Guida ad Analog

Versione 1.0 – 4/7/2005
© *Antonio Frecentese*

Analog è un programma che analizza i file di log dei server web e produce dei report con le statistiche di accesso. In questo modo è possibile studiare, entro certi limiti, il comportamento degli utenti del nostro sito e migliorare la navigabilità e/o i contenuti del sito stesso. Il programma è disponibile per diverse piattaforme e soprattutto è gratuito (licenza GNU GPL).

Questa guida vuole fornire le basi per poter configurare il programma affinché produca le statistiche a noi necessarie. Io lo uso sotto Windows, ma non ci sono differenze usando altri sistemi operativi: basta eseguire Analog ed esso leggerà il file di log e produrrà il report seguendo le indicazioni fornite nel file di configurazione, `analog.cfg`. La versione usata per questa guida è la 6. Inoltre, come esempio, esamineremo un file di configurazione che avevo utilizzato per il mio sito; sentitevi liberi di adattarlo a seconda di cosa vi serve.

Secondo me il metodo migliore per imparare a usare il programma è questo:

- leggete questa guida da cima a fondo per farvi un'idea di cosa potete fare con Analog;
- studiate la documentazione allegata al programma e in particolare i file di configurazione di esempio. A mio avviso il più utile di questi è `bigbyrep.cfg`, non solo perché ha molte opzioni ma soprattutto perché esse sono divise per report, quindi si fa meno fatica a individuare le opzioni da modificare;
- create infine il vostro file di configurazione secondo le vostre esigenze.

Un'ultima nota prima di cominciare: il file di log usato per questa guida è in formato ECLF (Extended Common Log Format), ma Analog è in grado di gestire anche altri formati.[1] Giusto per fare un esempio di come si presenti una line di un file di log in formato ECLF, la seguente è una linea di un logfile del mio sito (tenete presente che in realtà è *tutto* scritto su *una sola linea*):

```
host222-45.pool8249.interbusiness.it - - [17/Jun/2005:18:35:55 +0200] "GET /
Nihongo/kanjitest.php HTTP/1.1" 200 28076
"http://frech.altervista.org/Nihongo.php" "Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; SV1; .NET CLR 1.1.4322)"
```

E ora buon lavoro!

Indice

[1] Se per caso il programma non riuscisse a interpretare correttamente il file di log, è possibile che il log sia stato creato in modo non standard o che sia di un tipo non noto ad Analog. Si può allora creare un formato personalizzato per indicare al programma come interpretare i vari campi. Tuttavia questo è un argomento piuttosto complesso, quindi non lo tratterò in questa guida.

Cosa fa Analog

Analog non crea i file di log, li legge soltanto. Bisogna allora procurarsi tali file. Se il vostro ISP lo ha installato e vi sono stati dati i permessi necessari, è possibile eseguire Analog sul server del vostro ISP; altrimenti è necessario scaricare il file di log sul vostro computer e lanciare Analog da lì.[2]

Sotto Windows basta cliccare due volte sull'icona di Analog: il programma produrrà i file delle statistiche dove e come specificato nel file di configurazione, `analog.cfg`. Quest'ultimo va modificato con un comune text editor (anche il "Blocco Note" va bene)[3] e deve trovarsi nel computer locale (Analog non usa FTP né HTTP per prelevarlo da Internet).

Il primo comando da scrivervi, il più importante, è allora `LOGFILE` seguito dal nome del file di log completo di percorso:[4]

```
LOGFILE logfilename     # percorso del file di log da analizzare
```

Notate che il carattere `#` dà inizio a un commento (che dura fino alla fine della riga), ignorato dal programma. Usate spesso i commenti, in modo da capire meglio anche in seguito perché avete scritto così e non cosà.

Analog può effettuare fino a 44 differenti report, a seconda della configurazione scelta e dei dati memorizzati nel file di log.[5] Ecco l'elenco completo:

SIGLA	NOME	TITOLO	COSA MOSTRA
x	GENERAL	General Summary	sommario generale
1	YEARLY	Yearly Report	report annuale
Q	QUARTERLY	Quarterly Report	report trimestrale
m	MONTHLY	Monthly Report	report mensile
W	WEEKLY	Weekly Report	report settimanale
D	DAILYREP	Daily Report	report giornaliero
d	DAILYSUM	Daily Summary	report per giorno della settimana (tutti i lunedì, tutti i martedì eccetera)
H	HOURLYREP	Hourly Report	report per ora
h	HOURLYSUM	Hourly Summary	riassunto per fascia oraria
w	WEEKHOUR	Hour of the Week Summary	riassunto per fascia oraria durante la settimana
4	QUARTERREP	Quarter-Hour Report	report per quarti d'ora
6	QUARTERSUM	Quarter-Hour Summary	riassunto per fasce di 15 minuti
5	FIVEREP	Five-Minute Report	report per 5 minuti
7	FIVESUM	Five-Minute Summary	riassunto per fasce di 5 minuti
S	HOST	Host Report	tutti i computer che hanno richiesto dei file dal sito
l	REDIRHOST	Host Redirection Report	computer che hanno incontrato rispettivamente redirezioni o errori
L	FAILHOST	Host Failure Report	
Z	ORGANISATION	Organisation Report	cerca di elencare le organizzazioni (ISP, compagnie, eccetera) sotto cui sono registrati i vari computer
o	DOMAIN	Domain Report	i Paesi in cui si trovavano tali computer
r	REQUEST	Request Report	quali files sono stati scaricati
i	DIRECTORY	Directory Report	in quali directory si trovavano tali files

[2] Chi, come me, ricorre a servizi di hosting gratuiti è quasi certamente obbligato a questa seconda via. Io ho impostato nel mio sito uno script che mi invia periodicamente il file di logo come allegato di un'e-mail e poi lo rimuove dal server, così posso usare Analog sul mio computer quando voglio.

[3] Se lo usate, però, ricordate di salvare il file col nome `analog.cfg` racchiuso tra virgolette e selezionando nel campo "Salva come..." l'opzione "Tutti i file".

[4] E' possibile indicare più di un file di log separandoli l'uno dall'altro con delle virgole, e anche usare delle wildcards (ad esempio: `LOGFILE new1.log,old*.log`).

[5] Infatti, per esempio, un file in formato CLF ha meno campi (e dunque riporta meno informazioni) rispetto a uno in formato ECLF, quindi alcune analisi non sono possibili.

t	FILETYPE	File Type Report	il tipo di quei files (in effetti, la loro estensione)
z	SIZE	File Size Report	le dimensioni dei files
P	PROCTIME	Processing Time Report	il tempo necessario per servire ogni file
E	REDIR	Redirection Report	i nomi che hanno provocato delle redirezioni; principalmente si tratta di directory indicate senza la barra finale
I	FAILURE	Failure Report	i file che hanno causato degli errori
f	REFERRER	Referrer Report	i referrer, cioè le pagine che avevano un collegamento al file in questione, da cui insomma proviene la richiesta
s	REFSITE	Referring Site Report	i server su cui si trovavano tali referrer
N	SEARCHQUERY	Search Query Report	elenco delle "query string", cioè dei differenti gruppi di parole usate dai visitatori per trovare il vostro sito
n	SEARCHWORD	Search Word Report	elenco di tutti i differenti termini usati dai visitatori per trovare il vostro sito
Y	INTSEARCHQUERY	Internal Search Query Report	elenco delle "query string", cioè dei differenti gruppi di parole usate dai visitatori negli script interni al vostro sito
y	INTSEARCHWORD	Internal Search Word Report	elenco di tutti i differenti termini usati dai visitatori negli script interni al vostro sito
k	REDIRREF	Redirected Referrer Report	i referrer che conducono a redirezioni
K	FAILREF	Failed Referrer Report	essenzialmente un elenco di collegamenti inesistenti (broken link)
B	BROWSERREP	Browser Report	versione del browser usata (o meglio, quella dichiarata)
b	BROWSERSUM	Browser Summary	ditta fornitrice del browser (per come esso è stato dichiarato)
p	OSREP	Operating System Report	cerca di elencare il sistema operativo su cui girava il browser del visitatore
v	VHOST	Virtual Host Report	l'attività dei domini virtuali
R	REDIRVHOST	Virtual Host Redirection Report	rispettivamente le redirezioni e gli errori incontrati in ognuno di quei domini virtuali
M	FAILVHOST	Virtual Host Failure Report	
u	USER	User Report	visitatori, se il server del sito richiede l'autorizzazione; forse il cookie o l'identificativo di sessione dell'utente
j	REDIRUSER	User Redirection Report	rispettivamente gli utenti che hanno incontrato redirezioni o errori
J	FAILUSER	User Failure Report	
c	STATUS	Status Code Report	numero dei codici di status HTTP che si sono avuti (es. 200 = "tutto ok", 404 = "pagina non trovata")

Notate come ogni report sia identificato anche da un numero o da una lettera (attenzione: maiuscolo e minuscolo indicano report diversi!). Queste lettere saranno usate per ordinare i report nell'ordine a noi più congeniale.

Configurazione: lettura dei log

Una volta indicato ad Analog il file da analizzare, dobbiamo istruirlo su come esso debba essere letto, cosa debba prendere in considerazione e cosa no.

Ogni volta che un browser richiede una pagina web, nel file di log viene registrata tutta una serie di richieste (ovvero la pagina in questione più tutti gli elementi in essa presenti, dalle immagini alle animazioni Flash), ma a qualcuno potrebbero non interessare le statistiche su certi tipi di dati. Oppure si potrebbe non voler considerare i referrer interni al proprio sito o le richieste provenienti da un certo indirizzo (ad esempio perché si ha un IP fisso e le connessioni fatte da quell'indirizzo IP sono fatte solo a scopo di test). Ecco allora che coi comandi `INCLUDE` ed `EXCLUDE` possiamo indicare ad Analog di ignorare o, al contrario, di considerare certi dati.

Ad esempio, la linea

```
REFEXCLUDE http://frech.altervista.org/*
```

indica di ignorare le richieste provenienti dall'interno del sito (o meglio, il cui referrer è una pagina interna al sito), mentre

```
HOSTEXCLUDE miocomputer.mioisp.it
```

esclude tutte le richieste provenienti da quel dato computer.

La cosa importante da ricordare è che i comandi `INCLUDE` ed `EXCLUDE` vengono considerati in ordine sequenziale e che ogni cosa indicata viene inclusa o esclusa in base all'ultima regola che la riguarda: è possibile allora includere, per esempio, tutti i file di una certa directory e in seguito escludere tutti quelli di una sua sottodirectory. Per quanto riguarda gli host, si possono indicare sia degli indirizzi IP numerici che i nomi di dominio.

Ecco la lista completa di questi comandi:

- `HOSTINCLUDE` e `HOSTEXCLUDE`;
- `FILEINCLUDE` e `FILEEXCLUDE`;
- `BROWINCLUDE` e `BROWEXCLUDE`;
- `REFINCLUDE` e `REFEXCLUDE`;
- `USERINCLUDE` e `USEREXCLUDE`;
- `VHOSTINCLUDE` e `VHOSTEXCLUDE`;
- `STATUSINCLUDE` e `STATUSEXCLUDE`.

Questi comandi sono generali, ma ve ne sono anche di specifici per i vari report, cioè che escludono le linee del file di log non da tutti i report ma solo da alcuni di essi. Quindi per esempio

```
REFREPEXCLUDE http://frech.altervista.org/*
```

esclude le righe con referrer interni solo dal "Referrer Report" ma non dagli altri report.

La lista completa di questi comandi specifici è:

- `REQINCLUDE` e `REQEXCLUDE`;
- `REDIRINCLUDE` e `REDIREXCLUDE`;
- `FAILINCLUDE` e `FAILEXCLUDE`;
- `TYPEINCLUDE` e `TYPEEXCLUDE`;
- `DIRINCLUDE` e `DIREXCLUDE`;
- `HOSTREPINCLUDE` e `HOSTREPEXCLUDE`;
- `REDIRHOSTINCLUDE` e `REDIRHOSTEXCLUDE`;
- `FAILHOSTINCLUDE` e `FAILHOSTEXCLUDE`;
- `DOMINCLUDE` e `DOMEXCLUDE`;
- `ORGINCLUDE` e `ORGEXCLUDE`;
- `REFREPINCLUDE` e `REFREPEXCLUDE`;
- `REFSITEINCLUDE` e `REFSITEEXCLUDE`;
- `SEARCHQUERYINCLUDE` e `SEARCHQUERYEXCLUDE`;
- `SEARCHWORDINCLUDE` and `SEARCHWORDEXCLUDE`;
- `INTSEARCHQUERYINCLUDE` e `INTSEARCHQUERYEXCLUDE`;
- `INTSEARCHWORDINCLUDE` e `INTSEARCHWORDEXCLUDE`;

- REDIRREFINCLUDE e REDIRREFEXCLUDE;
- FAILREFINCLUDE e FAILREFEXCLUDE;
- BROWSUMINCLUDE e BROWSUMEXCLUDE;
- BROWREPINCLUDE e BROWREPEXCLUDE;
- OSINCLUDE e OSEXCLUDE; VHOSTREPINCLUDE e VHOSTREPEXCLUDE;
- REDIRVHOSTREPINCLUDE e REDIRVHOSTREPEXCLUDE;
- FAILVHOSTREPINCLUDE e FAILVHOSTREPEXCLUDE;
- USERREPINCLUDE e USERREPEXCLUDE;
- REDIRUSERREPINCLUDE e REDIRUSERREPEXCLUDE;
- FAILUSERINCLUDE e FAILUSEREXCLUDE.

C'è ancora una cosa importante da dire, almeno per quanto riguarda un sito come il mio. Analog considera come pagine i file scritti in HTML, ma il mio sito è composto da pagine in PHP. Bisogna allora informarlo che anche i file .php sono da considerarsi pagine:

```
PAGEINCLUDE *.php
```

Ovviamente, per chi ha pagine ASP il discorso è lo stesso, basta cambiare l'estensione.

Argomenti delle ricerche

Spesso si rivela utile conoscere in base a quali parole-chiave i visitatori sono riusciti a individuare il nostro sito tra milioni di altri, o quali parole-chiave vengono più usate dai visitatori nel motore di ricerca interno del nostro sito per cercare l'articolo o il prodotto di loro interesse. Sarà sufficiente indicare ad Analog quali sono i campi che corrispondono all'argomento della ricerca. Per esempio nel referrer

```
http://www.altavista.com/cgi-bin/query?pg=q&kl=XX&q=carrot+cake
```

si possono notare:

- lo script che esegue la ricerca, http://www.altavista.com/cgi-bin/query;
- il punto interrogativo, che separa lo script dai suoi parametri;
- i vari parametri, ovvero pg, kl e q, seguiti ognuno dal segno di uguale (=) e dai valori da essi assunti; i parametri sono separati l'uno dall'altro dal carattere ampersand (&).

In particolare, nell'esempio appena visto il termine della ricerca è contenuto nel campo q ed è composto da due parole unite dal segno +. Si usa allora il comando SEARCHENGINE nel modo seguente:

```
SEARCHENGINE http://www.altavista.com/cgi-bin/query q
```

Se si vogliono considerare tutti i mirror nei vari Paesi, si può usare la forma

```
SEARCHENGINE http://*altavista.*/* q
```

Per quanto riguarda il motore di ricerca interno (ma anche più semplicemente uno script del sito che prevede degli argomenti), il funzionamento è lo stesso ma viene usato il comando INTSEARCHENGINE. Ad esempio nel caso di un referrer come

```
/cgi-bin/search?trm=chocolate+cake
```

si usa

```
INTSEARCHENGINE /cgi-bin/search trm
```

Configurazione: stesura dei report

Adesso è giunto il momento di indicare ad Analog cosa vogliamo sapere e come vogliamo che venga creato il report.

E' bene che nel titolo del file di report compaia il nome e l'indirizzo del nostro sito. Ecco un esempio di come vanno segnati:

```
HOSTNAME "La Soffitta"
HOSTURL http://frech.altervista.org/
```

Nel nome del sito ci sono degli spazi, ecco perché esso viene racchiuso da virgolette. Posso anche usare

```
HOSTURL none
```

per non creare un collegamento. Riguardo al nome dell'host: Analog traduce i caratteri nella codifica usata per l'HTML, quindi se si vogliono inserire dei caratteri particolari come delle lettere accentate bisogna farli precedere da una barra:

```
HOSTNAME "M\&uuml;ller"
```

Scegliamo poi che tipo di file di output vogliamo: è possibile produrne di 7 tipi, ovvero `XHTML` (quello di default), `HTML`, `PLAIN`, `ASCII`, `XML`, `LATEX` e `COMPUTER`. Per specificare quale ci occorre basta usare il comando `OUTPUT` seguito dal tipo di report (es. `OUTPUT HTML`). Io non uso questo comando perché perché mi va bene il formato di default.

Bisogna poi usare il comando

```
OUTFILE nome_file # esempio: OUTFILE Report.html
```

per specificare il nome del file da produrre.[6] Da notare anche che si possono inserire delle particolari sequenze di caratteri in modo da creare dei report nel cui nome è compresa la data di creazione. Per esempio,

```
OUTFILE stats%y%M%D.html
```

produrrà un file col nome del tipo `stats990501.html`. Noterete che `%y` crea l'anno (le ultime due cifre), `%M` il mese (due cifre) e `%D` il giorno (due cifre).

Si può anche cambiare la lingua in cui dovrà essere scritto il report (di default è l'inglese):

```
LANGUAGE ITALIAN
```

imposterà la lingua italiana.

Soprattutto per chi lavora per una ditta e deve mostrare i report al capoufficio o al responsabile, può essere una buona idea inserire il logo dell'azienda al posto di quello di default di Analog. Per farlo, bisogna innanzitutto dire dove Analog dovrà trovare/creare le immagini del report. Attenzione: è necessario indicare un URL con la barra alla fine, e NON una directory fisica all'interno del computer:

```
IMAGEDIR img/   # URL relativo, nella stessa directory dell'output
IMAGEDIR http://www.myother.server.com/img/  # su un altro server
```

Dopodiché si può specificare l'immagine da usare come logo:

```
LOGO picture.gif
LOGO /images/picture2.gif
LOGO none         # per non usare nessun logo
```

Si presume che l'immagine sia all'interno della directory specificata con `IMAGEDIR`, a meno che non cominci con una barra o contenga `://`.

Ora si può anche specificare l'indirizzo Internet a cui collegare l'immagine:

```
LOGOURL http://www.la_mia_ditta.com/
LOGOURL none   # per non attivare collegamenti
```

Attenzione: `LOGOURL` funziona solo specificando il formato di output `XHTML`, non `HTML`.

Quali grafici inserire?

Consiglio a tutti coloro che usano Analog per la prima volta di settare `ALL ON` e poi lanciare il programma. In questo modo avrete un file di report molto lungo ma potrete farvi un'idea di quello che potete ottenere. Osservate bene tutti i report che vengono così prodotti e decidete quali vi servono; prendete anche nota di come vengono presentati i dati: per esempio, facendo come vi ho

[6] E' possibile indicare il percorso completo in cui salvarlo. Se non viene fatto, il file verrà salvato nella cartella in cui il programma si aspetta di salvarli (sotto Windows è la stessa cartella in cui si trova il programma).

suggerito vengono indicate solo le pagine del sito che hanno ricevuto più di un certo numero di richieste (20, se non ricordo male), ma a voi potrebbero servire tutte, oppure una selezione ancora più ristretta o ancora basata su un altro criterio. Impostate poi `ALL OFF` e abilitate uno per uno i report che vi occorrono.

A me piace avere una breve descrizione del contenuto dei report, come promemoria:

```
DESCRIPTIONS ON
```

Non voglio invece vedere dopo ogni titolo la sfilza di link agli altri report presenti nel file:

```
GOTOS OFF
```

Non voglio nemmeno sapere quanto ci impiega il programma ad analizzare i log (non ci mette mai molto):

```
RUNTIME OFF
```

Si può anche cambiare l'ordine in cui i report dovranno essere visualizzati, usando il comando `REPORTORDER` ed elencando le lettere di tutti i report possibili nell'ordine in cui li vogliamo avere. I caratteri non alfanumerici vengono ignorati, quindi possono essere usati come separatori. Ad esempio si può scrivere:

```
REPORTORDER x-1QmdDhHw4567W-cPz-ritEIYy-SlLZo-sNnfKk-ujJ-vMR-bBp
```

Report relativi al tempo

Io non voglio il "Daily Summary" (riassunto per giorno della settimana) ma il "Daily Report" (per ogni singolo giorno mese registrato nel file di log) mi interessa:

```
DAILYSUM OFF
DAILYREP ON
```

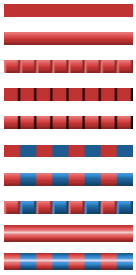
Per quest'ultimo, specifico quali colonne visualizzare nella tabella finale:

```
DAYREPCOLS RP
DAYREPGRAPH P
```

La `R` indica il numero di richieste, la `P` il numero delle pagine. Posso anche indicare la `B` (numero di bytes trasferiti); le lettere minuscole indicano invece la percentuale: `r` è la percentuale di richieste, e così via.

Un'altra cosa riguardo questo tipo di report: essi usano delle barre colorate per indicare il numero delle richieste, e con il comando `BARSTYLE` è possibile indicare quale stile si vuole usare (nel caso usiate un browser testuale, questo comando non avrà effetto). Ecco le opzioni disponibili:

```
BARSTYLE a
BARSTYLE b
BARSTYLE c
BARSTYLE d
BARSTYLE e
BARSTYLE f
BARSTYLE g
BARSTYLE h
BARSTYLE i
BARSTYLE j
```



Lo stile di default è il `b`, ma io preferisco il `j`:

```
BARSTYLE j
```

Altri comandi per i report

Anche gli altri report hanno dei comandi `COLS`, ma sono previste delle colonne aggiuntive. Ecco tutte quelle disponibili:

`R` Numero di richieste
`r` Percentuale delle richieste
`S` Numero di richieste negli ultimi 7 giorni
`s` Percentuale delle richieste negli ultimi 7 giorni
`P` Numero delle pagine richieste
`p` Percentuale delle pagine richieste
`Q` Numero delle pagine richieste negli ultimi 7 giorni
`q` Percentuale delle pagine richieste negli ultimi 7 giorni
`B` Numero di bytes trasferiti
`b` Percentuale di bytes trasferiti

`C` Numero di bytes trasferiti negli ultimi 7 giorni
`c` Percentuale di bytes trasferiti negli ultimi 7 giorni
`d` Data dell'ultimo accesso
`D` Data e ora dell'ultimo accesso
`e` Data del primo accesso
`E` Data e ora del primo accesso
`N` Numero di elementi nella lista

Inoltre si può eventualmente usare il comando `SORTBY` per specificare come vanno ordinati i report:

`REQUESTS`	numero totale di richieste
`REQUESTS7`	richieste negli ultimi 7 giorni
`PAGES`	richieste totali per pagina
`PAGES7`	richieste per pagina negli ultimi 7 giorni
`BYTES`	totale dei bytes trasferiti
`BYTES7`	bytes trasferiti negli ultimi 7 giorni
`FIRSTDATE`	data della prima richiesta
`DATE`	data della richiesta più recente
`ALPHABETICAL`	ordine alfabetico
`RANDOM`	casuale (può essere utile nel caso di report lunghi per velocizzarli)

Per esempio,

```
HOSTSORTBY ALPHABETICAL
```

ordinerà lo "Host Report" in ordine alfabetico (io preferisco per numero di richieste).

Per molti report è possibile poi specificare il `FLOOR` (in inglese "pavimento"), cioè il limite minimo da prendere in considerazione (quelle al di sotto di tale limite verranno raggruppate in fondo come "altro"). A seconda del report, si può indicare un `FLOOR` di bytes trasferiti, di pagine richieste o di percentuale.

Report relativi ai visitatori

Altri report ottenibili riguardano i visitatori del sito. Ecco quelli che mi servono:

```
DOMAIN ON
ORGANISATION ON

HOST ON
HOSTCOLS R
HOSTSORTBY REQUESTS

REFSITE ON
```

Io voglio anche il report sui referrer, ma le opzioni selezionate di default non mi soddisfano: di default vengono visualizzati solo quelli con almeno 20 richieste, mentre io le voglio tutte. Devo allora specificare il limite minimo:

```
REFERRER ON
REFFLOOR 0R
```

La `R` dopo lo `0` indica le richieste (potevo scegliere anche `P` per le pagine, o `Mb` per i Megabytes), lo `0` non impone limiti. Se avessi scritto `5R`, il limite minimo sarebbe stato quindi `5` richieste.

Non mi interessa però il grafico di questo report, perché diventa brutto da vedere se ci sono molte pagine con una sola richiesta:

```
REFCHART OFF
```

Non mi interessano i referrer interni al mio sito, mi basta sapere da quali siti/pagine provenivano i visitatori (cioè come hanno fatto a trovare il mio sito):

```
REFREPEXCLUDE http://frech.altervista.org/*
REFREPEXCLUDE http://www.frech.altervista.org/*
REFSITEEXCLUDE http://frech.altervista.org/
REFSITEEXCLUDE http://www.frech.altervista.org/
```

Potreste chiedervi come mai alla fine delle ultime due righe, dopo la barra, non sia stato messo l'asterisco. Il motivo è che si tratta di *siti*, e io voglio escludere il mio. Le prime due righe

riguardano invece delle *pagine*: per quel report io voglio escludere ogni singola riga in cui compaia l'indirizzo del sito nel campo dei referrer, quindi l'asterisco deve essere incluso nel comando.

Proseguiamo poi con i report relativi ai browser e ai sistemi operativi usati dai visitatori:

```
BROWSERREP ON
BROWREPCHART OFF

BROWSERSUM ON
BROWSUMCHART OFF

OSREP ON
OSCHART OFF
```

Qui non ho bisogno di alcuna configurazione extra, mi vanno bene quelle di default.

Report relativi alle richieste dei visitatori

Voglio anche un report sulle pagine richieste dai visitatori, tutte quante, senza limiti minimi di richieste:

```
REQUEST ON
REQFLOOR 0R
```

Anche qui c'è qualcosa da cambiare: selezionando solo queste opzioni, viene prodotta una tabella che riporta anche la data dell'ultima visita, ma a me questo non interessa. Devo allora cambiare l'elenco delle colonne:

```
REQCOLS R
```

Volendo, potrei anche richiedere statistiche sui tipi di file richiesti o sulla loro dimensione, ma a me non occorrono analisi del genere.

Altre configurazioni: ricerche DNS

A volte, a seconda di come è configurato il server che ospita il vostro sito, il file di log non riporta gli indirizzi di provenienza dei visitatori sotto forma di nome di dominio bensì in forma di indirizzo IP numerico. Si può dire ad Analog di effettuare questa conversione per noi, ma su alcuni sistemi questa ricerca non è possibile. Inoltre queste operazioni rendono le analisi più lente, dato che Analog deve connettersi alla rete e cercare di convertire gli indirizzi uno ad uno; pertanto ricorre all'uso di un file per non doverli cercare ogni volta. Questo file lo si indica ad esempio così:

```
DNSFILE dnscache.txt
```

Se non viene specificato un percorso completo, il file viene creato nella cartella in cui Analog si aspetta di trovarlo (per esempio sotto Windows si tratta della stessa cartella del programma).

Non basta però indicare semplicemente il file, bisogna anche specificare in che modo Analog lo debba usare. Ci sono quattro possibilità:

- `DNS NONE` nessuna ricerca viene effettuata;
- `DNS READ` Analog legge il file ma non effettua nuove ricerche (comodo per chi usa Analog senza connettersi ad Internet);
- `DNS WRITE` Analog legge il file vecchio, effettua la ricerca e aggiunge i risultati al file; in questo caso, ovviamente la prima volta si riceverà un avviso perché il file manca, ma esso esisterà le volte successive;
- `DNS LOOKUP` Analog legge il file vecchio ed effettua le ricerche senza però aggiungere i risultati al file, quindi essi andranno perduti al successivo utilizzo; normalmente questa opzione non viene usata, ma se per qualche motivo `DNS WRITE` dovesse fallire Analog la userà.

Bisogna ricordarsi che se si usa un comando `HOSTEXCLUDE` bisogna escludere l'indirizzo IP numerico se esso non può essere convertito, il nome in caso contrario. Insomma, bisogna escludere il modo in cui quel dato host è conosciuto nello "Host Report".

Se due copie di Analog dovessero scrivere nello stesso file DNS allo stesso tempo, queto file potrebbe essere danneggiato, perciò Analog usa un *file lock* per far sì che le altre copie del programma non scrivano in quel file nello stesso momento (cioè fa loro eseguire `DNS LOOKUP` invece di `DNS WRITE`). Il file *lock* viene specificato con

`DNSLOCKFILE filename`

Naturalmente bisogna fare in modo che tutte le copie di Analog usino lo stesso file di *lock*, perlomeno se hanno tutte lo stesso file di DNS.

Un'altra cosa: Analog non rimuove mai nulla dal file di DNS, quindi quest'ultimo continuerà a crescere. Bisognerà allora rimuovere le prime righe ogni tanto.

Ci sono poi due parametri per indicare quando devono "durare" i risultati di queste ricerche di DNS.

`DNSGOODHOURS 672`

farà in modo che le ricerche che hanno avuto successo vengano ricontrollate ogni 672 ore, cioè 4 settimane. Il valore di default è un numero molto grande, di modo che non vengano ricontrollate per molto tempo (o finché non vengono rimosse dal file di DNS).

`DNSBADHOURS 500`

farà lo stesso per le ricerche fallite, dopo 500 ore (il default è 336 ore, cioè 2 settimane).

Capire i report di Analog

La struttura dei report

I numeri tra parentesi nel "General Summary" sono per gli ultimi 7 giorni. Se tutte le richieste (o nessuna di esse) ricade in tale intervallo, non vi saranno numeri tra parentesi.

Nel "Domain Report", la dicitura "`domain not given`" significa che lo hostname non aveva al suo interno un punto, mentre "`Unknown domain`" significa che aveva il punto ma che il programma non ha trovato il dominio nel file che li elenca o che non è riuscito a leggere tale file.

Alcuni report hanno una struttura gerarchica. Ad esempio, nel deport dei domini potreste avere molte richieste da parte di siti `.com`, ma in tabella troverete, al di sotto di tale dicitura, quelle più "particolari". Ecco un esempio:

```
Listing the first 5 domains by the number of requests, sorted by the number of requests.
no.| #reqs| domain
---| -----| ------
  1| 13243| .com (Commercial)
   |  1262|   aol.com
  2| 11783| .jp (Japan)
   |  9592|   ad.jp
   |  1043|   co.jp
  3| 10073| .net (Network)
   |  1926|   uu.net
  4|  9657| [unresolved numerical addresses]
  5|  7388| .uk (United Kingdom)
   |  5792|   ac.uk
   |  1510|   co.uk
   | 18502| [not listed: 82 domains]
```

Noterete che i domini elencati sono 5 (rispettando quanto scritto nell'intestazione) e che al di sotto di ogni dominio vi sono i sotto-domini. Alla fine della tabella si trovate anche la dicitura "`not listed`": essa raggruppa quegli elementi che non hanno ricevuto un numero sufficiente di richieste per superare il limite minimo (`FLOOR`) richiesto per l'analisi; *non comprende* invece quelli esclusi esplicitamente con `EXCLUDE`: questi ultimi non sono stati nemmeno presi in considerazione, come se non fossero mai stati presenti fisicamente nel file di log.

Alcuni potrebbero voler visualizzare nel "Referrer Report" i referrer con tanto di argomenti di ricerca: bisogna allora dare a `REFARGSFLOOR` un valore basso. Se invece non si vuole vedere tali argomenti, si può impostare un valore alto oppure usare `REFARGSEXCLUDE`.

Gli script che causano una redirezione saranno visualizzati nel "Redirection Report", *non* nel "Request Report".

Coi file PDF si possono avere dei dati "falsati", perché il reader li scarica in piccole parti ognuna delle quali conta come richiesta. Non è nemmeno detto che ognuna di queste parti corrisponda a una pagina: il programma potrebbe anche scaricare l'intero documento in un colpo solo, e del resto il visitatore potrebbe voler vedere solo una parte del documento.

Cosa si intende per...?

Per *host* si intende il computer che ha richiesto quel dato file. Tale file può essere una *pagina* o anche un'altra cosa (come un'immagine). Di default i nomi che terminano in `.html`, `.htm` oppure `/` contano come pagine (come vi ho detto prima, però, possiamo includere altri file come pagine con il comando `PAGEINCLUDE`).

Per *richieste totali* si intendono tutti i file richiesti (pagine, immagini...: tutto). Un utente può generare molte richieste sia richiedendo molti differenti files sia richiedendo lo stesso file più volte.

Il *referrer* per una richiesta è il posto da cui l'utente ha richiamato quel file: se ha richiamato la pagina X cliccando su un collegamento presente sulla pagina Y allora il referrer è Y; se si tratta di un'immagine presente in una pagina X, il referrer per tale immagine è la pagina X stessa.

Come funziona Internet

Adesso vediamo di capire alcune cose che non dipendono da Analog né da nessun altro programma di questo tipo, ma che possono aiutarci a capire meglio le statistiche.

Supponiamo che qualcuno trovi il mio sito su un motore di ricerca, clicchi sul collegamento per visitare la mia home page, gironzoli qua e là per un po' e poi esca. Ebbene, dal log io posso sapere a che ora e in che giorno quel visitatore si è collegato alla mia home page e provenendo da quale pagina/sito, l'indirizzo IP del suo computer, che sistema operativo e che browser usa. Nient'altro. Scordatevi quindi di poter conoscere l'e-mail del visitatore. E scordatevi anche di capire che percorso ha seguito un certo visitatore all'interno del sito da quando ci è entrato a quando ne è uscito.

Un appunto riguardo al numero delle richieste: se, ad esempio, la mia home page ha 5 immagini, e qualcuno la visita, nel log verranno registrate ben 6 richieste, cioè una per la pagina principale e una per ogni immagine presente nella pagina. Il referrer per tali immagini sarà la pagina principale stessa. Ecco allora che il numero di richieste con questo referrer possono trarre in inganno.

Normalmente un browser mette le immagini e le pagine in una directory detta cache per non doverle scaricare di nuovo alla visita successiva. Ho detto *normalmente* perché un browser può anche essere impostato in modo tale da ricaricare le pagine ogni volta. Inoltre, dato che praticamente ogni ISP ha una sua cache, se un visitatore guarda la mia home page e un altro visitatore dello stesso ISP ha fatto di recente la stessa cosa, la cache potrebbe averla salvata e quindi potrebbe mostrarla al secondo visitatore senza che io ne sappia nulla (questo indipendentemente dalle impostazioni dei browser, perché queste ultime lavorano solo sul computer dell'utente e non su quello dell'ISP).

Molti browser "mentono" sulla propria identità, o lasciano che l'utente configuri il nome di tale browser. Per non parlare di quei servizi di "anonymizers" per navigare senza essere identificati, facendo quindi registrare nei log browser e referrer fasulli.

Nel caso di connessioni dial-up, cioè quelle effettuate con un normale modem analogico, uno stesso indirizzo IP può essere assegnato a due persone diverse che si connettono dopo un certo tempo l'una dall'altra (si parla infatti di indirizzi IP dinamici). Non è quindi detto che trovare nel logfile molte richieste provenienti da uno stesso indirizzo IP, ad una certa distanza l'una dall'altra, significhi che si tratti dello stesso visitatore.

Insomma, Analog (e qualunque programma del genere) si limitano ad analizzare dei file di log e presentare delle statistiche sul loro contenuto in una forma a noi comprensibile. Ma sappiamo bene che "la statistica è quella scienza secondo cui se una persona ha mangiato un pollo e un'altra non ne ha mangiato neanche uno, entrambe hanno mangiato mezzo pollo", quindi si parla di approssimazioni; si parla di ciò che è registrato nel log, cioè di ciò che è avvenuto nel server, non di ciò che gli utenti hanno fatto effettivamente. E' utile studiare le statistiche, ma ricordiamo che non rappresentano appieno la realtà.