UNINOVE

O conceito de indexação de conteúdo na web

ENTENDER O CONCEITO DE INDEXAÇÃO, OS MOTORES DE BUSCA, O PRINCÍPIO DOS ROBÔS DE RASTREAMENTO NA WEB.

AUTOR(A): PROF. GERSON RISSO

O conceito de indexação, história dos motores de busca

O que é indexação? Qual é o significado dessa palavra?

Possivelmente essa palavra tem aplicações entre outras áreas, além da área de TI, na área econômica sempre ouvimos falar da tal indexação de preços ou algo semelhante a isso.

Na área de TI, o conceito é extensamente utilizado, mais precisamente para aplicações em armazenamento de arquivos em sistemas operacionais, em banco de dados. Em estruturas de dados, quando armazenamos dados em arrays, estamos associando os dados aos índices desse array. Portanto, estamos indexando os dados aos índices dessa estrutura de dados. A indexação facilita a ação de busca de dados em uma estrutura.

Segundo a definição do dicionário on line infopédia, que inclusive apresenta o verbete em duas áreas distintas, justamente nas áreas de economia e TI, indexação é um método de organização utilizada nas bases de dados, que permite maior facilidade na organização da informação.

Legenda: INDEXAçãO DE DADOS

Em termos de internet, o mundo web, há quantidade de conteúdo é incalculável, que só faz aumentar, em frações pequenas de tempo. Então, realizar a busca eficiente nesse conteúdo de forma eficiente é extremamente importante e a base de uma busca são justamente os sites de busca. Nesse contexto, a indexação é a uma estruturação a fim de organizar arquivos em uma base de dados. À cada item, várias tags são associadas ou outros aspectos, determinados pelos sites de busca, que identifiquem um arquivo, entre muitos (Google, 2018).

Do surgimento da internet, por volta de 1945 até os dias atuais, a evolução dos sites de buscas foi brutal. Antes dos sites de buscas, nos limitávamos a digitar endereços eletrônicos de sites conhecidos por nós e a partir daí iniciava-se a navegação na internet. E hoje já nem memorizamos endereços eletrônicos, basta

ir no site de busca e digitar as palavras chave daquilo que desejamos localizar (Google, 2018).

Mas, a história dos sites de busca não começa com o site da Google, no passado muito recente, o site da Google nem existia, quanto mais ser essa gigante empresa dos dias atuais.

Segundo Cendón, uma das ferramentas mais antigas é o Archie, ele buscava arquivos do tipo FTP (File Transfer Protocol), organizados em diretórios, com evolução da internet, outras ferramentas surgiram.

Os diretórios são arquivos estruturados hierarquicamente por categorias e subcategorias e o usuário pode percorrer essa estrutura, procurando o assunto nas várias categorias e subcategorias. Alternativamente aos diretórios, surgiram os motores de busca, a medida que a internet se expandia. Eles utilizam-se de robôs para percorrer a imensas bases de dados colecionadas pelos sites de busca (Cendón, 2001).

Os robôs interagem com usuários que realizam a busca na internet, no momento em que solicita uma busca, entram em ação os robôs. Eles são programas utilizados para rastrear a web, coletando informações relevantes, segundo critérios baseados em inteligência artificial, sobre os recursos que encontram. Esses softwares também são conhecidos como spiders crawlers. O início da busca ocorre através de uma lista de endereços e a partir desse ponto, executam a varredura de páginas pelos links definidos nos documentos. Eles verificam se determinado documento foi visitado pelo robô, anteriormente e em caso afirmativo, verifica se ocorreram atualizações nesse documento, se sim, realiza a indexação. As informações relevantes são indexadas em uma imensa base de dados que será percorrida utilizando um site de busca (Deters, 2003).

Legenda: ESTRUTURA DE UM FTP

Legenda: BUSCA NA WEB

As bases de dados dos sites de busca - Informações adquiridas na varredura da web, pelos robôs, são direcionadas aos indexadores que coletam informações dos documentos e sem seguida armazenam na base dados. A amplitude da base de dados confere maior popularidade do site de busca, uma vez que, define a extensibilidade da busca, portanto, quanto mais recursos de documento uma base de dados tiver, maior será as informações retornadas (Deters, 2003).

O Altavista foi lançado em 1995 e possibilitava ao usuário digitar uma frase ao invés de apenas palavras chave, conforme (Laudon, 2007), revolucionou a busca na internet porque os sites dee busca precursores eram índices com palavras chaves que ocorriam maior frequência nos sites visitados por eles e apresentava uma lista de links que não poderia não ser relevante à pesquisa realizada (Laudon, 2007).

Por volta de 1998, foi fundado o Google, esse site de busca usava estratégia distinta dos demais sites de pesquisa, porque além da indexação de palavras chaves dos documentos visitados, também criava uma lista de resultados da busca (PageRanking System). Esse ranking, determina, basicamente, a popularidade de sites através do cálculo de visitas que já ocorreram para aquele documento web. Os desenvolvedores do Google projetaram um robô que indexava, palavras chave e arranjos de palavras envolvendo autores, os títulos dos seus documentos (Laudon, 2007).

Segundo Laudon, o prcesso de busca do Google, funciona basicamente, da seguinte forma, o usuário faz a consulta através de um software de interface digitando as palavras da busca. Os servidores web recebem a requisição HTTP. O Google se utiliza de uma infra-estrutura enorme e complexa para administrar as consultas e retornar os resultados.

A consulta é enviada para servidores de índice do Google (Laudon, 2007) que basicamente, descrevem quais páginas contém as palavras-chave da consulta que fora realizada. E onde essas páginas estão localizadas dentro dos servidores. Usando recurso do PageRanking, descrito anteriormente, o sistema determina a popularidade de cada página. Através de cálculos complexos e como resultado, se tem as melhores páginas para consulta naquele caso (Laudon, 2007).

PARA ASSISTIR A HISTóRIA DOS BUSCADORES...

Acesse o link:

https://sites.google.com/site/historiasobreossitesdebusca/historia-site/indexacao (https://sites.google.com/site/historiasobreossitesdebusca/historia-site/indexacao)

Objeto disponível na plataforma

Informação:

GOOGLE - O SITE DE BUSCA MAIS UTILIZADO ATUALMENTE

Os motores de busca, segundo Cendón, criam arquivos invertidos utilizados para otimizar a busca de informações, nas bases de dados. Ou seja, utiliza palavras chave para indexar documentos.

Os índices contém termos da pesquisa de informação e a URL das páginas que cotém tais termos. Podem ser armazenados dados sobre as posições das palavras em um documento web, por exemplo no título ou no sub-título a fim de melhorar o retorno de informações relevantes (Cedón, 2001).

Os motores de busca indexam em seu índice apenas a URL, as palavras que ocorrem com maior frequência no documento web. Podem, incluir também elementos do documento web que não é visível. Como as metatags definidas no cabeçalho, contendo palavras chave. Os metagas de descrição da página. O texto alt da tag imagem, que se refere ao texto exibido caso a imagem não seja carregada.

```
1
2  <!DOCTYPE html>
3  <html>
4      <head>
5          <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
6          <title>JSP Page</title>
7      </head>
8      <body>
9          <h1>
10             Bem vindo ao JSP
11         </h1>
12     </body>
13 </html>
14
15
```

Legenda: UMA AMOSTRA DE UM ARQUIVO HTML

Os critérios para indexação de páginas de documentos é importante pois, sites de buscas otimizam esse tipo de tarefa, a fim de não armazenar todas as páginas, como regra geral. Alguns motores de busca incluem nas suas bases de dados, apenas a página principal. Motores podem indexar também, além do documentos HTML, documentos XML, PDF entre outros (Cendón, 2001).

Uma parte da web (deep web) não inclui páginas indexadas pelos mototres de buscas e, portanto, ficam invisíveis para os navegadores. Não são anexadas por várias razões, frames e páginas dinâmicas, por exemplo (Cendón, 2001). Os frames em páginas dificultam a indexação pelos motores de busca.

Páginas dinâmicas, segundo Cendón, apresentam dificuldades para os robôs, normalmente são executadas por conta de alguma requisição do usuário. Segundo, Cendón, essas páginas se caracterizam-se por conter um ponto de interrogação na URL, os robôs evitam esses endereços, para evitar indexar muitas páginas semelhantes, com URL diferentes.

Legenda: CRITéRIOS DE INDEXAçãO

Um site WordPress pode ser indexado na busca do Google através do Googlebot. Ele é o robô de rastreamento do Google, percorre a web para coletar informações dos sites visitados, a fim de indexá-los nas bases de dados do Google. Depois que a informação foi coletada pelo robô do Google, essa informação é processada (indexada). Nesse processo que a qualidade do conteúdo é determinada para ser ranqueadas adequadamente nas páginas do Google. O Googlebot começam a pesquisa a partir das páginas visitadas anteriormente e em seguida verifica se ocorreram atualizações ou se é uma página nova.

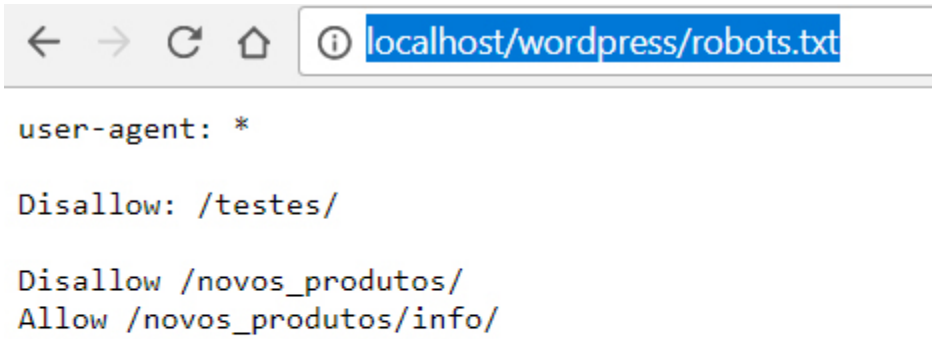
No WordPress, os motores de busca acessam o arquivo yourdomain.com/robots.txt a fim de obter informações sobre os sites Os motores de busca lêem um arquivo em robots.txt para obter informações sobre o que os robôs Bing, Yahoo, Google etc, mas, conforme documentação oficial do WordPress, alguns robô ignoram esse arquivo. Com base nesse arquivo, os robôs são informados do que podem ou não fazer dentro do seu site. Pois, o primeiro arquivo que eles vão ler é justamente o arquivo robotstxt. Na falta desse arquivo, os robôs indexam todo o conteúdo site.

A seguir é apresentado um arquivo típico do arquivo robots.txt. Que pode ser ecrito em um bloco de notas. Salvando-o com o nome robots.

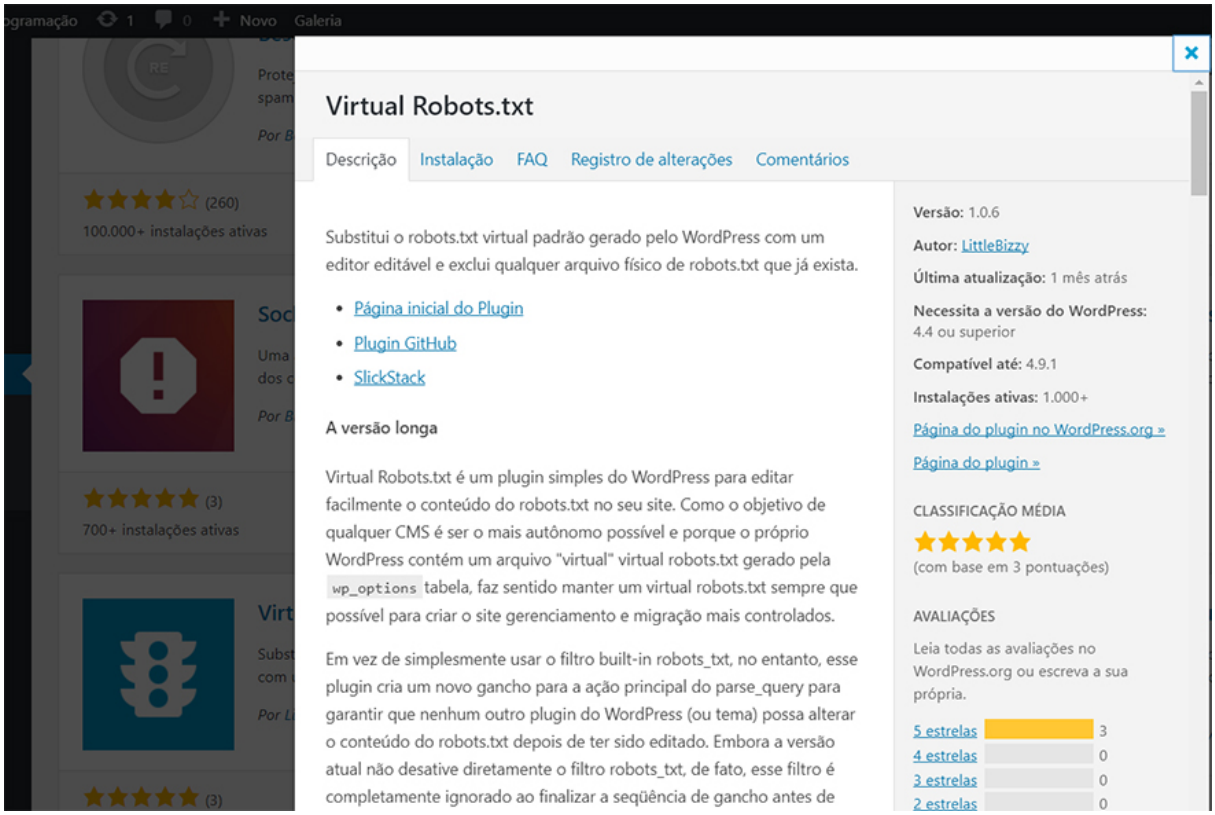
PARA MAIS INFORMAçõES SOBRE O GOOGLEBOT...

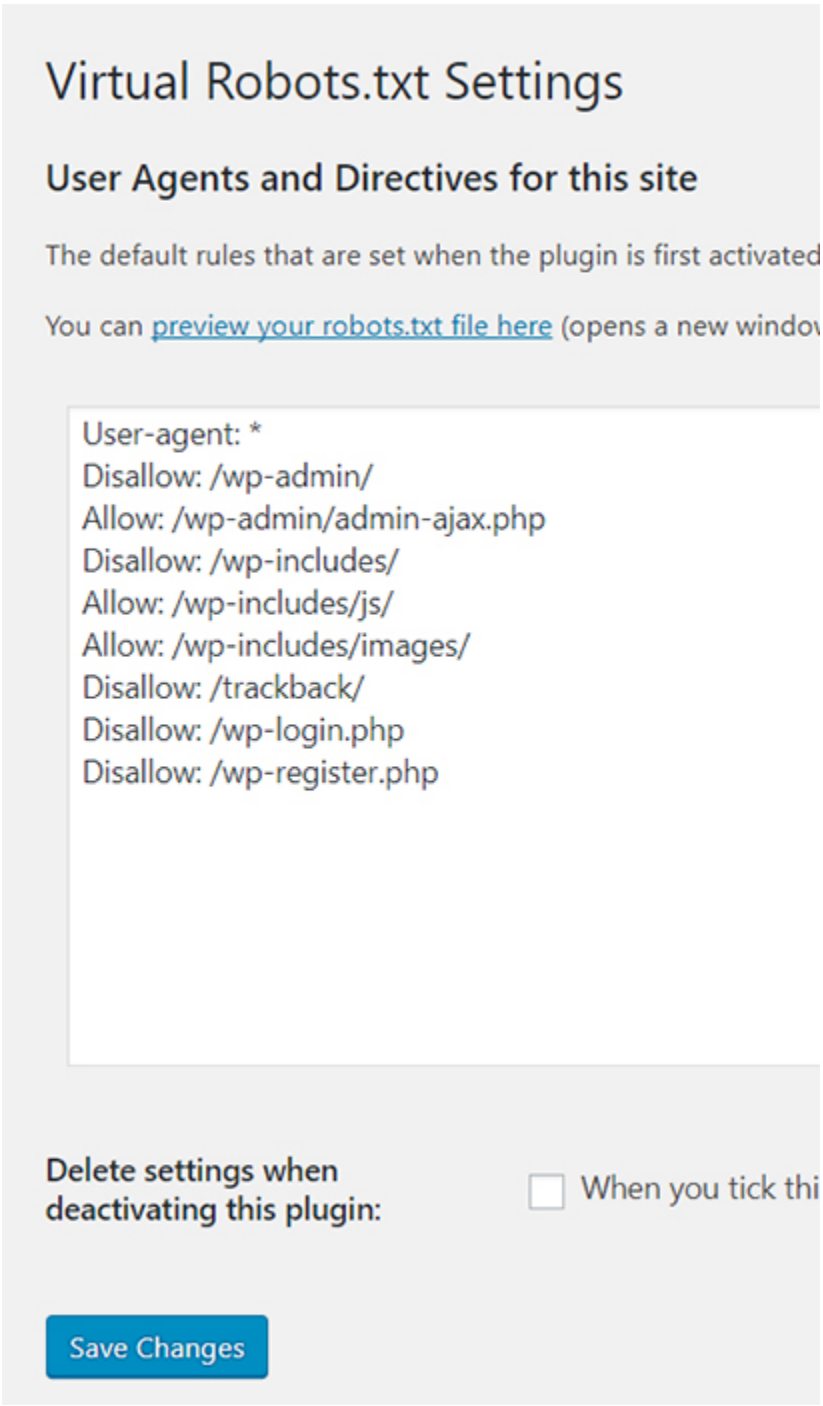
> Acesse o link:
> https://support.google.com/webmasters/answer/182072?hl=pt-BR
> (https://support.google.com/webmasters/answer/182072?hl=pt-BR)

```
1. user-agent: *
2.
3. Disallow: /testes/
4.
5. Disallow /novos_produtos/
6. Allow /novos_produtos/info/
```



Na linha 1 estamos informando que qualquer robô que acesse o site deve seguir as regras definidas na sequência.
Na linha 3 definimos que há uma pasta denominada teste que não deve ser indexada pelo robô de rastreamento.
Na linha 5 definimos também que não é permitido a indexação da página novos_produtos. Porém, pela regra de baixo, é permitido que od robôs de busca acessem a pasta info contido na pasta novos_produtos.
Salve o arquivo na pasta dedica ao WordPress e para acessar o arquivo, digite http://localhost/wordpress/robots.txt. Então o arquivo robots.txt será exibido na tela. Na verdade ele é um arquivo de livre acesso, em qualquer você pode digitar o endereço e no final inserir o nome robots.txt.


Virtual Robots.txt
Descrição
Instalação
FAQ
Registro de alterações
Comentários
Substitui o robots.txt virtual padrão gerado pelo WordPress com um editor editável e exclui qualquer arquivo físico de robots.txt que já exista.
Página inicial do Plugin
Plugin GitHub
SlickStack
A versão longa
Virtual Robots.txt é um plugin simples do WordPress para editar facilmente o conteúdo do robots.txt no seu site. Como o objetivo de qualquer CMS é ser o mais autônomo possível e porque o próprio WordPress contém um arquivo "virtual" virtual robots.txt gerado pela wp_options tabela, faz sentido manter um virtual robots.txt sempre que possível para criar o site gerenciamento e migração mais controlados.
Em vez de simplesmente usar o filtro built-in robots_txt, no entanto, esse plugin cria um novo gancho para a ação principal do parse_query para garantir que nenhum outro plugin do WordPress (ou tema) possa alterar o conteúdo do robots.txt depois de ter sido editado. Embora a versão atual não desative diretamente o filtro robots_txt, de fato, esse filtro é completamente ignorado ao finalizar a seqüência de gancho antes de
Versão: 1.0.6
Autor: LittleBizzy
Última atualização: 1 mês atrás
Necessita a versão do WordPress: 4.4 ou superior
Compatível até: 4.9.1
Instalações ativas: 1.000+
Página do plugin no WordPress.org »
Página do plugin »
CLASSIFICAÇÃO MÉDIA
(com base em 3 pontuações)
AVALIAÇÕES
Leia todas as avaliações no WordPress.org ou escreva a sua própria.
5 estrelas 3
4 estrelas 0
3 estrelas 0
2 estrelas 0



É possível gerar o arquivo robots.txt através plugins: Virtual Robots.txt gera o arquivo com algumas regras básicas, mas que você pode configurar também.

Vá no menu Plugins e adicionar novo. Digite no campo de pesquisa o nome do plugin e clique em instalar, na sequência clique em Ativar plugin.

Após esse processo, clique em Plugins instalados, localize o plugin e clique em settings.

VEJA UM TUTORIAL PARA…

Criar e configurar um arquivo robots.txt

(ht
tp:
//s
av
efr
o
m.
ne
t/?
url
=h
ttp
s
%
3A
%
2F
%
2F
yo
ut
u.
be
%
2F
Qk
nii
qX
VE
jw
&
ut
m
_s
ou
rc
e=

ch
a
m
ele
on
&
ut
m
_
m
ed
iu
m
=e
xt
en
si
on
s
&
ut
m
_c
a
m
pa
ig
n=
lin
k_
m
od
ifi
https://youtu.be/QkniiqXVEjw (https://youtu.be/QkniiqXVEjw)er)

NESSE TóPICO VOCê APRENDEU...

O conceito de indexação de arquivos e dados.

Os processos de busca na web.

A história dos buscadores.

Sobre o robô de busca do Google.

A criar e configurar o arquivo robots.txt no WordPress.

ATIVIDADE

Analise as afirmações a seguir e assinale a alternativa correta:

I) Indexação é um conceito muito específico da web.

II) Indexação é um conceito específico de TI.

III) É um conceito amplo, aplicado em áreas distintas do conhecimento.

A. Somente a afirmação I é verdadeira.

B. Somente a afirmação II é verdadeira.

C. Somente a afirmação III é verdadeira.

D. Todas as afirmações estão erradas.

ATIVIDADE