukuje indeks, aby znaleźć i uporządkować strony najbardziej adekwatne do tego zapytania. Proces ten nazywany jest rankingiem. Algorytmy Google sprawdzają czynniki takie jak treść zapytania, trafność, przydatność stron, profesjonalizm źródła oraz ustawienia użytkownika i lokalizacje. Waga przypisywana czynnikom różni się w zależności od charakteru zapytania np. dla zapytania o bieżące wydarzenia większą rolę odgrywa czas zamieszczenia treści niż w zapytaniu o definicje słownikowe.

Główne czynniki, pomagające określić jakie wyniki zostaną wyświetlone dla zapytania:

- Znaczenie - Aby wyświetlić użytkownikowi odpowiednie wyniki, Google tworzy modele językowe, aby jak najlepiej dopasować treść wyszukiwania do wyników. Obejmuje to czynności od rozpoznawanie literówek po modele rozpoznające znaczenie i intencję słów. Systemy Google dążą także do zrozumienia tego, jakiego rodzaju informacje są poszukiwane, np. gdy użytkownik użyje słów “gotowanie” i “zdjęcia” to systemy pokażą mu przepisy i grafikę. Jeśli zapytanie zostanie napisane w jakimś języku, system pokaże wyniki w tym języku. Systemy Google potrafią pokazywać odpowiedzi w kontekście lokalizacji użytkownika, rozumieją również słowa zyskujące w danym czasie popularność, pokazując użytkownikowi najświeższe informacje.
- Trafność - Systemy Google analizują treści, aby określić, czy zawierają one informacje, o które zapytał użytkownik. Podstawowym czynnikiem jest ilość występujących w wyniku słów kluczowych. Systemy korzystają też z danych o interakcjach, aby ocenić trafność zapytań. Te dane są przekształcane w sygnały, które pomagają systemom uczyć się coraz lepiej oszacowywać trafność.
- Jakość - Gdy systemy rozpoznają, że dane treści są trafne starają się też nadać im odpowiedni priorytet. Systemy Google wyposażone są w mechanizmy identyfikacji sygnałów, które pomagają wskazać treści wiarygodne. Jednym z czynników istotnych w tym procesie jest obecność linków lub odniesień do tych treści na innych stronach. Google dba o to, aby zachować równowagę między trafnością i rzetelnością informacji.
- Użyteczność - Systemy sprawdzają czynniki techniczne strony np. jej dostosowanie do urządzeń mobilnych, szybkość ładowania treści.
- Kontekst - Google korzysta z informacji o lokalizacji, historii wyszukiwania oraz ustawień wyszukiwarki. Pomagają one odpowiadać na zapytania bardziej trafnymi i użytecznymi treściami. System poznaje zainteresowania, aby proponować trafne wyniki, równocześnie nie szuka informacji poufnych.

Google wykorzystuje zaawansowane algorytmy, które analizują setki czynników w celu ustalenia kolejności wyświetlanych wyników.

RankBrain - pierwszy system uczenia głębokiego wykorzystywany przez wyszukiwarkę Google, wykorzystywany od 2015 roku. Pomógł skojarzyć wykorzystane słowa z ideą jaka za

nimi idzie. Dało to możliwość dopasowania odpowiednich treści nawet jeśli nie zawierają one słów użytych w pytaniu np. pytając o konsole Sony dostaniemy informacje o konsoli PlayStation pomimo, że słowa PlayStation nie było w pytaniu. Algorytm ten daje również możliwość poprawnego wyszukiwania w przypadkach użycia homonimów, czyli wyrazów które mają wiele znaczeń (np. zamek). Mimo nadal używany jest przez silnik.

PageRank - system nadający indeksowanym stronom internetowym określoną wartość (od 0 do 10), która oznacza jakość tej strony. Opracowany przez twórców Google Larry'ego Page'a i Sergeya Brina. Algorytm opiera się na zasadzie mówiącej, że jakość tekstu jest proporcjonalna do tego, ile innych tekstów się na niego powołuje. W odniesieniu do sieci chodzi o ilość odnośników na innych stronach do danej strony. Twórcy Google chcąc ulepszyć algorytm zaproponowali, aby waga odnośnika z danej strony na inną była większa, jeśli ta strona, na której istnieje odnośnik, ma wysoką wartość PageRank. Implementacja algorytmu pozostaje tajna i jest chroniona przez Google

BERT - Model językowy stworzony przez badaczy Google w 2018 roku. Dzięki wykorzystaniu architektury transformatora (architektura uczenia maszynowego), potrafi analizować zapytania utworzone przez użytkownika w ich całym kontekście. BERT analizując zapytanie nie analizuje je słowo po słowie, zamiast tego analizuje każde słowo w odniesieniu do słów użytych przed nim i po nim. Dzięki temu wyszukiwarka jest w stanie dać nam trafniejsze odpowiedzi dla naszych pytań.

Transformer- architektura głębokiego uczenia maszynowego opracowana w 2017 przez badaczy w Google roku oparta na zrównoleglonym mechanizmie uwagi (mechanizm uwagi to mechanizm uczenia maszynowego, który symuluje mechanizm ludzkiej uwagi. Działa on nadając wszystkim słowom w zdaniu wagi i na tej podstawie je analizuje).

Czynniki wpływające na indeksowanie

Fundamentem indeksacji strony jest jej poprawność techniczna. Utrzymując witrynę internetową należy pamiętać o cechach takich, jak:

- Poprawny kod HTML: Strona powinna być napisana zgodnie z semantyką i zasadami języka HTML, co usprawnia Googlebotom interpretowanie treści witryny.
- Responsywność: Strona powinna być przystosowana do urządzeń mobilnych, ponieważ Google stosuje mobile-first indexing.
- Czas ładowania: Wolno ładujące się strony mogą być trudniejsze do pełnego przeszukania.

Podsumowanie

W pracy przedstawiono narzędzia i techniki, jakie wykorzystuje Google, aby efektywnie indeksować strony. Zaprezentowano sposób w jaki system wyszukiwarki zbiera, analizuje i wykorzystuje dane o witrynach. Opisano wykorzystywane przez Google algorytmy przetwarzania języka naturalnego oraz wskazano czynniki techniczne, jakie wpływają na jakość indeksowania.

Bibliografia

[1] Porządkowanie informacji – jak działa wyszukiwarka Google
[2] https://developers.google.com/search/docs
[3] https://blog.google/products/search/how-ai-powers-great-search-results
[4] http://infolab.stanford.edu/~backrub/google.html