

Motori di ricerca

Search...

- Non confondiamo i motori di ricerca (search engine) con le funzioni di ricerca all'interno dei siti.
- Le funzioni di ricerca all'interno sono in genere costituite da "script" che permettono di fare ricerche testuali nei file che compongono il sito...
- Per siti più strutturati le informazioni sono mantenute mediante particolari organizzazioni di dati: human-powered directories...

Human-powered directories

- Permettono di:
 - organizzare i documenti all'interno dei siti in modo strutturato
 - Associare ai documenti un "abstract" e una serie di parole chiave
- Allo scopo di facilitare la reperibilità dei documenti.

Alcuni prodotti

- Search server Hummingbird
- Oracle
- Look Smart

Open Directory Project

- E' una organizzazione che fa il censimento dei siti che spontaneamente si iscrivono e fornisce informazioni su tali siti.
- Crea manualmente gli elenchi dei siti Internet

crawler

- I crawler-based search engine invece creano i loro elenchi in modo automatico.
- Navigano nei siti e catturano le informazioni...

Search engine misti

- Alcuni motori di ricerca sono di tipo misto, ovvero usano alcune informazioni statiche e alcune informazioni reperite dinamicamente mediante crawlers.
- Ad esempio MSN Search che ha la parte statica gestita mediante Look Smart e la parte dinamica gestita mediante il crawler Inktomi.

Crawler-based search engine

- Basati su tre parti:
 - Lo spider/crawler
 - L'indice
 - Il search engine

Lo spider

- Visita i siti web
- Segue tutti i link presenti nelle pagine
- Esegue questa operazione periodicamente sui siti (ogni mese o due)

L'indice

- L'indice è come un grande elenco che contiene tutte le pagine visitate dallo spider
- L'elenco viene mantenuto localmente
- A volte succede che lo spider reperisce nuove pagine web, ma queste pagine non vengono inserite nell'indice immediatamente: solo quando saranno nell'indice saranno disponibili per il search engine.

Search engine

- Esegue le ricerche sulle pagine memorizzate nell'indice
- Associa ad ogni pagina un "punteggio" (rank) in funzione di quello che viene ritenuto **rilevante** sulla pagina ai fini della ricerca eseguita.
- A volte il punteggio viene espresso mediante una percentuale (100%, 75%...)

Metodi per attribuire il punteggio

- Localizzazione e frequenza delle parole chiave
 1. Parole presenti nel titolo della pagina (html title tag)
 2. Frequenza delle parole chiave nelle prime righe
 3. Frequenza delle parole chiave nelle pagine

I metodi “segreti”

- I metodi di localizzazione e frequenza vengono usati un po' da tutti i search engine, ma ognuno da risultati diversi.
- Ogni search engine ha un suo algoritmo speciale per determinare il ranking e generalmente questi algoritmi sono segreti e possono determinare il successo di un motore piuttosto che un altro.

I metodi “segreti”

- Periodicità di indicizzazione.
- Penalizzazione delle pagine che fanno “spamming” delle parole chiave
- Eliminazione di alcune pagine che sono strutturate in modo da sembrare più importanti di quello che in realtà sono (pagine con molti falsi links...)
- Misura di quali pagine vengono scelte dai visitatori che fanno le ricerche (clickthrough)

Crawling	Yes	No	Notes
Deep Crawl	AllTheWeb, Google, Inktomi	AltaVista, Teoma	
Frames Support	All	n/a	
robots.txt	All	n/a	
Meta Robots Tag	All	n/a	
Paid Inclusion	All but...	Google	
Full Body Text	All	n/a	Some stop words may not be indexed
Stop Words	AltaVista, Inktomi, Google	FAST	Teoma unknown
Meta Description	All provide some support, but AltaVista, AllTheWeb and Teoma make most use of the tag		
Meta Keywords	Inktomi, Teoma	AllTheWeb, AltaVista, Google	Teoma support is "unofficial"
ALT text	AltaVista, Google, Teoma	AllTheWeb, Inktomi	
Comments	Inktomi	Others	

Caratteristiche

- Deep crawl: lo spider/crawler naviga in profondità nel sito e porta via molte pagine.
- Frames support: capacità di navigare un sito anche se è organizzato in frames.
- Robot.txt è un mezzo per tenere lontani i search engines, specifica una politica di accesso per i robots (crawlers) e permette loro di accedere solo a una pagina del sito, quindi non possono navigare.

Caratteristiche (2)

- Meta tags:

```
<HEAD>
```

```
<TITLE>Il mondo dei francobolli</TITLE>
```

```
<META name="description" content="Tutto  
quello che avreste voluto sapere sui  
francobolli, dal prezzo alla storia >
```

```
<META name="keywords"  
content="francobolli, collezione francobolli,  
storia dei francobolli, prezzo dei francobolli,  
francobolli da comprare>
```

```
</HEAD>
```

Caratteristiche (3)

- Inclusione a pagamento
 - Molti search engine offrono un programma a pagamento per mantenere il sito nell'indice...
 - O addirittura posizionato nelle ricerche (google fa vedere sulla destra i link sponsorizzati)

Caratteristiche (4)

- Full body text
 - Tutti i maggiori search engine dichiarano di indicizzare la parte visibile del testo (non i meta tags per intenderci)
 - Inoltre vengono prese le parti iniziali delle pagine, ad esempio google prende solo i primi 101K delle pagine lunghe.

Caratteristiche (5)

- Stop words
 - Sono le parole che non hanno peso nella ricerca (di a da con che su giù...), alcuni search engine non indicizzano le stop words
- Meta description, Meta keywords
- Alt text comments: commenti associati alle immagini