

DIogene

Il motore di ricerca multisito configurabile

1. Cos'è Diogene?

Diogene è il motore di ricerca sviluppato dal Cineca. Le informazioni vengono recuperate grazie a un software, *robot* o *spider*, che percorre continuamente lo spazio web dei **siti prescelti** e, in modo **completamente configurabile**, riporta solo le pagine di interesse. Le informazioni acquisite dallo spider sono gli elementi necessari alla costruzione dell'indice che contiene tutti i termini significativi e le url delle pagine in cui sono presenti.

Diogene può rispondere alle richieste degli utenti secondo due modalità:

- *distribuita*: il motore di ricerca reperisce il documento direttamente presso il sito che lo ha pubblicato.
- *centralizzata*: all'utente viene fornita una copia locale mantenuta direttamente dal motore di ricerca permettendo di evitare possibili problemi di connessione.

2. Struttura di Diogene

Diogene è così strutturato:

■ **Spider**: è il modulo incaricato di visitare periodicamente un elenco di siti predefiniti (collezioni) alla ricerca di informazioni sulle pagine. Le URL dei siti di interesse vengono specificate grazie a una interfaccia web di amministrazione. Per ognuno dei siti indicati è possibile specificare quali informazioni prelevare attraverso un file di configurazione il cui mantenimento, così come il funzionamento di Diogene, può essere:

- *distribuito*: lo spider obbedisce alle istruzioni contenute in un file di configurazione residente presso il sito di interesse.
- *centralizzato*: è lo stesso amministratore del sistema che scrive il file di configurazione. Una volta lanciato, lo spider punta ad un gateway locale sul quale sono mantenuti i file di configurazione per tutti i siti da considerare. Successivamente, il reperimento dei documenti avviene in maniera analoga al funzionamento distribuito.

Il file di configurazione rispetta lo standard robots.txt, utilizzato dai motori di ricerca Standard Internet.

L'accesso alle funzionalità dello spider è completamente integrato con **Ianus**, la tecnologia sviluppata dal CINECA che assicura il controllo e la gestione dei servizi erogati via internet.

■ **Inserimento documenti**: una volta che lo spider ha terminato la procedura di reperimento e caching dei file, l'inserimento nell'indice del motore di ricerca avviene in base ai seguenti criteri:

- ogni documento nuovo viene direttamente inserito
- se già presente nel database, il suo contenuto viene confrontato con quello della copia nuova in base ad alcuni parametri specifici. Se è uguale il documento viene ignorato altrimenti ha luogo l'aggiornamento.

Conclusa la procedura di inserimento, i documenti vengono mantenuti se **Diogene** lavora in maniera centralizzata altrimenti, nel caso di funzionamento distribuito, vengono conservate solo le informazioni strettamente necessarie.

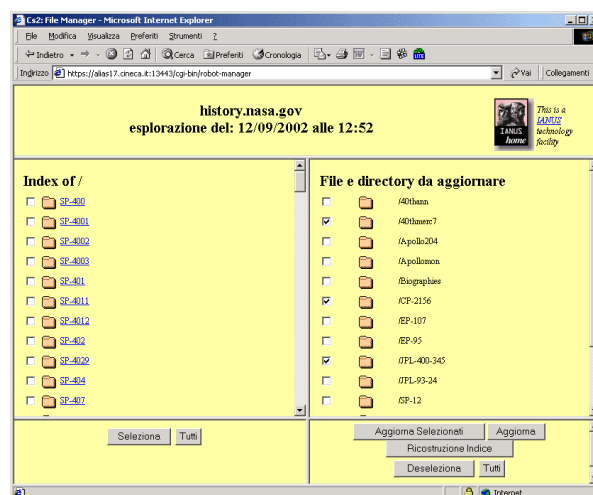


Figura 1 - Interfaccia web per la navigazione e la scelta dei documenti di un dominio da indicizzare

■ **Ricerca**: l'interfaccia web del motore di ricerca è personalizzabile. Generalmente offre all'utente due vie per reperire le informazioni desiderate: una modalità "semplice" e una "avanzata".

Il metodo più semplice dà la possibilità di scegliere se cercare le informazioni solo all'interno del sito principale o se inglobare nell'area di ricerca anche tutti i siti ad esso collegati. In seguito è sufficiente compilare il campo di testo con i termini oggetto della ricerca e, dopo averla avviata, attendere la pagina con i risultati ottenuti grazie al lavoro del motore.

La modalità "avanzata" permette di istruire il motore di ricerca affinché, attraverso indicazioni più precise come, ad esempio, distanza massima, sequenza parole, frase esatta, ecc., i risultati vengano selezionati più efficacemente.

In entrambi i casi è poi possibile intervenire sul modo in cui i risultati vengono presentati all'utente specificando un parametro di ordinamento dei risultati (per rilevanza, per data, in ordine alfabetico, per frequenza di consultazione).



Figura 2 - Diogene per Normeinrete

La Figura 2 mostra un esempio di interfaccia di uno dei motori di ricerca sviluppati dal CINECA. Il layout grafico è gestito completamente dalla tecnologia **Backstage Director**, lo strumento di *Web Content Management* del CINECA per la creazione e la manutenzione dei siti in modo facile ed immediato.

■ **Trattamento della meta-informazione:** è possibile creare un'interfaccia di ricerca specifica che consenta di sfruttare la meta-informazione estratta dai documenti, oltre a ricercare per testo libero. Il trattamento della meta-informazione è configurabile in funzione della tipologia di rappresentazione scelta: testo strutturato, marcatura XML o altro.

All'interno del file di configurazione si può indicare dove si trovano i campi che rappresentano la meta-informazione ed il loro significato. Dopo l'estrazione, l'informazione viene inserita in banca dati, l'indice viene aggiornato ed il motore è immediatamente utilizzabile per le ricerche.

3. Configurazione del motore di ricerca

E' possibile configurare in remoto il funzionamento degli aspetti principali del motore di ricerca in modo da variare il suo comportamento dinamicamente:

- attivare o disattivare la ricerca sui titoli dei documenti e decidere il tasso di prevalenza delle chiavi di ricerca presenti nel titolo rispetto al resto del testo del documento.
- decidere se mostrare il peso di ciascuna delle chiavi della ricerca che viene eseguita per stabilire quali siano più selettive e quali meno.

- configurare la lunghezza dell'abstract e delle frasi che compongono ciascuno dei risultati da presentare.

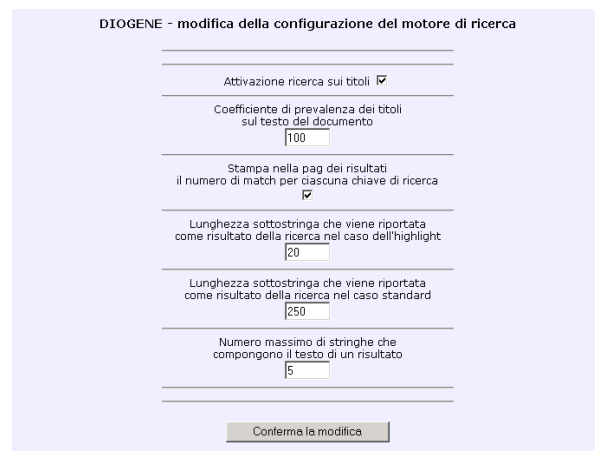


Figura 3 - Interfaccia web per la configurazione dell'output dei risultati e del calcolo della rilevanza

Infine, è possibile distinguere insieme rappresentativi di documenti che compongono il dominio entro il quale avviene la ricerca. Tali insiemi vengono chiamati *collezioni*.

Per ciascuna di esse si può specificare un fattore che influenza l'importanza della collezione in modo tale da aumentare a piacimento la visibilità di alcuni insiemi e diminuire quella di altri.

Diogene 2 - Modifica del peso delle collezioni

Collezione	Codice	Bonus	
axpbib.pd.infn.it	2	1	modifica
cisas.unipd.it	3	1.1	modifica
dmp.unipd.it	5	1	modifica
dpg.psy.unipd.it	6	1.1	modifica
dpss.psy.unipd.it	7	1.2	modifica
dssp.scipol.unipd.it	8	1	modifica
esterni.www.unipd.it	100	0.5	modifica
fac.psy.unipd.it	9	1	modifica
giove.pd.astro.it	16	0.9	modifica
infostudent.scform.unipd.it	10	1.6	modifica
mercurio.cheg.unipd.it/impianti	11	1.1	modifica
pdmecc8.mecc.unipd.it	17	1	modifica
www-fog.bio.unipd.it/ebs	61	1	modifica
www-tiresias.bio.unipd.it	62	1.2	modifica

Figura 4 - Tabella che permette di assegnare a ciascuna collezione un valore al fattore di rilevanza

Attribuzione dei bonus alla rilevanza per ciascuna collezione

Il valore di default è 1. Una collezione con bonus pari a uno non riceve alcuna modifica alla rilevanza dei suoi documenti che hanno soddisfatto i criteri di ricerca. E' possibile però specificare valori maggiori o minori in modo da aumentare o diminuire la possibilità che documenti appartenenti alla collezione compaiano prima o dopo gli altri nell'ordine di presentazione dei risultati.