Consulenza SEO Google Ads Training Guide Blog About me Contatti



Sei qui: EVE Milano / Posizionamento Motori di Ricerca / Cos'è un Crawler e come funzionano gli spider?

Cos'è un Crawler e come funzionano gli spider?

Autore: Giovanni Sacheli Data di pubblicazione: 28/09/2023 Tempo di lettura: 12 minuti Commenti: 🖼



Introduzione

La guida che segue è stata concepita come risorsa completa per coloro che desiderano comprendere a fondo i meccanismi del crawling e gli spider nei motori di ricerca.

Comprendere <u>come funzionano i motori di ricerca</u> è imperativo per chi opera nel settore SEO e per chiunque desideri avere una presenza online efficace nei motori di ricerca. Attraverso questa guida, verranno fornite risposte dettagliate a domande comuni sul crawling e gli spider, illustrando come questi elementi siano fondamentali per il funzionamento dei motori di ricerca.

Cos'è un Motore di Ricerca?

Un <u>motore di ricerca</u> è un sistema software sofisticato che facilita la ricerca di informazioni su Internet. La sua funzione primaria è quella di **indicizzare una vasta quantità di dati web**, organizzarli in modo strutturato e fornire risultati rilevanti in risposta alle query degli utenti.

Questo processo si articola in diverse fasi chiave:

- Crawling: I motori di ricerca utilizzano programmi automatizzati chiamati crawler o spider per esplorare il web in modo sistematico e continuo. Questi crawler scandagliano le pagine web esistenti e nuove, seguendo i link da una pagina all'altra e raccogliendo dati sul contenuto di ciascuna pagina. Per i contenuti generati con Javascript avviene una fase supplementare di <u>rendering</u>, in modo da poter creare un <u>DOM</u> completo di tutti gli elementi presenti in pagina da poter indicizzare.
- Indicizzazione: Dopo il crawling, le informazioni raccolte vengono elaborate e
 indicizzate. L'indicizzazione implica l'organizzazione dei dati in strutture efficienti
 che permettono una rapida ricerca e recupero delle informazioni. Durante questa
 fase, i motori di ricerca creano un indice che associa termini e concetti chiave alle
 pagine web pertinenti.
- Processamento delle Query: Quando un utente inserisce una query, il motore di ricerca processa questa richiesta confrontando i termini della query con l'indice creato durante la fase di indicizzazione. Questo processo implica l'analisi semantica e sintattica della query per comprendere l'intento dell'utente e identificare le pagine web più rilevanti.
- Ranking: Dopo aver identificato un insieme di pagine web pertinenti, il motore di ricerca le ordina in base a vari criteri di ranking. Questi criteri possono includere la rilevanza del contenuto rispetto alla query, la qualità e l'autorevolezza del sito web, la freschezza del contenuto, la posizione geografica dell'utente, e molti altri fattori specifici del motore di ricerca in uso.
- Presentazione dei Risultati: Infine, i risultati ordinati vengono presentati all'utente in una pagina di risultati del motore di ricerca (SERP). Ogni risultato include solitamente un titolo, un URL e una breve descrizione o snippet del contenuto della pagina.

I motori di ricerca utilizzano algoritmi complessi e tecnologie avanzate come
l'apprendimento automatico e l'intelligenza artificiale per migliorare continuamente
l'efficacia e la precisione del processo di ricerca e recupero delle informazioni. L'obiettivo è
fornire risposte accurate e rilevanti che soddisfino le esigenze
modo rapido ed efficiente.

Questo sito o gli strumenti terzi da questo
utilizzati si avvalgnon di cookie necessari
al funzionamento ed utili alle finalità
illustrate nella cookie policy. Se vuoi
saperne di più o negare il consenso a tutti
o ad alcuni cookie, consulta la cookie
policy. Chiudendo questo banner,
scorrendo questo banner,
scorrendo questo banner,
scorrendo questo pagina, cliccando su un
linko proseguendo la navigazione in altra

Richiedi un preventivo

rmativa estesa CHIL

attraverso il quale vengono scoperte, scansionate nuove pagine web o aggiornamenti di pagine esistenti. Questa operazione è condotta da *agenti* software specializzati chiamati crawler o spider.

Ecco una descrizione più approfondita delle fasi e delle caratteristiche chiave del crawling:

- Inizio del Crawling: Il processo di crawling inizia con una lista di URL conosciuti, che
 potrebbero provenire da crawling precedenti o da sitemap fornite dai webmaster.
 Questi URL rappresentano il punto di partenza per gli spider.
- Esplorazione delle Pagine: Gli spider visitano questi URL, esaminando il contenuto delle pagine web, la struttura, i metadati e gli altri elementi cruciali come l'HTML, gli header, i tag, le immagini, e i video.
- Salvataggio dei Link: Durante l'esplorazione, gli spider salvano i link presenti nelle pagine, sia interni che esterni, e li seguiranno per scoprire ulteriori pagine. Questa è una caratteristica fondamentale che permette ai motori di ricerca di navigare la vasta rete di pagine interconnesse su internet.

- Rispetto delle Direttive: Gli spider dei MdR generalmente rispettano le direttive fornite dai webmaster, come quelle presenti nel file robots.txt o nei meta tag, che possono includere istruzioni su quali pagine scansionare o evitare.
- Ottimizzazione del Crawl Budget: I motori di ricerca hanno un budget di crawling che determina la frequenza e la profondità del crawling di un sito web. Gli spider sono programmati per utilizzare questo budget in modo efficiente, evitando pagine duplicate o contenuti irrilevanti.
- Aggiornamento Continuo: Il processo di crawling è continuo e ripetitivo, con gli spider che ritornano su siti precedentemente scansionati per identificare nuovi contenuti o aggiornamenti, assicurando così che l'indice del motore di ricerca rimanga aggiornato e rilevante.
- Rilevamento di Errori e Problemi: Durante il crawling, gli spider possono anche rilevare problemi come errori 404, pagine duplicate, o problemi di reindirizzamento, che possono poi essere segnalati ai webmaster attraverso strumenti come Google Search Console.

Il processo di crawling è fondamentale per il funzionamento dei motori di ricerca e per la visibilità online di siti web e contenuti. Una comprensione approfondita del crawling è essenziale per i professionisti SEO e i webmaster per ottimizzare la presenza online e garantire che i loro siti web siano facilmente accessibili e indicizzabili dai motori di ricerca.

Esempi di Crawler

- Googlebot: Googlebot è il crawler ufficiale di Google che esplora il web per indicizzare nuove pagine e aggiornamenti. Utilizza algoritmi sofisticati per determinare quali siti visitare, la frequenza delle visite, e il numero di pagine da recuperare da ciascun sito.
- Bingbot: Bingbot è il crawler di Bing che, similmente a Googlebot, raccoglie dati dal web per aggiornare l'indice di Bing.
- SEMrush Bot: SEMrush Bot è utilizzato dall'omonima piattaforma SEMrush per raccogliere dati che alimentano gli strumenti di analisi SEO e di concorrenza della piattaforma.

Come conoscere il nome di uno spider?

Conoscere il nome di uno spider (o user-agent) può essere utile per vari motivi, inclusa la gestione dei bot sul tuo sito web. Ecco alcune strategie per identificare il nome di uno spider:

Richiedi un preventivo

• Analisi dei Log del Server: I log del web server registrano tutte le richieste fatte al tuo sito web, inclusi gli accessi da parte degli spider dei motori di ricerca. Ogni voce nei log del server include l'User Agent, che identifica lo spider. Esempio di una voce del log del server. Nell'esempio che segue, l'user-agent è: Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html).

```
"-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
1523
```

• Integrazione di Software di Rilevamento Bot: Integra software o plugin specifici sul tuo sito che rilevano e registrano l'attività dei bot, fornendo dettagli sul nome dello spider e altre informazioni correlate.

Attraverso queste metodologie, sarà possibile identificare i nomi degli spider che interagiscono con il tuo sito web, permettendoti di gestire meglio le loro attività e di ottimizzare la tua strategia SEO.

Differenza tra Web Crawling e Web Scraping

Il Web Crawling e il Web Scraping sono due processi distinti. Il crawling è l'esplorazione sistematica del web per indicizzare le pagine, mentre lo scraping è l'estrazione di dati specifici da pagine web, spesso senza permesso e per scopi mirati.

Il Web Scraping è il processo di estrazione di dati da siti web utilizzando script o bot. A differenza del crawling, che indicizza il web, lo scraping raccoglie dati specifici. Per esempio, uno scraper potrebbe estrarre tutte le informazioni sui prodotti da un sito di ecommerce, come prezzi, descrizioni e immagini. Questi dati possono poi essere utilizzati per analisi di mercato, comparazione prezzi, o aggregazione di informazioni.

Gestione dei Bot per un Sito Web: Pratiche Consigliate

La gestione ottimale dei bot è cruciale per garantire una corretta indicizzazione del tuo sito web, non sovraccaricare inutilmente il web server e per mantenere un ambiente online sicuro ed efficiente. Ecco alcune strategie avanzate e dettagliate per la gestione dei bot sul tuo sito:

- 1. Implementazione del File Robots.txt: Il file robots.txt è il primo punto di contatto tra il tuo sito e i motori di ricerca. Configura accuratamente questo file per indicare ai bot quali sezioni del sito possono accedere e quali no. Assicurati che le direttive siano chiare e aggiornate, evitando blocchi inutili che potrebbero impedire l'indicizzazione o il rendering di pagine importanti.
- 2. Utilizzo dei Meta Tag Robots: Implementa i meta tag robots all'interno dell'head HTML delle pagine per fornire istruzioni specifiche ai bot su come interagire con singole pagine. Utilizza noindex per prevenire l'indicizzaz Questo sito o gli strumenti terzi da questo di cookie necessari prevenire il trasferimento di <u>PageRank</u> attraverso i link.

al funzionamento ed utili alle finalità illustrate nella cookie policy. Se vuoi saperne di più o negare il consenso a o ad alcuni cookie, consulta la cookie 3. Creazione e Manutenzione delle Sitemap XML: Fornisci u policy. Chiudendo questo banner, scorrendo questa pagina, cliccando su linko o proseguendo la navigazione in altra maniera, acconsenti all'uso dei cookie.

Richiedi un preventivo

attraverso gli strumenti per webmaster dei motori di ricerca.

- 4. Monitoraggio attraverso Strumenti per Webmaster: Utilizza strumenti come Google Search Console o Bing Webmaster Tools per monitorare l'attività dei rispettivi bot sul tuo sito ed identificare problemi di crawling o di indicizzazione.
- 5. Analisi dei Log del Server: Analizza regolarmente i log del web server per valutare l'attività dei bot sul tuo sito, identificare bot dannosi e capire come i motori di ricerca interagiscono con il tuo contenuto.
- 6. Blocco dei Bot Dannosi: Implementa misure di sicurezza come firewall o sistemi di rilevamento e blocco bot per prevenire accessi non autorizzati e attività dannose da parte di bot malevoli.
- 7. Implementazione di Verifiche Captcha: Utilizza sistemi di verifica come reCAPTCHA per distinguere tra traffico legittimo e bot dannosi, prevenendo attività malevole.
- 8. Manutenzione e Aggiornamenti di Sicurezza: Mantieni aggiornati tutti i sistemi, plugin e altre tecnologie utilizzate sul tuo sito per prevenire vulnerabilità che potrebbero essere sfruttate da bot malevoli.
- 9. Formazione e Consulenza Professionale: Mantieniti aggiornato sulle ultime minacce e sulle migliori pratiche per la gestione dei bot, e considera la consulenza di esperti

SEO o di sicurezza web per garantire una gestione efficace e professionale dei bot sul tuo sito.

La gestione accurata dei bot è essenziale per garantire che il tuo sito sia indicizzato correttamente, risponda efficacemente alle query degli utenti e rimanga protetto da attività online malevole.

Come si fa un Crawl

Il processo di crawling può essere effettuato sia dai motori di ricerca che manualmente, per scopi SEO o di analisi.

Esistono software che permettono di emulare il comportamente dei bot dei motori di ricerca sul nostro sito web. Strumenti come Screaming Frog ci aiutano ad analizzare e diagnosticare criticità sui nostri siti, eseguando la scansione delle pagine proprio come farebbe un motore di ricerca.

Lanciando la scansione possiamo analizzare il contenuto delle pagine e identificare vari elementi, come la struttura dei link o i problemi tecnici che potrebbero influenzare la visibilità del sito nei risultati di ricerca.

Cos'è il Crawl Budget

Il <u>crawl budget</u> è un concetto che si riferisce al **tempo** che un bot decide di trascorrere su un sito web scansionando le sue pagine. Gestire efficacemente il crawl budget è cruciale per garantire che le pagine più importanti del sito vengano crawlate e indicizzate in modo tempestivo.

Conclusioni

La comprensione del crawling e degli spider è fondamentale per chiunque operi nel campo della SEO o desideri ottimizzare la visibilità online. Attraverso una gestione efficace del crawling e una comprensione approfondita dei meccanismi dei Questo sito o gli strumenti terzi da questi utilizzati si avvalgono di cookie necessari al funzionamento ed utili alle finalità illustrate nella cookie policy. Se vuoi saperne di più o negare il consenso a tutti o ad alcuni cookie, consulta la cookie policy. Chiudendo questo banner, scorrendo questa pagina, cliccando su un link o proseguendo la navigazione in altra maniera, acconsenti all'uso dei cookie. possibile migliorare significativamente la presenza online e rag ampio.

Richiedi un preventivo

Come posso controllare quando il mio sito è stato scansionato l'ultima volta?

È possibile verificare l'ultima data di crawling di un sito attraverso gli strumenti per webmaster forniti dai motori di ricerca, come Google Search Console. Qui, nella sezione "Statistiche di scansione", si possono vedere le date di ultimo crawling per le pagine del tuo sito.

Posso impedire il crawling di alcune pagine del mio sito?

Sì, è possibile impedire il crawling di specifiche pagine utilizzando il file robots.txt con la direttiva disallow applicata a determinate sezioni del sito.

Cosa sono i meta tag robots e come influenzano il crawling?

I meta tag robots sono istruzioni inserite nell'header HTML di una pagina che indicano ai motori di ricerca se scansionarla o indicizzarla. Esempi comuni includono "noindex" per impedire l'indicizzazione e "nofollow" per impedire il seguimento dei link.

È possibile influenzare la frequenza di crawling del mio sito?

La frequenza di crawling può essere influenzata da vari fattori, come la qualità e la freschezza del contenuto, la struttura del sito e la velocità del server.

Cosa significa quando una pagina è "non trovata" durante un crawl?

Quando una pagina è segnata come "non trovata" durante un crawl, significa che il crawler non è stato in grado di accedere a quella pagina, spesso a causa di un errore 404.

Come può un sito web ottimizzare il suo crawl budget?

Ottimizzare il crawl budget può includere azioni come migliorare la velocità del sito, ridurre errori di pagina, strutturare il sito in modo logico e utilizzare il file robots.txt in modo efficace per guidare i crawler verso le pagine più importanti.

Cosa sono i crawler di terze parti e come differiscono dai crawler dei motori di ricerca?

I crawler di terze parti sono programmi automatizzati creati da entità esterne ai motori di ricerca per raccogliere dati dal web. Possono essere utilizzati per scopi vari, come la ricerca accademica o l'analisi del mercato (ad esempio Semrush bot e AhrefsBot), e possono non aderire alle stesse politiche di crawling dei motori di ricerca (nel senso che potrebbero non rispettare i tuoi disallow).

Come posso testare il mio sito per assicurarmi che sia facilmente crawlable?

Utilizzando strumenti di analisi SEO e crawling come URL inspector di Google Search Console, Screaming Frog o Sitebulb, è possibile eseguire crawl simulati per identificare eventuali problemi che potrebbero impedire un crawling efficace da parte dei motori di ricerca.

Qual è la differenza tra crawling, indicizzazione e ranking?

Il crawling è il processo di scoperta di nuove pagine o aggiorna l'indicizzazione è l'atto di aggiungere queste pagine all'indice de l'atto di aggiungere que all'indice de l'atto di aggiungere queste pagine all'indice de l'atto di aggiungere que all'in

Qual è il ruolo dei file sitemap.xml nel processo di crawling?

Le <u>sitemap.xml</u> aiutano i motori di ricerca a comprendere la struttura del sito e a trovare nuove pagine più rapidamente. Fornire una sitemap aggiornata è una best practice SEO che può facilitare un crawling più efficace e tempestivo del sito.

Richiedi un preventivo

Risorse Utili

In questa sezione trovi link a strumenti, guide, articoli e altre risorse utili per approfondire ulteriormente gli argomenti trattati nella guida:

- https://it.wikipedia.org/wiki/Crawler
- https://www.cloudflare.com/learning/bots/what-is-a-web-crawler/ <a href="https://www.cloudflare.com/learning/bots/what-i
- https://developers.google.com/search/docs/crawling-indexing/overview-googlecrawlers?hl=it ₽
- https://www.evemilano.com/valutare-indicizzazione/
- https://www.evemilano.com/indicizzare-sito-google/



Guide correlate a Knowledge

- o Linee guida SEO per siti di Affiliazione
- o SEO Skill, le competenze che un SEO dovrebbe avere
- o Significato di KPI Key Performance Indicators
- o Scopri il Significato di SEO: Si dice IL o LA?
- o Come funziona Google (Easy Mode)
- o Perché penso che Google ignori i fattori di usabilità del tuo sito
- o Le GARANZIE nella SEO non esistono
- o A cosa serve la cache di Google e come usarla
- o Come diventare un SEO professionista
- o Strategie per aumentare il traffico sul sito web

Autore

Mi chiamo Giovanni Sacheli, sono un consulente in Search Marketing (SEO e PPC) e lavoro nella mia agenzia Searcus Swiss Sagl. Co-autore del libro SEO Audit Avanzato, appassionato di **analisi SEO tecniche** e campagne con Google Ads. Dal 2010 condivido in questo blog guide SEO e nozioni tecniche per consulenti, webmaster e addetti ai lavori.

LASCIA UN COMMENTO

Questo sito o gli strumenti terzi da questo utilizzati si avvalgono di cookie necessari al funzionamento ed utili alle finalità al funzionamento ed utili alle finalità illustrate nella cookie policy. Se vuoi saperne di più o negare il consenso a tutti o ad alcuni cookie, consulta la cookie policy, Chiudendo questo banner, scorrendo questa pagina, cliccando su un linko proseguendo la navigazione in altra maniera, acconsenti all'uso dei cookie.

	Richiedi un preventivo	Informativa estesa	CHIUDI
Nome *			
Email *			

Invia commento

RICHIEDI UN PREVENTIVO

		Nome e Cognome
		Telefono
		Telerono
		La tua email aziendale
	:	Sito Web
	:	Sono interessato a
		□ Consulenza SEO
		☐ Consulenza Google Ads
		□ Corsi SEO
		□ Corsi Google Ads
	· ·	_ corsi doogle Aus
		Budget annuale per SEO e PPC
		Piccole e Medie Aziende: da 10.000€ a 2 💙
		Descrivi la tua richiesta
Richiedi un preventivo	Questo sito o gli strumenti terzi da questo utilizzati si avvalgono di cookie necessari al funzionamento ed utili alle finalità illustrate nella cookie policy. Se vuoi saperne di più o negare il consenso a tutti o ad alcuni cookie, consulta la cookie policy. Chiudendo questo banner, scorrendo questa pagina, cliccando su un link o proseguendo la navigazione in altra maniera, acconsenti all'uso dei cookie. Informativa estesa CHIUDI	Vorrei ricevere la newsletter mensile Acconsento al trattamento dei dati personali* Quale numero è maggiore: 5 o 7?
		Invia la richiesta * Campo obbligatorio
		CLASSIFICA POSIZIONAMENTO

CLASSIFICA POSIZIONAMENTO MOTORI DI RICERCA

♥ 83 <u>web</u>	Guida completa alla migrazione SEO di un sito
♥ 46 web	Ottimizzare i link interni e la struttura del sito
♥ 45	PageRank, scopriamo l'algoritmo più famoso

♥ 44 <u>Screaming Frog Guida Completa in italiano</u> professionale ♥ 36 Guida a Microsoft Excel per la SEO ♥ 35 SEO Checklist: i fattori SEO da Ottimizzare ♥ 28 Guida completa al Robots meta tag NOINDEX ♥ 25 <u>Indicizzazione su Google – significato ed</u> <u>analisi</u>

AGGIORNAMENTI IN QUESTA CATEGORIA

Cos'è un Crawler e come funzionano gli spider?

Come impedire a GPT l'accesso ai contenuti del tuo

Link pruning: analizzare e correggere i link in ingresso (backlinks)

COMMENTI

Giovanni Sacheli su <u>Analisi dei contenuti - Come</u> eseguire un Content Audit SEO

Carlo Di Somme su <u>Analisi dei contenuti - Come</u> eseguire un Content Audit SEO

Giovanni Sacheli su Guida all'ottimizzazione delle Breadcrumb per la SEO

Questo sito o gli atrumenti terzi da questo utilizzati si avvalgono di cookie necessari al funzionamento de utili alle finalità illustrate nella cookie policy. Se vuoi lustrate nella cookie policy. Se vuoi lustrate nella cookie policy. Se vuoi tutti o ad alcuni cookie consulta la cookie policy Chiudendo questo banner, scorrendo questa pagina, cliccando su un linko proseguendo la navigazione in altra maniera, acconsenti all'uso dei cookie.

Richiedi un preventivo

ARCHIVIO

Seleziona il mese

EVE MILANO BY SEARCUS

Sacheli Giovanni Consulente SEO e PPC Email: info(at)evemilano.com La società titolare di questo sito web è Searcus Swiss Sagl 🗗

Agenzia SEO e Google Ads Searcus Swiss SAGL

PARTNER GOOGLE ADS



Cos'è un Crawler e come funzionano gli spider?

UID: CHE-180.818.315

Num reg comm: CH-501.4.019.466-4

Amministrazione: Via dei Faggi 4V, Pazzallo - 6912

(CH

Ufficio: Via Penate 16, Mendrisio - 6850 (CH)

Email: giovanni(at)searcus.ch Mobile: <u>+39 3393668879</u> Tel: <u>+41 (0)79 8643035</u> Sei interessato ad una consulenza per la gestione di campagne Google Ads? Mi chiamo Giovanni Sacheli e sono membro del programma di Certificazione Google Ads Partner per garantire ai miei clienti competenze e professionalità.

SEMINARI SISTRIX.COM



Una volta al mese presento Sistrix, lo strumento SEO dei professionisti. Il seminario dura dalle 10:00 alle 13:30. La partecipazione è gratuita. Per informazioni clicca qui.

CONNECT

linkedin 🗗 - discord 🗗

 $Copyright © 2010 - 2023 \ EVE\ Milano\ Consulenza\ SEO\ Consulenza\ SEO\ by\ Searcus\ Swiss\ Sagl.\ Tutti\ i\ diritti\ riservati\ |\ Privacy\ |\ Sitemap$

Questo sito o gli strumenti terzi da questo utilizzati si avvalgono di cookie necessari al funzionamento ed utili alle finalità illustrate nella cookie policy. Se vuoi saperne di più o negare il consenso a tutto ad alcuni cookie, consulta la cookie policy, Chiudendo questo banner, scorrendo questa pagina, cliccando su un linko proseguendo la navigazione in altra maniera, acconsenti all'uso dei cookie.

Richiedi un preventivo

Informativa estesa