Indeksowanie stron internetowych przez Google – raport

Stanisław Greń

Informatyka, rok 3

26.10.2024r.

1. Abstrakt

Problem odpowiedniego pozycjonowania strony internetowej dotyka zarówno przedsiębiorców, jak i administratorów dużych serwisów internetowych. Jedynym sposobem na polepszenie swoich wyników jest dogłębna analiza pracy maszyny indeksowania wyszukiwarki Google. Praca przedstawia sposób w jaki realizowane jest indeksowanie oraz jak zmaksymalizować szanse na wysoką pozycje. Opisane w niej zostały istotne elementy składające się na pozycjonowanie stron przez wyszukiwarkę, takie jak: crawling czy indexing, czyli mechanizmy, które przy zachowaniu odpowiednich praktyk pomogą stronie znaleźć się wyżej w wynikach wyszukiwania.

2. Cel

Projekt ma na celu dostarczenie kompleksowej wiedzy, na temat procesu indeksowania stron, wpływu dobrej konfiguracji strony internetowej na jej pozycjonowanie (pliki: robots.txt, sitemap.xml czy metatagi języka HTML), oraz, na tej podstawie dostarczyć wskazówki, jak poprawić jakość swoich witryn i zwiększyć ich widoczność w wynikach wyszukiwania.

3. Zakres

a. Analizę procesu indeksowania przez Google, z przedstawieniem konkretnych etapów: crawling, indexing, wyświetlanie wyników.
b. Przedstawienie elementów wpływających na skuteczne indeksowanie stron.
c. Opis wyzwań i barier, takich jak problemy z treściami dynamicznymi i duplikatami.
d. Przegląd narzędzi, które wspierają webmasterów w optymalizacji witryn.
e. Przedstawienie najlepszych praktyk SEO i zaleceń optymalizacyjnych.

4. Metodyka

Praca podczas realizacji tematu opierała się o oficjalne materiały google, własne doświadczenie w pracy ze stronami internetowymi, dokumentacje techniczne narzędzi do pracy z ulepszaniem SEO oraz materiały dotyczące działania konkretnych narzędzi znalezione w internecie.

I. Część Teoretyczna

1. Proces indeksowania stron internetowych przez google

a. Czym jest indeksowanie?

Indeksowanie strony internetowej przez Google to proces, w którym Google analizuje i zapisuje informacje o stronie internetowej w swojej bazie danych, zwanej indeksem wyszukiwarki. Dzięki temu strona może być wyświetlana w wynikach wyszukiwania, gdy użytkownicy wpisują odpowiednie zapytania. Google do indeksowania wykorzystuje autorskiego robota o nazwie Googlebot, Robot ten przeszukuje sieć w poszukiwaniu nowych lub zaktualizowanych treści.

b. Googlebot

Googlebot to robot indeksujący firmy Google, który automatycznie przeszukuje strony internetowe w celu dodania ich zawartości do indeksu wyszukiwarki Google. Jest to kluczowe narzędzie w procesie zbierania danych o stronach internetowych. Googlebot przetwarza miliardy stron dziennie, Popularne strony mogą być odwiedzane przez Googlebota nawet kilkaset razy dziennie, szczególnie gdy ich treść jest często aktualizowana.

Istnieje kilka typów, np. Googlebot Desktop (symulujący przeglądarki na komputerach) oraz Googlebot Mobile (przeszukujący strony z perspektywy urządzeń mobilnych).

Googlebot ustala priorytety stron na podstawie:

- Popularności witryny,
- Aktualności treści,
- Jakości linków prowadzących do danej strony. Strony, które są trudno dostępne lub zablokowane w pliku robots.txt, mogą zostać pominięte w procesie crawlingu.

c. Crawling

Robot podąża za linkami, analizując zarówno linki wewnętrzne (prowadzące do innych stron w tej samej witrynie), jak i zewnętrzne (prowadzące do innych domen). Przy tym tworzy sieć połączeń miedzy zasobami strony. W skórcie, crawling to zbieranie informacji o stronie internetowej. Jest to pierwszy etap który pozwala na ocenę atrakcyjności strony.

Bot indeksujący pobiera kod strony, przy tym renderując ją w oparciu o najnowszą wersje przeglądarki Chrome. W taki sposób robot może zrozumieć jej strukturę i zawartość. Istotna w tym wypadku jest szybkość wygenerowania danych na stronie.

d. Indexing

Indexing to inaczej rozpoznanie tematyki strony. Robot analizuje jej zawartość w oparciu o analizę pliku html. Określa również czy strona jest kanoniczna, czyli stronę najbardziej reprezentatywną w zbiorze duplikatów. W uproszczeniu można powiedzieć, że to proces eliminacji duplikatów.

Kluczowe elementy uwzględniane w analizie to:

- Meta tagi (np. meta description, meta robots),
- Nagłówki HTML (H1, H2 itd.),
- Treść strony (w tym unikalne słowa kluczowe),
- Pliki graficzne z opisami (alt, text).
- Jeśli strona zawiera elementy oznaczone jako „noindex”, nie zostanie ona dodana do indeksu.

Indexing to też zapis wynikow wykonanej pracy przez Googlebota do baz danych. Indeks to nazwa właśnie tej bazy w której zapisywane są strony. Warto mieć na uwadze to, że nie wszystkie strony zostaną tam zapisane. Wszystko zależy od jakości naszej strony, na przykład tego czy nie jest duplikatem, w taki sposób google unika wyświetlaniu spamu.

Typowe problemy związane z indeksowaniem to między innymi:

- Niska jakość treści na stronie
- Dodane reguł zabraniających indeksowania
- Ekosystem projektu, który może to utrudniać

e. Wyświetlanie wyników

Wyświetlanie wyników to ostatni z elementów procesu indeksacji stron. Jest to moment w którym użytkownik wpisuje hasło w wyszukiwarkę. Google, na podstawie informacji, które zebrał podczas poprzednich procesów, wyświetla użytkownikowi wyniki, dostosowane do jego potrzeb.

Wyniki są wyświetlane na podstawie dużej liczby czynników, a najbardziej podstawowe to takie jak język, lokalizacja czy jakość strony. Dlatego wpisując to samo hasło, będąc w dwóch różnych państwach otrzymamy inne wyniki, to właśnie zasługa indeksowania. Przy wpisaniu tego samego hasła przez dwie osoby, również możemy otrzymać różne wyniki.

2. Kluczowe elementy techniczne

a. Plik robots.txt to plik tekstowy, który znajduje się w głównym katalogu witryny. Jego celem jest instruowanie robotów wyszukiwarek, które obszary witryny mogą być przeszukiwane, a które powinny zostać zignorowane.

   - User-agent: Określa, do którego robota odnosi się instrukcja (np. Googlebot). Symbol „*” oznacza wszystkie roboty.
   - Disallow: Wskazuje sekcje strony, które mają zostać pominięte przez robota.
   - Allow: Pozwala na indeksację określonych obszarów, mimo że nadrzędna sekcja jest zablokowana.

   Dlaczego robots.txt jest ważny?

   - Pozwala zarządzać zasobami witryny, ograniczając crawling mniej istotnych sekcji, co oszczędza zasoby Googlebota.
   - Zapobiega indeksacji stron administracyjnych, np. admin.

b. Plik sitemap.xml to mapa witryny w formacie XML, która zawiera listę wszystkich stron w witrynie, wraz z dodatkowymi informacjami, takimi jak częstotliwość aktualizacji, priorytet strony oraz data ostatniej zmiany.

   Dlaczego sitemap.xml jest ważny?

   - Pomaga robotom wyszukiwarek zidentyfikować wszystkie ważne strony w witrynie, nawet jeśli nie są one łatwo dostępne przez linki.
   - Zawiera dodatkowe dane, które umożliwiają Google efektywniejsze zarządzanie procesem indeksacji.
   - Jest szczególnie przydatny dla dużych witryn lub stron z dynamiczną strukturą.

- Plik sitemap.xml poprawia widoczność witryny w wyszukiwarkach, zapewniając, że wszystkie kluczowe strony są uwzględnione w procesie indeksowania.

c. Treść i jej jakość

Treści unikalne, aktualne i dostosowane do potrzeb użytkowników są kluczowe. Strony zawierające duplikaty treści, niskiej jakości artykuły lub treści generowane automatycznie nie są atrakcyjne dla algorytmów, oraz mogą spowodować zakwalifikowanie naszej strony jako spam.

d. Linki

- Linki wewnętrzne

  Są to linki umieszczone na stronie, wpływają na łatwość nawigacji i strukturę witryny. Przyjazne URL-e, to takie które tworzą strony z logiczną i czytelną strukturą adresów URL, takie też są łatwiej indeksowane.
- Linki zewnętrzne

  Są to linki z innych stron (zwane również backlinkami), budują autorytet witryny i zwiększają jej widoczność.

3. Wyzwania w indeksowaniu

- Treści dynamiczne

  Strony generowane w czasie rzeczywistym są trudniejsze do indeksowania. Strona która będzie wolno się renderować może być pominięta w trakcie indeksowania. Dlatego należy dbać o wydajność kodu w javascripcie podczas pisania strony.
- Duplikaty treści

  Prowadzą do marnowania zasobów Googlebota na skanowanie tych samych informacji. Takie treści nie są kanoniczne i strona może zostać potraktowana jako spam.
- Problemy techniczne

  Wolne ładowanie strony, wspomniane wczesnień, brak atrybutów przy tagach typu img, błędy w pliku robots.txt lub brak mapy witryny mogą negatywnie wpłynąć na indeksowanie.

4. Narzędzia wspierające indeksowanie

a. Google Search Console

Google Search Console to darmowe narzędzie od Google, które pozwala właścicielom stron internetowych monitorować i zarządzać obecnością ich witryn w wynikach wyszukiwania Google. Jest to kluczowe narzędzie dla SEO, ponieważ dostarcza informacji o tym, jak Google indeksuje i wyświetla Twoją witrynę w wyszukiwarce.

- Pozwala monitorować, które strony zostały zaindeksowane.
- Informuje o błędach w plikach sitemap oraz problemach z dostępnością stron.
- Identyfikowanie błędów indeksowania, takich jak niedostępne URL-e.
- Zgłaszanie map witryn.

b. Lighthouse

Przeglądarka Chrome dostarcza narzędzie Lighthouse, pozwala ono na wygenerowanie raportu SEO, oraz identyfikacje błędów w strukturze kodu. Narzędzie dostarcza dane dotyczące: wydajności, ułatwień dostępu, sprawdzonych metod, oraz SEO. Dodatkowo można wygenerować raport zarówno dla urządzeń desktopowych i mobilnych.

5. Rekomandacje i dobre praktyki

1. Poprawa pliku robots.txt – Unikaj blokowania kluczowych treści.
2. Aktualizacja mapy witryny – Regularnie weryfikuj jej poprawność i zgłaszaj do Google Search Console.
3. Optymalizacja techniczna – Zadbaj o szybkość ładowania, responsywność i przyjazne URL-e.
4. Dynamiczne treści – Używaj technik prerenderingu lub serwerowego renderowania treści, aby ułatwić ich indeksowanie.

II. Podsumowanie

Projekt przedstawił przegląd procesu indeksowania przez Google, wskazując na znaczenie optymalizacji stron internetowych, zarówno pod kątem treści, jak i technologii. Dzięki

omówieniu etapów procesu, czynników wpływających na indeksowanie oraz narzędzi wspierających developerów, udało się zaproponować praktyczne wskazówki i rozwiązania dla poprawy widoczności witryn w wyszukiwarce.

III. Bibliografia

1. In-depth guide to how Google Search work: https://developers.google.com/search/docs/fundamentals/how-search-works?hl=pl
2. Organizing Informations: https://www.google.com/search/howsearchworks/how-search-works/organizing-information/
3. Co to jest skanowanie stron – crawling: **https://www.sempire.pl/co-to-jest-skanowanie-stron-crawling.html**
4. https://edgemesh.com/blog/how-google-crawls-your-website-and-indexes-your-pages
5. https://www.semrush.com/blog/google-search-console/