***** QUESTO DOCUMENTO E' INCOMPLETO ED IN FASE AMPLIAMENTO CONTINUO. *****

PREFAZIONE

Questo progetto nasce perchè mi sono reso conto dell'importanza che i motori di ricerca hanno su Internet. Internet rappresenta una miniera infinita di informazioni, ma come tutti ben sappiamo, non sempre avere troppe informazioni significa avere un aiuto, infatti quando dobbiamo risolvere un problema oppure stiamo cercando qualcosa, ci sono sostanzialmente 3 casi che si possono verificare:

1 Non abbiamo alcuna informazione, e non sappiamo come trovarle.
2 Abbiamo una marea di informazioni, sappiamo che la nostra informazione è li, ma non sappiamo come recuperarla. (il classico problema dell'ago nel pagliaio).
3 Non abbiamo bisogno di nessuna informazione, perchè abbiamo già risolto il problema.

Escludendo il caso 3, negli altri 2 casi si capisce subito come, pur disponendo di una risorsa come Internet, non sempre siamo in grado di utilizzare tale risorsa, perchè non sempre siamo in grado di selezionare le informazioni che abbiamo a disposizione, oppure non sempre siamo in grado di trovare delle informazioni.

In questi 2 casi ci vengono in aiuto i così detti "Motori di Ricerca", che possono essere di vario tipo, e che possono aiutarci in differenti casi, una cosa però che accomuna tutti i motori di ricerca attuali, è comunque in pratica l'incapacità di offrire una soluzione globale al problema della ricerca su Internet, tanto è vero che si stima che il più potente motore di ricerca per il web attuale(Google), non è in grado di accedere al 100% delle risorse web(pagine http), pubblicate su Internet, e anche mettendo insieme i risultati dei maggiori motori di ricerca, si può arrivare ad un 60%, quindi resta sempre fuori un 40% di informazioni. Ma non finisce qui il problema, perchè questi motori, solitamente si occupano solo del web, ci sarebbero poi gli altri servizi Internet, come newsgroup(Google offre anche un servizio di ricerca per i newsgroup), irc, o altri, e poi i contenuti di database non accessibili dagli agenti dei motori di ricerca.

Un'altro problema particolarmente importante è il problema della freschezza delle informazioni, ovvero i motori di ricerca grossi, e generici, solitamente impiegano giorni e alcuni addirittura mesi e mesi per completare un ciclo di indicizzazione, per poi offrire la possibilità di cercare agli utenti, questo tempo rende a volte vecchie ed inutili alcuni tipi di informazioni, come le news, oppure come le offerte di lavoro, gli appalti pubblici, i concorsi!

E poi c'è il problema che i grossi motori di ricerca, potrebbero favorire alcune informazioni a scapito di altre, volontariamente oppure involontariamente, a chi non è capitato di cercare documentazione tecnica e trovarsi di fronte, centinaia di migliaia di links di università, dove però la documentazione tecnica cercata non è presente, ma sono presenti i programmi con i titoli che parlano dell'argomento sul quale si è interessati?

L'utente dovrebbe poter scegliere di scartare i siti che vuole, o di selezionare i siti che desidera, è per questo motivo che ho deciso di cominciare a creare un modello nuovo di "Motore di Ricerca", non più un "Motore di Ricerca" globale, ma un "Motore di Ricerca" totalmente sotto il controllo dell'utente, che così può scegliere autonomamente, come gestire i risultati delle ricerche, cosa filtrare e cosa prendere in considerazione.

INDICE DEI CAPITOLI

1) COME FUNZIONA UN MOTORE DI RICERCA

Ci sono vari tipi di motori di ricerca, ognuno con le sue caratteristiche, e con le sue funzionalità, in questo documento mi occuperò dei motori di ricerca più comunemente conosciuti, ovvero quelli per trovare informazioni sul web, tramite chiavi di ricerca, oppure parole chiavi. Generalmente si possono divedere i motori di ricerca in meta motori, e motori di ricerca veri, entrami a loro volta si possono dividere in base al tipo di database sul qualse si basano, ci sono i motori di ricerca che si appoggiano a database relazionali, e motori di ricerca per indici, le differenze sono grandi, in questo documento manuale, illustrerò come utilizzare questo software creato da me, per creare entrambi i modelli di motore. Per il motore di ricerca con database relazionale, si farà riferimento ad un database MySql, ma non è importante, perchè teoricamente potrebbe essere cambiato con qualsiasi altro database.

1.1) MOTORE DI RICERCA VAMPIRO

Questa tipologia di motori di ricerca, non sono altro che semplici coperchi, che sfruttano semplicemente altri motori di ricerca, filtrando poi i risultati non ritenuti rilevanti, ed ordinando i risultati a proprio piacimento, questi motori non verranno trattati in questo manuale, perchè non sono devi veri e propri motori di ricerca, e sono dal mio punto di vista di scarso interesse.

1.2) MOTORE DI RICERCA CHE SI BASA SUI META TAGS

Questi motori di ricerca, sono motori che si basano su informazioni date dagli utenti, vengono principalmente utilizzati nei casi dove non è possibile gestire tutta l'infrastruttura di un normale motore di ricerca, per ragioni di risorse, o di tempo, ma si vuole lo stesso avere la possibilità di effettuare delle ricerche sugli indirizzi web raccolti. Questi motori non sono il massimo dell'affidabilità, tuttavia si possono comunque creare dei servizi di directory o di ricerca utili utilizzando questo modello di motore di ricerca, il software da me offerto nella versione PROFESSIONAL ed ENTERPRISE, mette a disposizione uno script completo per gestire una directory con ramificazioni, ed un motore di ricerca che usa i META TAGS, ed altre informazioni, sfruttando naturalmente il mio Spider per recuperare le informazioni sui siti.
Vedere i paragrafi successivi per imparare ad impostare lo Spider per lavorare con questo tipo di motori, in ogni caso ripeto che è necessaria la versione Professional o Enterprise del mio software per avere lo script in Php/MySql che si può interfacciare con lo Spider, e che permette di gestire una directory di links. IL mio spider può catturare TITOLO, e i meta tags KEYWORDS e DESCRIPTION, nel caso il TITOLO o i META non fossero nella pagina, è possibile impostare lo spider per scartare la pagina, oppure per recuperare del testo dalle pagine successive, oppure dal resto della pagina, per riempire gli appositi spazi.

1.3) MOTORE DI RICERCA CHE SI BASA SUL CONTENUTO DI TUTTA LA PAGINA

Questo tipo di motori, non prendono in considerazione solo i Meta Tags, ma tutto il contenuto della pagina, ovvero Titolo, Meta Tags, i campi Alt, i links, il testo, ecc.. ecc..
Questo tipo di motori può funzionare sia con database relazionali, sia con altri tipi di database, ma solitamente si usano database che si basano sugli indici delle parole.
IL mio spider può funzionare con entrambe le soluzioni se opportunamente impostato. Parliamo comunque del metodo degli indici delle chiavi, sul quale si basano i grossi motori di ricerca, e che consiste nel creare degli indici per ogni parola trovata nel testo recuperato dagli spider. In particolar modo in questo tipo di motore di ricerca, diventano rilevanti tre componenti, ovvero tre moduli, che sono lo SPIDER, l'INDICIZZATORE, ed infine il CERCATORE.

1.4) SPIDER, COSA E' E A COSA SERVE

Questa componente, che è la componente principale dei motori di ricerca grossi, è un software che ha lo scopo di recuperare i documenti web, o pagine web, recuperando nuovi indirizzi (secondo una politica scelta dal motore di ricerca) e proseguendo il suo lavoro di recupero di documenti, e di nuovi links da seguire, a seconda della politica di spidering, uno spider potrebbe seguire una lista di siti, oppure una lista di siti con tutte le sotto pagine, oppure potrebbe proseguire all'infinito, trovando sempre nuov indirizzi da seguire, oppure in base ad altre regole. IL mio spider ha diverse opzioni che permettono di scegliere diverse politiche di SPIDERING, alcune incompatibili fra loro, e altre che si possono combinare, ma parlerò nei successivi parafrafi dedicati alla configurazione, su come impostare alcune politiche di SPIDERING. Lo spider una volta raccolta una pagina, ne fa il PARSING, raccoglie quello che deve raccogliere, ovvero i nuovi links, e passa il resto all'INDICIZZATORE, oppure passa la pagina senza PARSING, dipende dalle decisioni che si vogliono prendere. IL mio spider per il momento, passa all'INDICIZZATORE, una pagina sulla quale è già stato fatto il PARSING, in futuro aggiungerò altre opzioni per personalizzare questo, ed aggiungerò la possibilità di fare l'INDICIZZAZIONE sulla pagina originale.

1.5) INDICIZZATORE, COSA E' E A COSA SERVE

Questa componente, è la componente più importante nei motori di ricerca che si basano sugli indici, ha lo scopo di organizzare gli indici per le parole chiavi, in pratica deve estrapolare le parole dal testo, creare degli indici, ci sono sostanzialmente 2 possibilità, o si indicizza una pagina già pulita (in questo modo lavora il mio motore di ricerca), oppure si indicizza la pagina originale per prendere in considerazione anche alcuni tag nella pagina (si suppone che Google lavora Google in questo modo). IL mio INDICIZZATORE, integrato nello SPIDER, lavora sulla pagina già pulita dal PARSER.

1.6) CERCATORE, COSA E' E A COSA SERVE

Questa ultima componente ha il compito di interrogare gli indici, ordinare le ricerche per rilevanza, creare eventualmente una copia/cache per le ricerche già elaborate in modo da non ripetere inutili elaborazioni, che possono costare risorse di macchina. Questa componente ha poi il compito di presentare i risultati all'utente che li chiede.
Nello spider è integrato un CERCATORE che è in grado di lavorare sugli indici creati, nella versione PROFESSIONAL ed ENTERPRISE viene invece fornito uno script aggiuntivo in Php che che è in grado di fare le stesse operazioni fatte dal CERCATORE integrato sullo spider.

2) OPZIONI E CONFIGURAZIONE DI GIOVANNI CEGLIA SPIDER V 1.0

OPZIONI > Recupera Links (Modalità Spider)

Questa opzione provvedere a recuperare i links all'interno di una pagina html, e a metterli in lista per essere esaminati, a seconda delle strategie di spidering, dipendeti dalle altre impostazioni.

OPZIONI > Attiva parametri nelle Url

Questa opzione provvedere a recuperare i link con tutti i parametri, salvo tagliare i links con eccessivi parametri o troppo lunghi, se questa opzione è disattivata lo spider, preleva solo links puliti, senza parametri.

OPZIONI > Crea indici per pagine scandagliate

Questa opzione è compatibile solo con la modalità "Enterprise" e provvede a creare gli indici per il motore di ricerca.

OPZIONI > Attiva modalità Concorrente

Questa opzione è valida per tutte le strategie di spidering, e serve a dare la possibilità allo spider di essere eseguito con più istanze sulla stessa macchina o su una rete, utilizzando un database degli indirizzi da esaminare comune.

OPZIONI > Modalità Enterprise

Questa modalità, è la modalità per scansionare grosse quantità di links, si appoggia ad un database basato sulla tecnica hash, che va preventivamente creato tramite gli appositi menu e funzioni dello spider.

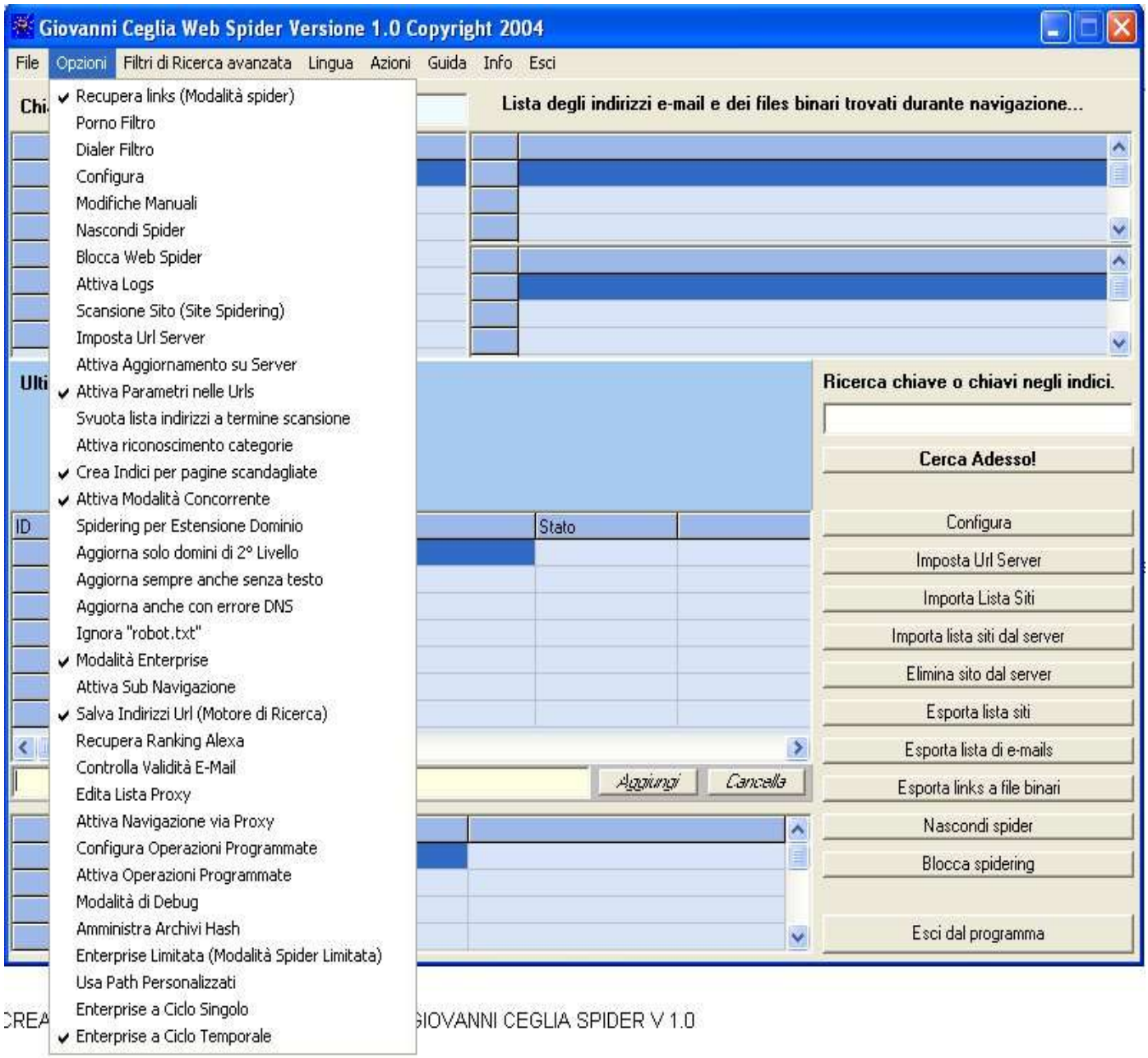
OPZIONI > Salva Indirizzi Url

Questa modalità, dice allo spider di salvare nel database per la modalità Enterprise, i nuovi indirizzi trovati, in modo che possono essere esaminati.

OPZIONI > Enterprise a Ciclo Temporale

Questa modalità, dice allo spider di effettuare una sola scansione degli indirizzi in un certo periodo di tempo, serve a non ripetere la richiesta di un indirizzo già esaminato da poco.

Per una immagine di come configurare il menù opzione:



Per informazioni su altre altre opzioni o modalità di spidering, contattare giovanniceglia@xungame.com.

3) CASI DI STUDIO E DI APPLICAZIONE

3.4) MOTORE DI RICERCA PER SITI E DOCUMENTI WEB

Questa breve guida illustra come impostare "Giovanni Ceglia Spider", per creare la base di un motori di ricerca per indici e parole chiavi. Come prima cosa avviare lo Spider, e a questo punto, impostare come vere, cioè devono avere la linguetta, le seguenti opzioni:

OPZIONI > Recupera Links (Modalità Spider)
OPZIONI > Attiva parametri nelle Url
OPZIONI > Crea indici per pagine scandagliate
OPZIONI > Attiva modalità Concorrente
OPZIONI > Modalità Enterprise
OPZIONI > Salva Indirizzi Url
OPZIONI > Enterprise a Ciclo Temporale

A questo punto bisogna creare un archivio per la modalità Enterprise, e bisogna settarlo di default. Per fare questo, cliccare su:

OPZIONI > Amministra archivi Hash

Si aprirà un form dove è possibile creare nuovi file .hash, che possono essere utilizzati per la modalità Enterprise. Quindi inseirre un nome per il file, inserire una descrizione, e quindi dove dice "records", inserire un numero alto (più siti si vogliono includere e più è consigliabile inserire un numero alto, altrimenti le operazioni di spidering e indexing diventeranno lente).

A questo punto selezionare l'archio hash di default.

Quindi salvare le impostazioni che abbiamo configurato, tramite:

FILE > Salva Impostazioni.

Quindi chiudere lo Spider, e riavviarlo.

A questo punto le impostazioni sono effettive, e possiamo importare una lista di siti da cui inziare, per creare un piccolo indice da far utilizzare al motore di ricerca. Per importare una lista di siti, potete, creare un file di testo (Plain Text) inserendo una lsita di siti, e poi lo potete importare tramite:

FILE > Importa per Modalità Enteprise

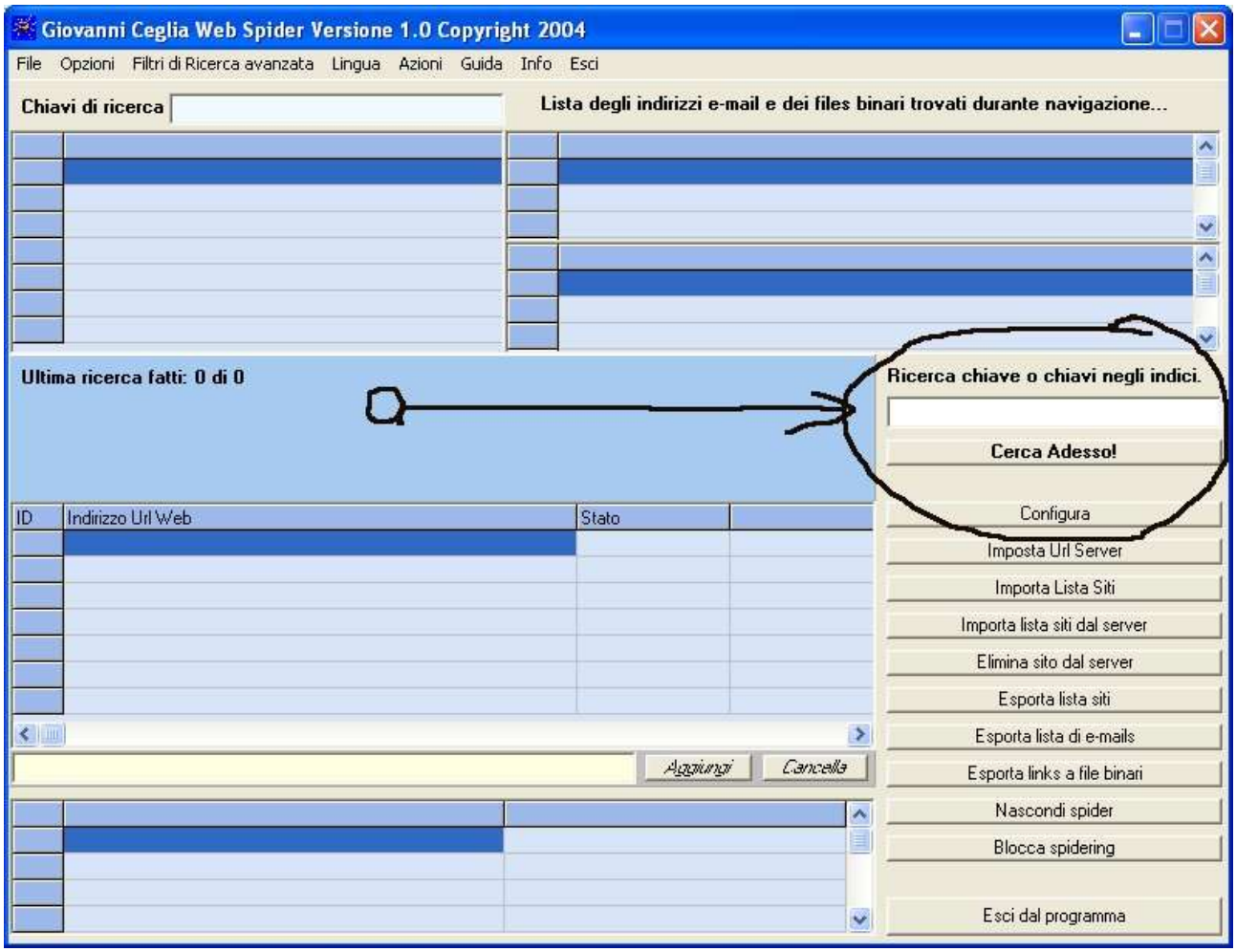
Gli indirizzi verranno importati nel file .hash che avete selezionato, come file di default.

A questo punto, potete accendere lo Spider, che comincerà il suo lavoro. Cliccate su:

AZIONI > Avvia Scansione.

Lo spider comincerà a scorrere l'archivio hash, e quando troverà degli indirizzi, provvederà a recuperare la pagina, fare il parsing, recuperare i links, creare gli indici per le parole trovate nella pagina.

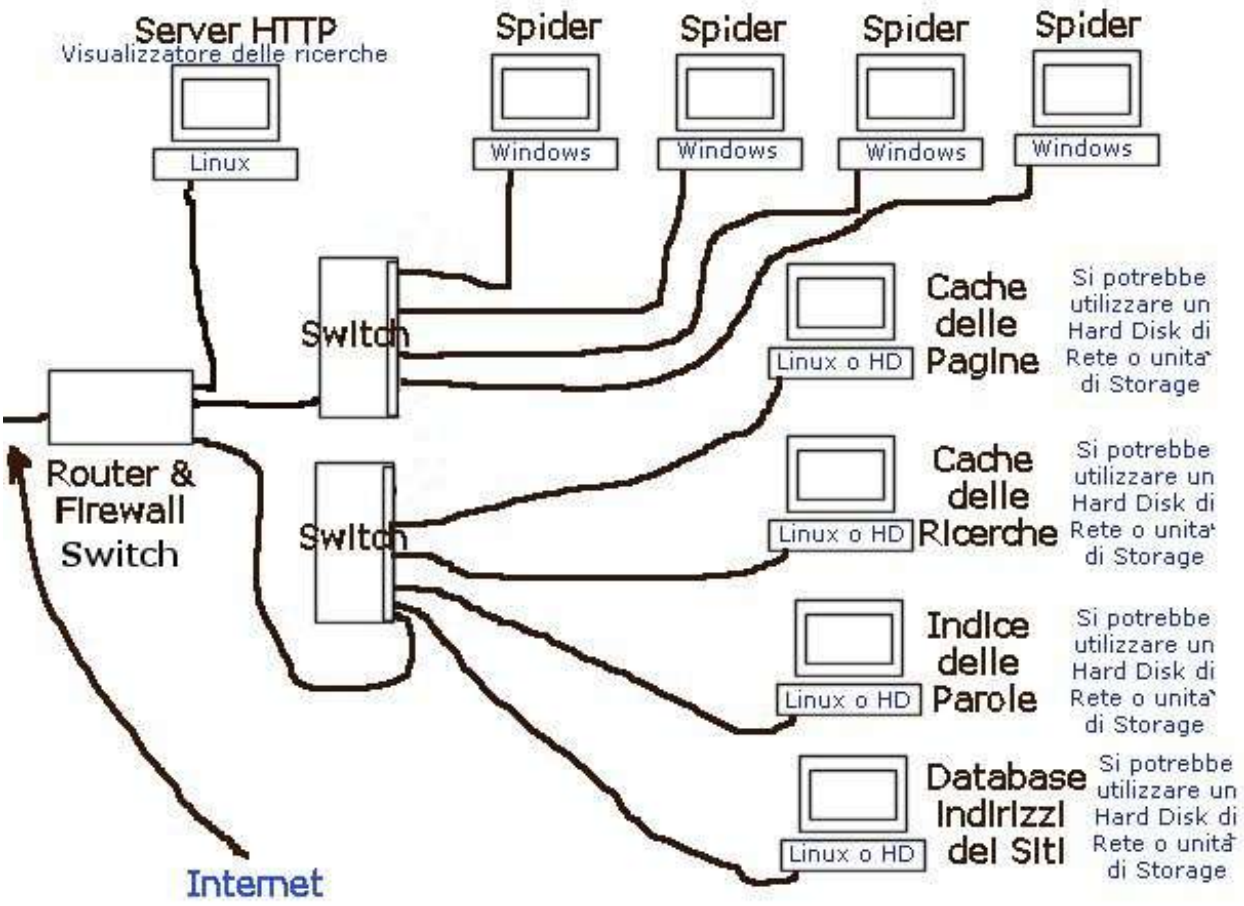
Una volta scansionata qualche pagina web, è possibile, provare la ricerca per singola parola chiave (per il momento), utilizzando il form a destra dello spider, esattamente come indicato in figura. I risultati verranno ordinati per rilevanza. La rilevanza è calcolata in base ad alcuni parametri, pochi per il momento, che includono la posizione nella pagina html, della parola (Titolo, Body, Meta Tags), la quantità di quelle parole, e se la parola è presente nell'indirizzo URL.



E' possibile ovviamente modificare le opzioni selezionate ed ottenere diverse modalità di spidering, a seconda delle opzioni selezionate, modalità che verranno spiegate, in questo documento, nei prossimi rilasci, e revisioni.

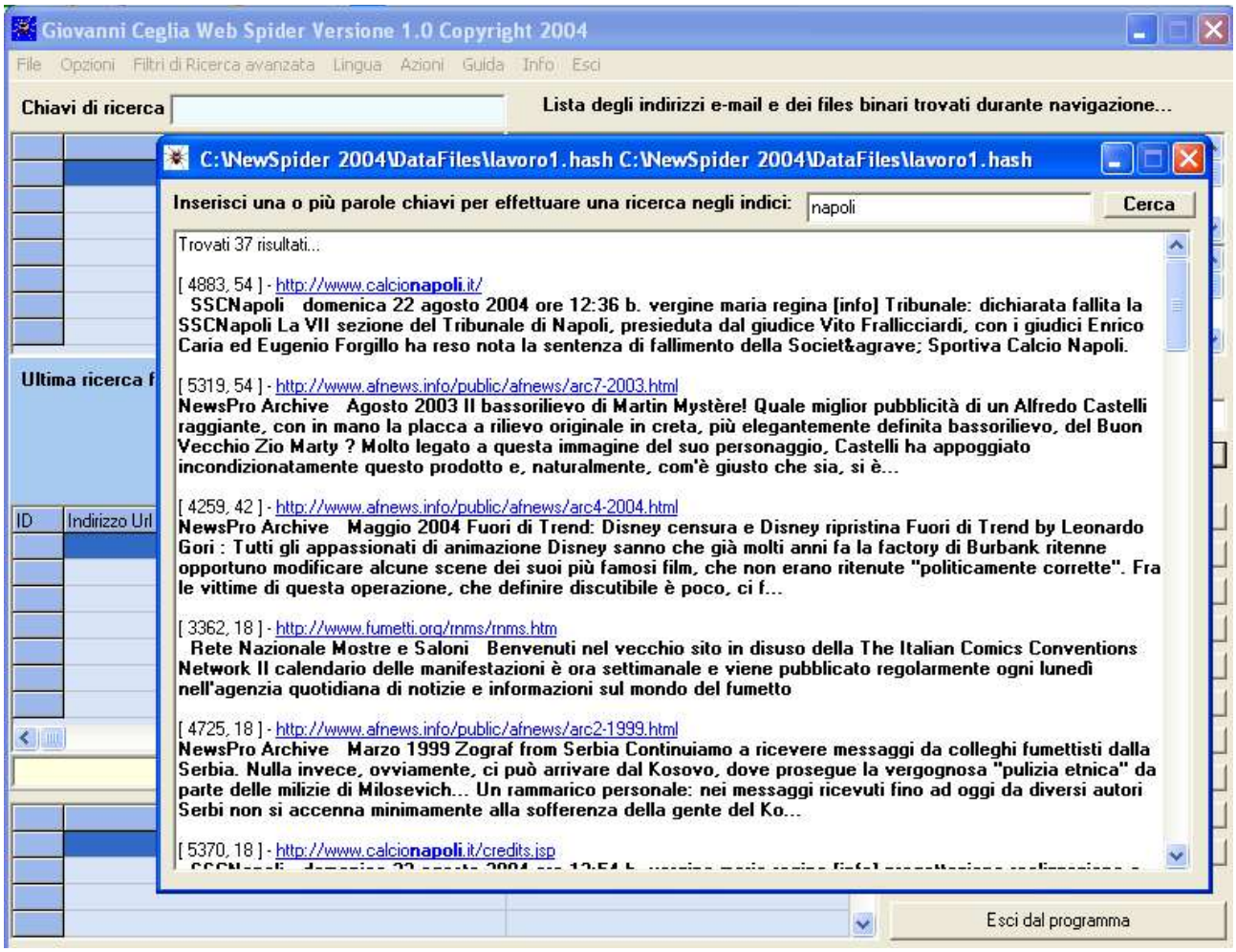
4) GUIDA ALLA CREAZIONE DI UN MOTORE DI RICERCA CON GIOVANNI CEGLIA SPIDER V 1.0

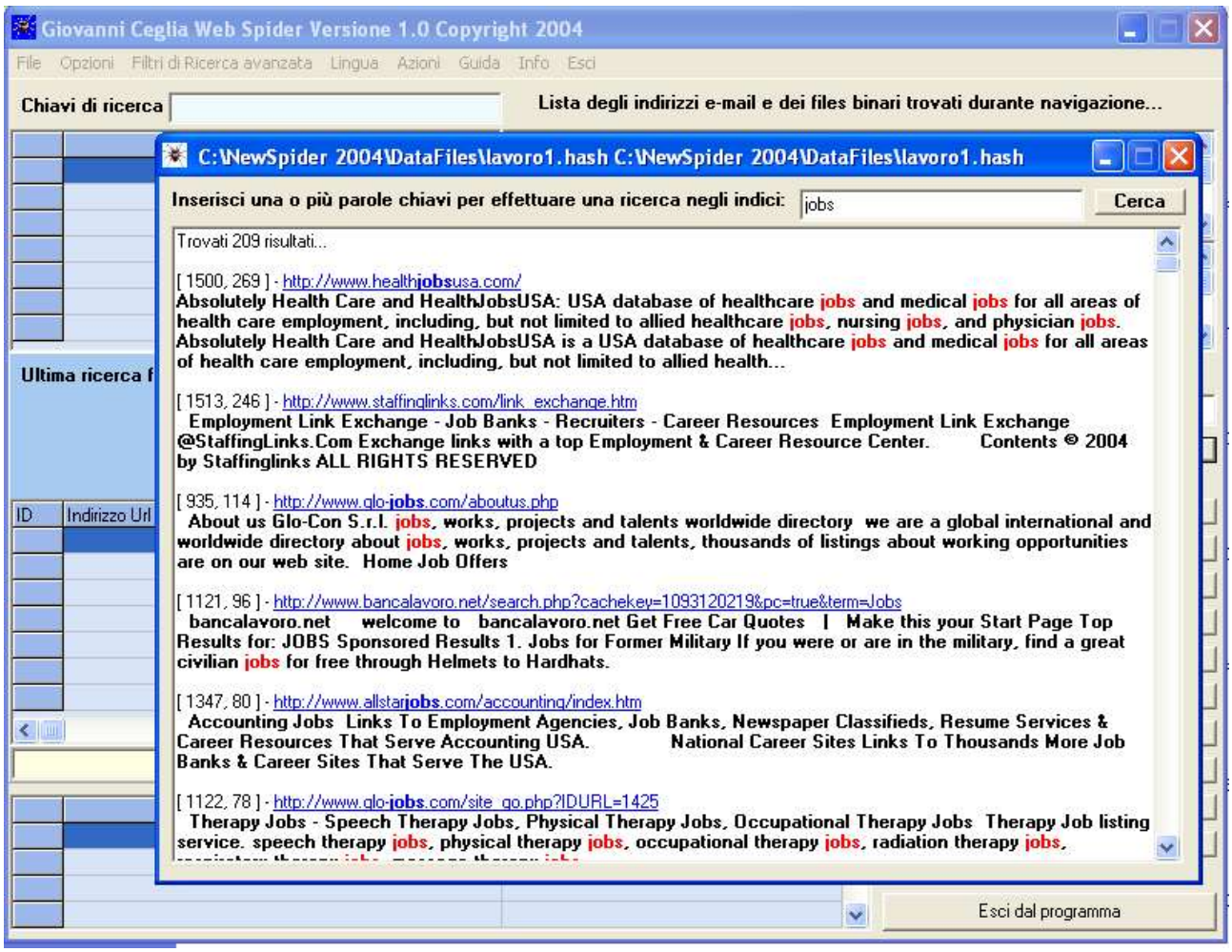
Potete visualizzare in questa immagine un modello di come potrebbe essere strutturato un piccolo motore di ricerca, utilizzando “Giovanni Ceglia Spider”.

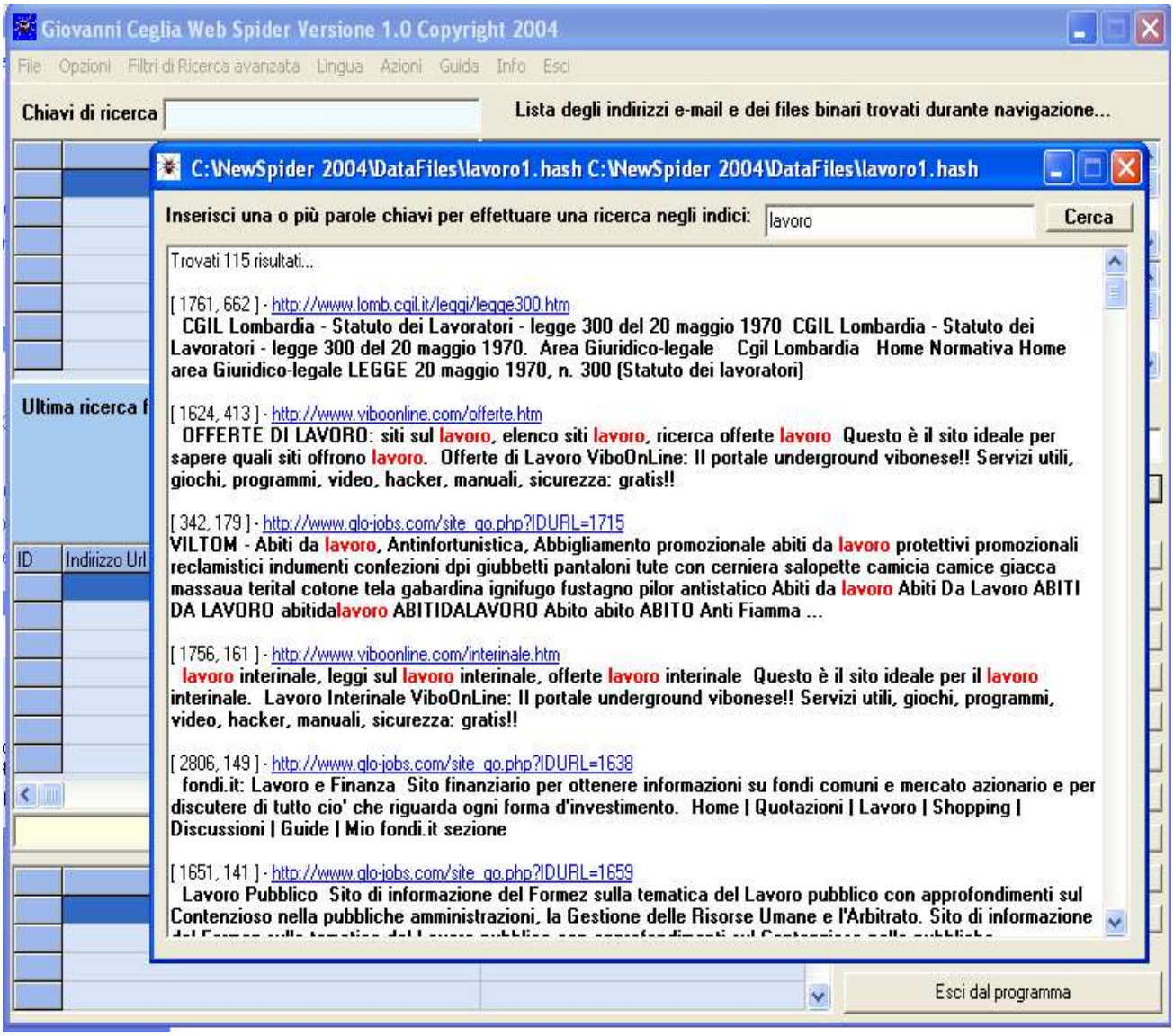


Per maggiori informazioni e versioni più aggiornate, contattare giovanniceglia@xungame.com

ALCUNI ESEMPI DI RICERCHE SU UNA PICCOLA BASE DI INDIRIZZI (25000)




Giovanni Ceglia Web Spider Versione 1.0 Copyright 2004
File Opzioni Filtri di Ricerca avanzata Lingua Azioni Guida Info Esci
Chiavi di ricerca
Lista degli indirizzi e-mail e dei files binari trovati durante navigazione...
C:\NewSpider 2004\DataFiles\lavoro1.hash C:\NewSpider 2004\DataFiles\lavoro1.hash
Inserisci una o più parole chiavi per effettuare una ricerca negli indici:
jobs
Cerca
Trovati 209 risultati...
[1500, 269] - http://www.healthjobsusa.com/
Absolutely Health Care and HealthJobsUSA: USA database of healthcare jobs and medical jobs for all areas of health care employment, including, but not limited to allied healthcare jobs, nursing jobs, and physician jobs. Absolutely Health Care and HealthJobsUSA is a USA database of healthcare jobs and medical jobs for all areas of health care employment, including, but not limited to allied health...
[1513, 246] - http://www.staffinglinks.com/link_exchange.htm
Employment Link Exchange - Job Banks - Recruiters - Career Resources Employment Link Exchange @StaffingLinks.Com Exchange links with a top Employment & Career Resource Center. Contents © 2004 by Staffinglinks ALL RIGHTS RESERVED
[935, 114] - http://www.glo-jobs.com/aboutus.php
About us Glo-Con S.r.l. jobs, works, projects and talents worldwide directory we are a global international and worldwide directory about jobs, works, projects and talents, thousands of listings about working opportunities are on our web site. Home Job Offers
[1121, 96] - http://www.bancalavoro.net/search.php?cachekey=1093120219&pc=true&term=Jobs
bancalavoro.net welcome to bancalavoro.net Get Free Car Quotes | Make this your Start Page Top Results for: JOBS Sponsored Results 1. Jobs for Former Military If you were or are in the military, find a great civilian jobs for free through Helmets to Hardhats.
[1347, 80] - http://www.allstarjobs.com/accounting/index.htm
Accounting Jobs Links To Employment Agencies, Job Banks, Newspaper Classifieds, Resume Services & Career Resources That Serve Accounting USA. National Career Sites Links To Thousands More Job Banks & Career Sites That Serve The USA.
[1122, 78] - http://www.glo-jobs.com/site_go.php?IDURL=1425
Therapy Jobs - Speech Therapy Jobs, Physical Therapy Jobs, Occupational Therapy Jobs Therapy Job listing service. speech therapy jobs, physical therapy jobs, occupational therapy jobs, radiation therapy jobs,
Ultima ricerca f
ID
Indirizzo Url
Esci dal programma


Giovanni Ceglia Web Spider Versione 1.0 Copyright 2004
File Opzioni Filtri di Ricerca avanzata Lingua Azioni Guida Info Esci
Chiavi di ricerca
Lista degli indirizzi e-mail e dei files binari trovati durante navigazione...
C:\NewSpider 2004\DataFiles\lavoro1.hash C:\NewSpider 2004\DataFiles\lavoro1.hash
Inserisci una o più parole chiavi per effettuare una ricerca negli indici:
lavoro
Cerca
Trovati 115 risultati...
[1761, 662] - http://www.lomb.cgil.it/leggi/legge300.htm
CGIL Lombardia - Statuto dei Lavoratori - legge 300 del 20 maggio 1970 CGIL Lombardia - Statuto dei Lavoratori - legge 300 del 20 maggio 1970. Area Giuridico-legale Cgil Lombardia Home Normativa Home area Giuridico-legale LEGGE 20 maggio 1970, n. 300 (Statuto dei lavoratori)
[1624, 413] - http://www.viboonline.com/offerte.htm
OFFERTE DI LAVORO: siti sul lavoro, elenco siti lavoro, ricerca offerte lavoro Questo è il sito ideale per sapere quali siti offrono lavoro. Offerte di Lavoro ViboOnLine: Il portale underground vibonese!! Servizi utili, giochi, programmi, video, hacker, manuali, sicurezza: gratis!!
[342, 179] - http://www.qlo-jobs.com/site_go.php?IDURL=1715
VILTOM - Abiti da lavoro, Antinfortunistica, Abbigliamento promozionale abiti da lavoro protettivi promozionali reclamistici indumenti confezioni dpi giubbetti pantaloni tute con cerniera salopette camicia camice giacca massaua terital cotone tela gabardina ignifugo fustagno pilor antistatico Abiti da lavoro Abiti Da Lavoro ABITI DA LAVORO abitidalavoro ABITIDALAVORO Abito abito ABITO Anti Fiamma ...
[1756, 161] - http://www.viboonline.com/interinale.htm
lavoro interinale, leggi sul lavoro interinale, offerte lavoro interinale Questo è il sito ideale per il lavoro interinale. Lavoro Interinale ViboOnLine: Il portale underground vibonese!! Servizi utili, giochi, programmi, video, hacker, manuali, sicurezza: gratis!!
[2806, 149] - http://www.qlo-jobs.com/site_go.php?IDURL=1638
fondi.it: Lavoro e Finanza Sito finanziario per ottenere informazioni su fondi comuni e mercato azionario e per discutere di tutto cio' che riguarda ogni forma d'investimento. Home | Quotazioni | Lavoro | Shopping | Discussioni | Guide | Mio fondi.it sezione
[1651, 141] - http://www.qlo-jobs.com/site_go.php?IDURL=1659
Lavoro Pubblico Sito di informazione del Formez sulla tematica del Lavoro pubblico con approfondimenti sul Contenzioso nella pubbliche amministrazioni, la Gestione delle Risorse Umane e l'Arbitrato. Sito di informazione
Ultima ricerca f
ID
Indirizzo Url
Esci dal programma