

Analisi del traffico su un sito Internet

- Perché è possibile
- Che cosa possiamo sapere
- Che cosa non possiamo sapere
- Perché è importante l'analisi
- Gli strumenti di analisi
- Un'analisi dettagliata

Perché è possibile analizzare il traffico di un sito

- Ogni sito ha un nome che corrisponde a un numero IP
- La navigazione di un utente su un sito corrisponde a una richiesta fatta al *web server* da parte del computer dell'utente che è identificato da un IP
- Il *web server* può memorizzare nel *log* quale IP lo ha interrogato e conservare traccia di altri dati caratteristici
- Confronto con Auditel per la TV

Che cosa possiamo sapere

- IP dell'utente
 - Posso dedurre qualcosa sulla località di provenienza della richiesta
- Data e ora
- In quale pagina c'era il collegamento che ha portato al nostro sito
- Quale ricerca sui motori ha portato al nostro sito
- Quanti dati sono stati trasmessi
- Il tempo impiegato
- Altri dati tecnici

Che cosa non possiamo sapere

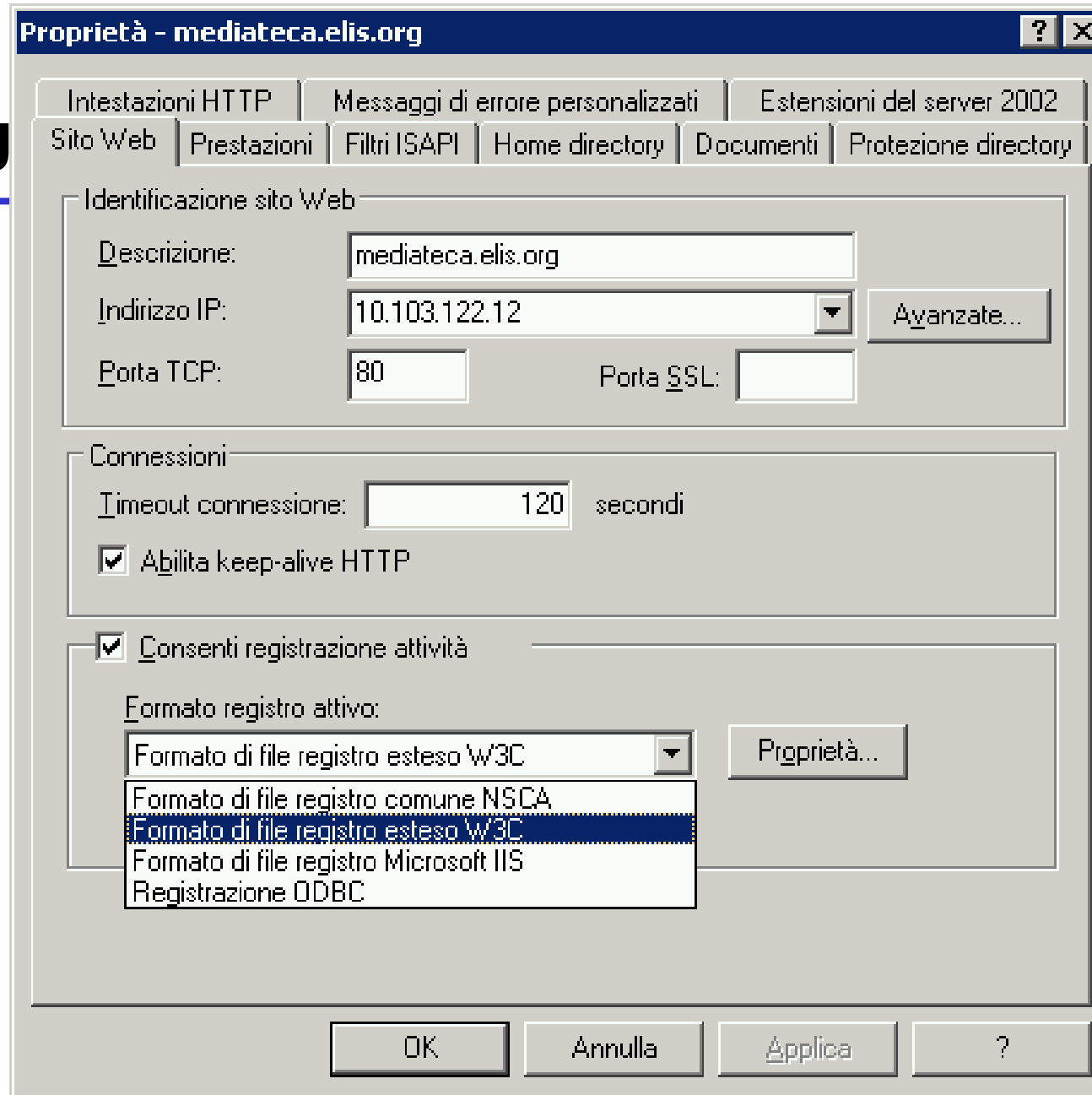
- L'identità del visitatore
- Il posto preciso da cui naviga
- Se ha letto il contenuto della pagina su cui ha navigato
- Il vero numero di visitatori
 - Un visitatore può usare IP diversi in momenti diversi
- Il vero numero di visite
 - Ogni immagine scaricata può contare come una visita
 - Un utente può usare la cache del suo browser per rivedere una pagina
 - Ci sono cache intermedie gestite dagli ISP
- Quanto tempo ha dedicato al mio sito

Perché è importante l'analisi

- Studiare la tendenza
 - Stiamo acquisendo popolarità
- Quali sono le parole chiave che la gente cerca per trovare il nostro sito
 - Diverse per ogni motore di ricerca
- Quali sono i siti che ci citano
 - Opportuni e inopportuni
- Quante e quali pagine lo stesso utente richiede in una sessione
 - Dove mettere i contenuti più attraenti

Esempio di log

Con IIS 6 di Windows Server 2003 si può scegliere il formato di *log*



IIS 6

Tutti i dati che si possono registrare con il *log* formato esteso W3C (standard modificato da Microsoft)

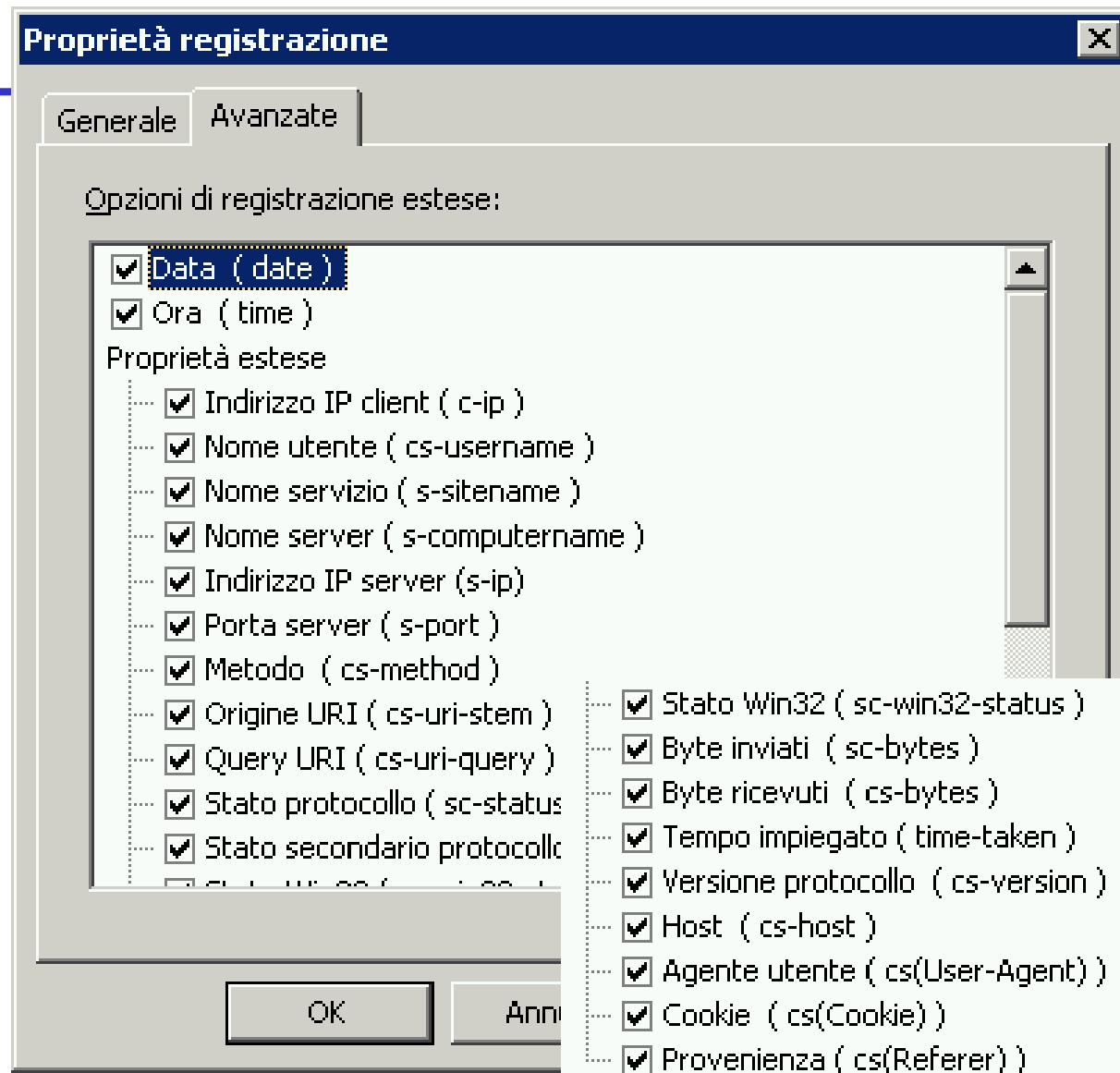
Azioni del

c=client

s=server

cs=client verso il server

sc=server verso il client



Le definizioni dei campi registrati

date	La data
time	L'ora del server
c-ip	L'indirizzo IP del client che ha acceduto al server
cs-username	Il nome dell'utente (quando c'è di mezzo una autenticazione). Se è anonimo (caso più frequente) c'è un -
s-sitename	Il servizio Internet come identificato dal server
s-computername	Il nome del server
s-ip	L'indirizzo IP del server
s-port	Il numero di porta del servizio sul server
cs-method	Il metodo di azione del client (per esempio un GET).
cs-uri-stem	La risorsa chiamata, per esempio la pagina HTML

Le definizioni dei campi registrati

cs-uri-query	La query o interrogazione (quando è il caso) che il client fa al server
sc-status	La risposta del server HTTP (operazione riuscita, fallita, ecc.)
sc-win32-status	La risposta del server secondo Microsoft
sc-bytes	Il numero di byte inviati dal server
cs-bytes	Il numero di byte ricevuti dal server
time-taken	La durata dell'azione in millisecondi
cs-version	La versione del protocollo HTTP usato
cs-host	Il nome del sito web
cs(User-Agent)	Il browser usato dal client
cs(Cookie)	Il contenuto del cookie inviato o ricevuto dal server
cs(Referer)	Il sito sul quale c'è il link usato dal client per raggiungere il server

Due record del log

Un utente scarica un file dalla pagina Download.asp della Mediateca-ELIS

date	31/03/2004	31/03/2004
time	22.48.30	22.49.01
s-sitename	W3SVC1159930175	W3SVC1159930175
s-computername	WEB-ELIS	WEB-ELIS
s-ip	10.103.122.12	10.103.122.12
cs-method	GET	GET
cs-uri-stem	/Download.asp	/software/Nm3.01.build3388.ita.exe
cs-uri-query	-	-
s-port	80	80
cs-username	-	-
c-ip	80.181.165.87	80.181.165.87
cs-version	HTTP/1.1	HTTP/1.1

Due record del log

cs(User-Agent)	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+.NET+CLR+1.1.4322)	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+.NET+CLR+1.1.4322)
cs(Cookie)	-	ASPSESSIONIDAQDTBDAR=KGAABAGBEE NFOPDIAGKDAJAI
cs(Referer)	http://www.google.it/search?hl=it&ie=UTF-8&oe=UTF-8&q=download+netmeeting+in+italiano&lr=	http://mediateca.elis.org/Download.asp
cs-host	mediateca.elis.org	mediateca.elis.org
sc-status	200	200
sc-substatus	0	0
sc-win32-status	0	0
sc-bytes	15250	1650764
cs-bytes	490	515
time-taken	421	25468

Da che città navigava l'utente

- Non riusciamo a capire la città, ma abbiamo normalmente dati per sapere la nazione
- Telecom Italia può saperlo e lo rivela all'autorità se richiesto

Report per 80.181.165.87 [host87-165.pool80181.interbusiness.it]								
Analisi: Pacchetti IP persi sulla rete "Interbusiness infrastrutturale" al hop 11. Informazioni insufficienti nella cache per la determinazione della rete successiva al hop 12.								
Hop	%Pers	Indirizzo IP	Nome nodo	Localizzazione	F.Ora	ms	Grafico	Rete
0		194.242.61.53	interhost.it	*			0 219	HostingSolutions.it
1		194.242.61.1	-	Florence, Italy	+01:00	0		HostingSolutions.it
2		62.94.32.145	-	(Italy)	+01:00	0		GENESYS INFORMAT
3		62.94.209.41	-	Arezzo, Italy	+01:00	7		Eutelia
4		62.94.31.38	a1-0-10.mi15.e	Arezzo, Italy	+01:00	15		Eutelia
5		217.29.66.35	interbusiness.	-		0		Milan Internet eXchang
6		151.99.98.225	-	(Netherlands)	+01:00	8		RIPE Network Coordin
7		151.99.75.157	r-mi258-vl3.op	Milan, Italy	+01:00	103		RIPE Network Coordin
8		151.99.98.97	r-rm213-mi258	Rome, Italy	+01:00	31		RIPE Network Coordin
9		151.99.101.19	r-fi63-rm213.oj	Florence, Italy	+01:00	22		RIPE Network Coordin
10		151.99.98.90	r-fi67-fi63.opb.	Florence, Italy	+01:00	28		RIPE Network Coordin
11		80.19.134.153	-	(Italy)	+01:00	24		Interbusiness infrastru
...								
?		80.181.165.87	host87-165.po	-				Telecom Italia S.p.A.

Da che sito proveniva l'utente

- Lo deduciamo dal campo cs(Referer)
 - **www.google.it**/search?hl=it&ie=UTF-8&oe=UTF-8&q=download+netmeeting+in+italiano&lr=
- Possiamo sapere anche che cercava le parole
 - *download netmeeting in italiano*
- Referer è il dato registrato più interessante ai fini della promozione di un sito
 - Studiare gli effetti delle parole chiave che abbiamo impostato
 - Qual è il motore di ricerca dal quale provengono più richieste
 - Andare a verificare in che posizione è il nostro sito in risposta a quelle parole

Alcuni strumenti per l'analisi

- Analog – www.analog.cx
 - Si installa sul server
 - Necessario poter accedere direttamente al server
 - Uno dei più diffusi
 - Gratuito per tutti i tipi di server
- Freestats – www.freestats.com
 - Si inserisce un codice nelle pagine del sito
 - È utilizzabile senza accedere al server
 - Gestito da Freestats
 - Gratis con pubblicità nelle statistiche
 - A pagamento servizi più completi

Analisi del traffico su mediateca.elis.org

Pagina html generata da Analog, eseguito periodicamente

Programma attivato Sab 29-Mag-2004 alle 00:24.

Analizzate le richieste da Lun 26-Gen-2004 alle 14:45 a Ven 28-Mag-2004 alle 22:16 (123,31 giorni).

Inizio | [Sommaro generale](#) | [Resoconto trimestri](#) | [Resoconto mesi](#) | [Resoconto settimane](#) | [Resoconto giorni](#) | [Sommaro giorni della settimana](#) | [Resoconto ore](#) | [Resoconto per ore della settimana](#) | [Sommaro ore del giorno](#) | [Resoconto domini](#) | [Resoconto organizzazioni](#) | [Resoconto host](#) | [Resoconto host richieste reindirizzate](#) | [Resoconto host richieste fallite](#) | [Resoconto utenti](#) | [Resoconto redirezione utenti](#) | [Resoconto insuccessi utenti](#) | [Resoconto provenienze richieste reindirizzate](#) | [Resoconto provenienze richieste fallite](#) | [Resoconto provenienze richieste](#) | [Resoconto siti di provenienza](#) | [Resoconto richieste ricerca](#) | [Resoconto parole ricerca](#) | [Resoconto browser](#) | [Sommaro browser](#) | [Resoconto sistemi operativi](#) | [Resoconto codici di stato](#) | [Resoconto tempi di esecuzione](#) | [Resoconto dimensione dei file](#) | [Resoconto tipi di file](#) | [Resoconto directory](#) | [Resoconto reindirizzamenti](#) | [Resoconto insuccessi](#) | [Resoconto richieste](#)

Statistiche di mediateca.elis.org

I valori in parentesi si riferiscono a i 7 giorni fino al 29-Mag-2004 00:24.

Richieste soddisfatte: 28.561 (1.381)

Media giornaliera di richieste soddisfatte: 231 (197)

Linee nel log file senza codice di stato: 9.279 (0)

Richieste di pagine soddisfatte: 22.734 (1.133)

Media giornaliera di richieste di pagine soddisfatte: 184 (161)

Richieste fallite: 4.115 (233)

Richieste reindirizzate: 2.500 (71)

File distinti richiesti: 933 (217)

Host distinti serviti: 6.444 (333)

Registrazioni nel log file non desiderate: 118.540

Quantità totale di traffico: 3,36 gigabytes (129,85 megabytes)

Traffico medio giornaliero: 27,88 megabytes (18,55 megabytes)

Andamento settimanale

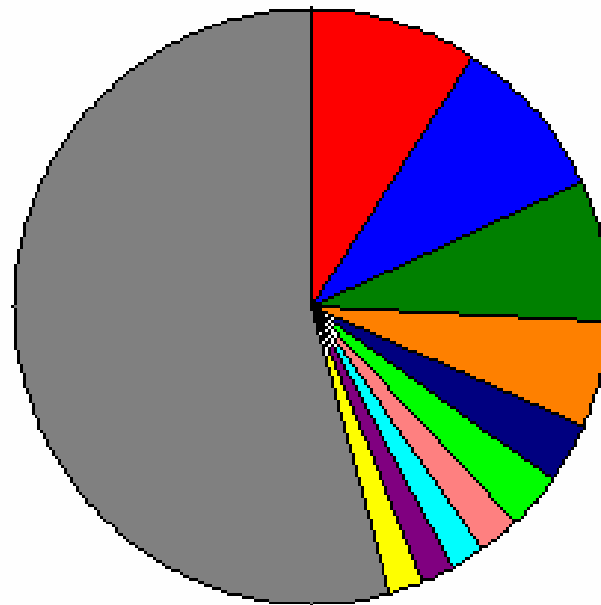
- La definizione di *pagine* è personalizzabile (.htm, .asp, ecc.)
- Permette di non contare le immagini che sono incluse fra le richieste

in. sett.	#rich	#pag.	
25-Jan-04	1334	1013	
1-Feb-04	1576	1234	
8-Feb-04	1798	1434	
15-Feb-04	1437	1152	
22-Feb-04	1728	1468	
29-Feb-04	1499	1282	
7-Mar-04	1746	1431	
14-Mar-04	1790	1501	
21-Mar-04	1406	1033	
28-Mar-04	1470	1169	
4-Apr-04	1082	847	
11-Apr-04	1309	981	
18-Apr-04	1590	1215	
25-Apr-04	2062	1563	
2-May-04	2604	2041	
9-May-04	1661	1353	
16-May-04	1234	1010	
23-May-04	1235	1007	

Le frasi più cercate

Non è l'obiettivo della Mediateca-ELIS mettere a disposizione *netmeeting* per tutti

Risulta ai primi posti su Google alla richiesta *netmeeting italiano*



- netmeeting download
- mediateca
- netmeeting italiano
- download netmeeting
- download real player
- download netmeeting italiano
- real player download
- compilatore c download
- download compilatore c
- netmeeting italiano download
- *

Le aree sono disegnate per numero di richieste.

elenco delle prime 30 ricerche in ordine di numero di richieste, in ordine di r

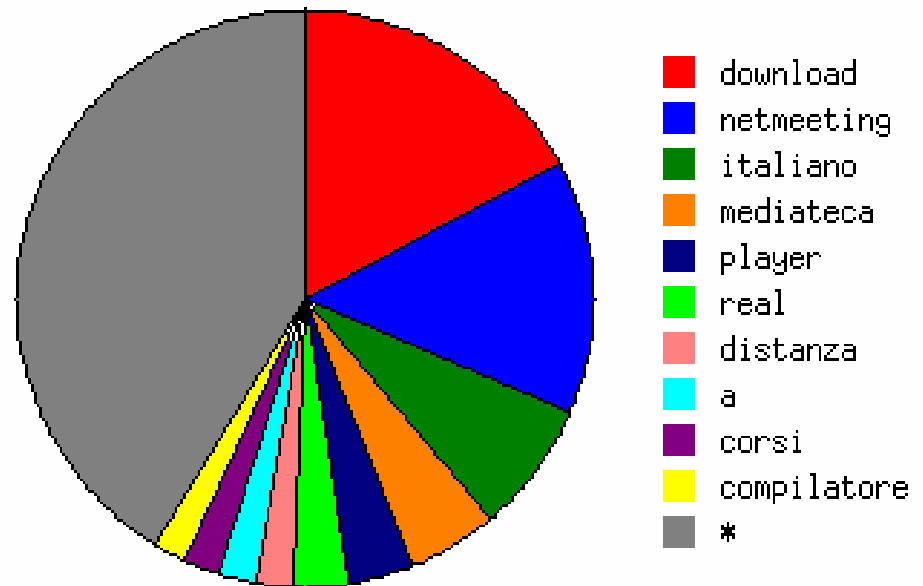
#rich	chiave ricerca
276	netmeeting download
269	mediateca
229	netmeeting italiano

Le singole parole più cercate

La parola *distanza* è presente fra le prime

E' un obiettivo rispondere a richieste del tipo *formazione a distanza*

La frase *corsi a distanza* ha 37 richieste



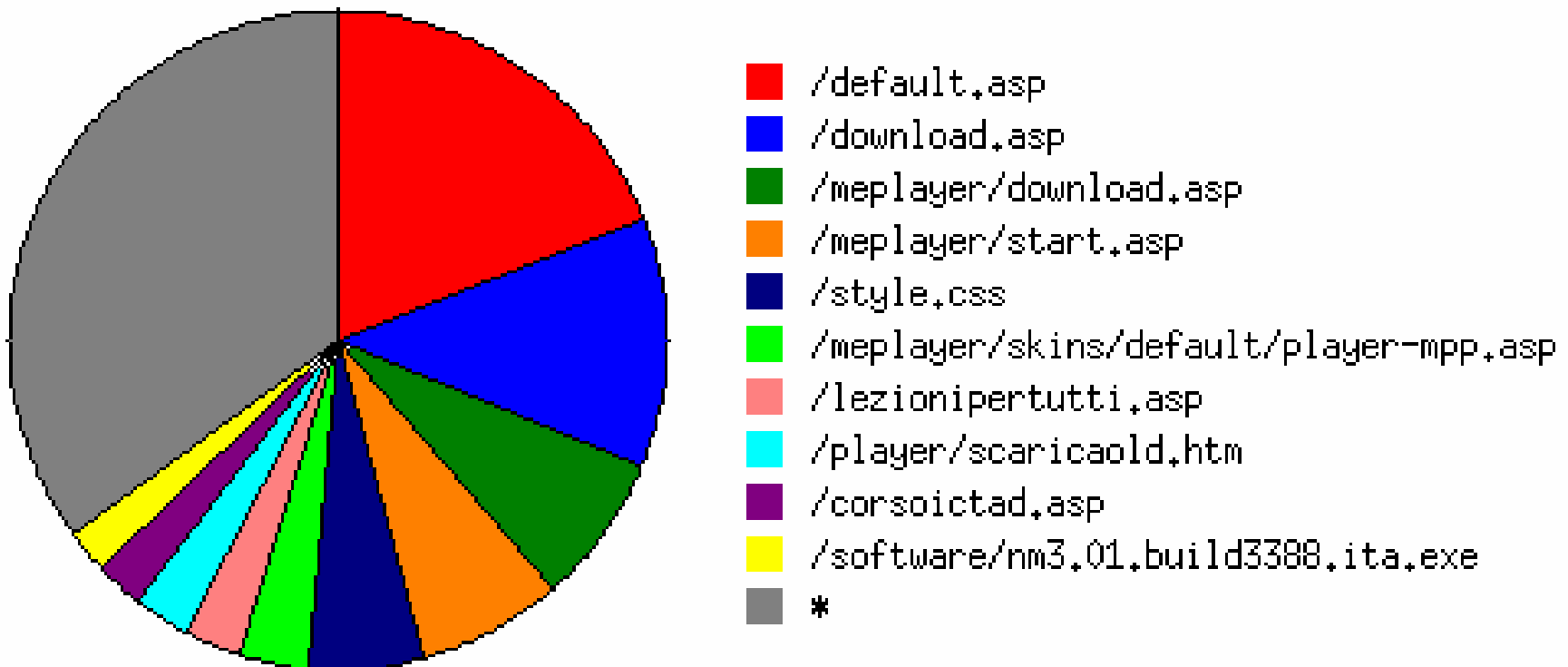
Le aree sono disegnate per numero di richieste.

elenco delle prime 30 parole cercate in ordine di numero

#rich	chiave ricerca
1363	download
1107	netmeeting
591	italiano

Le pagine più richieste

- Non tutti partono dalla pagina principale *default.asp*
- La pagina *style.css* potrebbe essere esclusa dalle statistiche perché inutile
- *corsoictad.asp* mostra il successo di un corso
- *lezionipertutti.asp* sono videolezioni aperte a chiunque

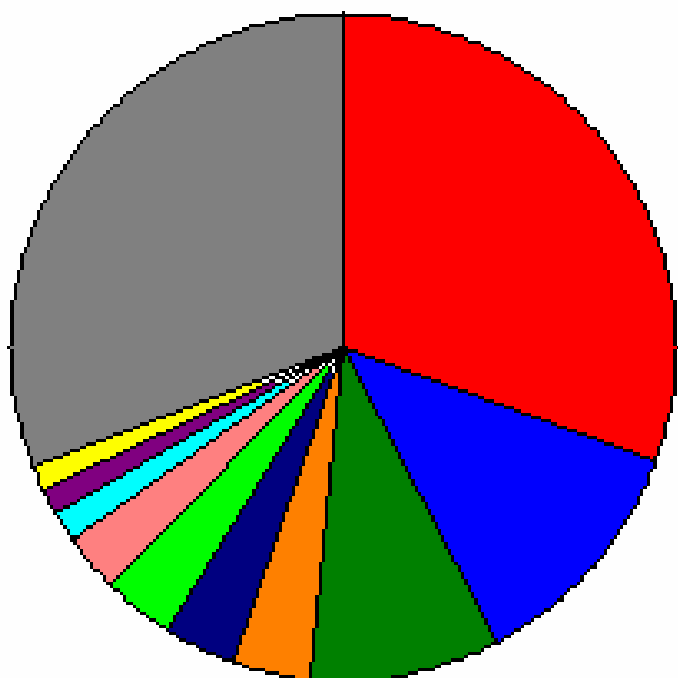


Domini di provenienza

#rich	%byte	dominio
13490	69,58%	.it (Italia)
3687	9,71%	[indirizzo numerico non risolto]
2587	8,28%	.net (Network)
4000	6,94%	.com (Commerciale)
2766	1,51%	[dominio non fornito]
905	0,82%	.org (Organizzazioni Non-Profit)
29	0,54%	.br (Brasile)
37	0,54%	.ch (Svizzera)
16	0,49%	.th (Thailandia)
7	0,29%	.sa (Arabia Saudita)
24	0,20%	.ar (Argentina)
2	0,13%	.jo (Giordania)
163	0,12%	.mx (Messico)

I siti di provenienza

- Google in testa, Virgilio e MSN tra i primi
- *131.x.x.x* e *mlsandboxint* sono i server del corso ICTAD2
- www.mediateca.com è un alias di mediateca.elis.org



- <http://www.google.it/search>
- <http://131.175.5.236/bin/common/lesson.pl>
- <http://www.mediateca.com/>
- <http://www.mediateca.com/Download.asp>
- <http://mlsandboxint.black....com/bin/common/lesson.pl>
- <http://www.elis.org/>
- <http://search.virgilio.it/search/cgi/search.cgi>
- <http://login.elis.org/Login.asp>
- <http://www.eforum.it/Methodologie/elearning.asp>
- <http://search.msn.it/results.aspx>
- *

Sistemi avanzati di analisi

- Permettono di tracciare il percorso di navigazione fatto dall'utente nel nostro sito
- Consentono di sapere quali parti della pagina attirano di più l'utente
- Approfondiscono la conoscenza del luogo di provenienza
 - Negli USA è più facile perché gli utenti hanno connessioni permanenti con IP fisso

Consigli finali

- Non fissarsi sui numeri assoluti ma analizzare la tendenza
 - Verificare l'effetto delle nostre campagne promozionali
 - Verificare l'effetto del cambiamento di posizione nei motori di ricerca
- Fare a gara con altri siti più importanti cercando di aumentare il traffico senza badare a chi frequenta il nostro sito può essere inutile