Istituto di Ricerche sulla Popolazione e le Politiche Sociali - CNR

IRPPS Working Papers

ISSN 2240-7332

Un anno in rete: analisi statistica del *report log* del sito www.irpps.cnr.it

Loredana Cerbara e

Maria Girolama Caruso

What is IRPPS?

IRPPS is an Interdisciplinary Research Institute that conducts studies on demographic and migration issues, welfare systems and social policies, on policies regarding science, technology and higher education, on the relations between science and society, as well as on the creation of, access to and dissemination of knowledge and information technology.

www.irpps.cnr.it

IRPPS WPs n. 02/2004

Questo lavoro è stato realizzato grazie al contributo del Servizio Reti di Comunicazione del CNR che ha fornito i dati di base per le elaborazioni. Il lavoro è il primo passo di un progetto più ampio che prevede sia il monitoraggio degli accessi e dell'utilizzo del sito dell'Istituto di Ricerche sulla Popolazione e le Politiche Sociali, sia lo sviluppo di tecniche innovative di *data web mining*.

Indice

1. Introduzione

La semplice pubblicazione in rete di pagine web non assicura che siano disponibili utenti interessati ad usufruirne. Di solito si ricorre a stratagemmi che aumentano la 'visibilità' di un sito, (indicizzazione sui principali motori di ricerca, inserimento di 'meta tag' contenenti parole chiave, campi di testo nella pagina, ecc.) per assicurarsi la migliore visibilità possibile. Ciò non garantisce comunque che il sito venga visitato e non fornisce informazioni circa gli utenti e l'utilizzo che se ne fa. Il reperimento e l'analisi di tali informazioni è di particolare interesse per il miglioramento continuo delle pagine pubblicate indipendentemente dal fatto che, anche se non fosse mai visitato, il sito di un istituto di ricerca avrebbe la sua ragione di esistere già nella stessa diffusione gratuita della descrizione delle attività di ricerca svolte.
Il server che ospita le nostre pagine web registra quotidianamente i dati delle visite alle nostre pagine in un file, detto file di log, che può essere messo a disposizione per analisi statistiche e che, in assenza di altri tipi di rilevazione di informazioni sui nostri visitatori (come ad esempio fanno i siti che richiedono iscrizioni volontarie dei visitatori chiedendo loro dati di tipo strutturale), è la nostra unica fonte di dati.
Ma mentre è prassi ormai consolidata quella di analizzare le tracce lasciate dai visitatori dei siti commerciali o di interesse generale che sono oggi presenti sul web, per gli istituti di ricerca, che agiscono in settori ad alta specializzazione, non si richiede di norma questo tipo di analisi. In realtà, dal momento che un istituto come l'IRPPS può presentarsi come fornitore di servizi da mettere a disposizione attraverso la rete mondiale, vale la pena di condurre anche in questo caso una analisi dei files di log e delle tracce lasciate su di essi dai visitatori del sito. Lo studio del profilo degli utenti può avere ricadute inaspettate, come quella di poter monitorare e veicolare un canale di comunicazione scientifica dalle caratteristiche eccezionali, come è quello offerto da Internet.
Con questo primo lavoro, che ha scopi essenzialmente esplorativi, vogliamo abbozzare una descrizione degli utenti del sito dell'IRPPS ed eseguire una preliminare analisi statistica del materiale più richiesto. Ciò, oltre a dare informazioni su quali siano i settori della ricerca dell'IRPPS di maggior interesse per il pubblico on line, può produrre anche linee guida per gli sviluppi futuri sia di tipo architetturale e strutturale del sito, che di tipo formale e di contenuto. Possiamo cioè tentare di rispondere a domande del tipo: quali utenti ci visitano?; che tipo di informazione ci richiedono?; quali servizi possiamo pensare per i nostri utenti?
Abbiamo la possibilità di usare tecniche di *data web mining* che sono oggi disponibili e che consentono di estrarre facilmente le informazioni sui comportamenti di visita dagli archivi contenuti nel server (i file di log) anche se la particolare struttura del sito, non ancora consolidata dal punto di vista tecnico, ne consente al momento un uso piuttosto limitato. Infatti l'analisi non ha potuto cogliere tutti gli aspetti di nostro interesse proprio a causa della struttura delle pagine web le quali, essendo composte di

frame[1], non consentono la corretta rilevazione della navigazione nel sito. Inoltre la pagina principale era fin troppo scarna di collegamenti alle sezioni interne al sito, cosa che può aver determinato una sorta di 'effetto offuscamento' dei prodotti e servizi offerti on line e di conseguenza non ha contribuito a favorire la fedeltà degli utenti visitatori. Ovviamente trattandosi di un sito di natura non commerciale, l'interesse principale è quello di agevolare il reperimento delle informazioni e di garantire la massima chiarezza, più che di catturare visitatori fedeli.
Tuttavia questa analisi ha avuto come prima conseguenza la generazione di una nuova versione del sito stesso, molto semplificata nella struttura, ma, a nostro avviso, molto più efficiente perché rende immediatamente disponibile l'informazione di base di tutto il sito, proprio come se si trattasse di un 'portale' (termine ormai molto diffuso in ambito web che sta a rappresentare un luogo di ingresso privilegiato al modo della Rete, dove ogni navigatore possa trovare inizio ma soprattutto guida nella sua ricerca sul Web) sulle ricerche e le attività dell'Istituto.

2. Statistiche generali

2.1 Dati di sintesi

La prima analisi da effettuare è senza dubbio quella che riguarda il calcolo e la lettura delle statistiche generali. Queste sintetizzano l'informazione sulle visite al sito in maniera eccessivamente spinta, ma riescono a dare una misura, seppur approssimativa, dell'uso che si è fatto delle pagine pubblicate durante la finestra temporale di osservazione.
Le pagine visitate sono in totale quasi 180000 con una media giornaliera di 457 pagine. Ogni visitatore[2] ha visto quasi quattro pagine per un totale di 46000 visitatori. Abbiamo avuto in media 118 visitatori al giorno con 22558 indirizzi IP[3] diversi. Questi dati di per se riescono a dire soltanto che il sito è stato visitato pressoché giornalmente, ma non danno informazioni su chi siano i visitatori e su quali siano le pagine visitate. Ciò significa che in questi valori sono compresi, come era ovvio attendersi, anche tutte le visite degli interni all'istituto che hanno solitamente il sito in esame come pagina principale predefinita nel browser e pertanto visitano giornalmente la home

[1] Un frame è una suddivisione di una pagina htm, cioè scritta in un linguaggio di programmazione adatto alla pubblicazione sul web, in più riquadri, in ognuno dei quali è possibile far apparire un documento diverso.
[2] Qui per visitatore si intende 'visita' nel senso che è possibile che un visitatore sia conteggiato più volte, tanti quanti sono i suoi accessi al sito.
[3] IP: Internet Protocol, un numero a 32 bit composto da 4 cifre comprese tra 0 e 255 separate da punti. È fornito dal proprio amministratore di rete o dal *provider* dei servizi di rete. Si tratta di un particolare numero che ci viene assegnato ogni qual volta ci connettiamo in rete e che permette molto spesso di identificare alcune caratteristiche dell'utente visitatore.

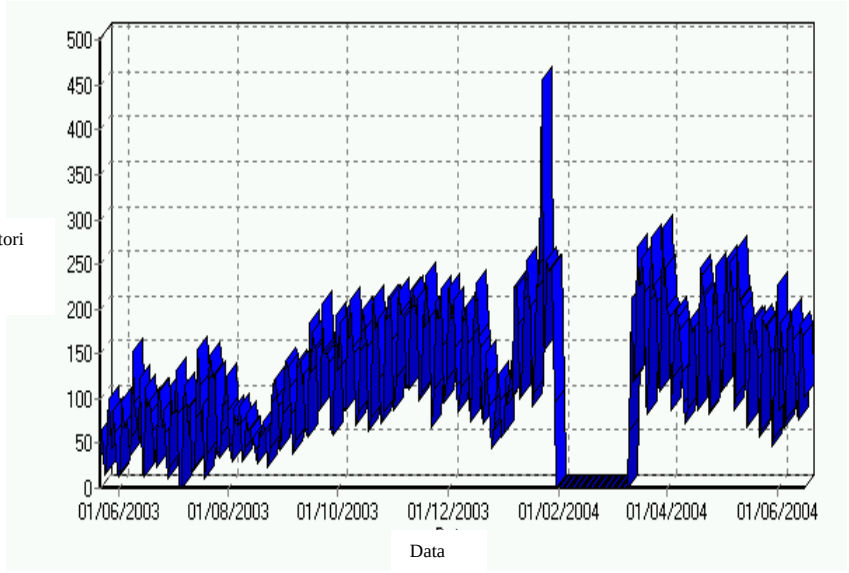
Pagine visitate dal 22/05/2003 al 16/06/2004	
Numero totale di pagine visitate	179.222
Numero medio di pagine visitate al giorno	457
Numero medio di pagine visitate per visitatore	3,87
Utenti visitatori dal 22/05/2003 al 16/06/2004	
Numero totale di visitatori	46.298
Numero medio di visitatori al giorno	118
Numero totale di IP	22.558

del sito non appena aprono il browser di navigazione in internet. Inoltre la struttura a frame del sito ha come effetto il fatto che semplicemente accedendo alla home page si visitino ben 3 pagine, quelle di cui è, appunto, composta la home page. Vedremo in seguito come sviscerare i dettagli di queste informazioni usando ulteriori output che siamo in grado produrre e come trarre da essi l'informazione che più ci interessa.

2.2 Profili temporali

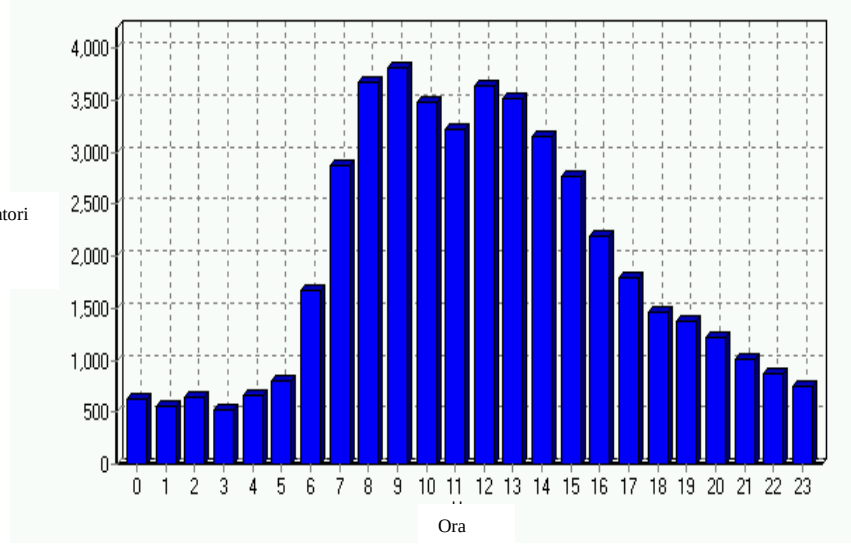
Il primo dato di dettaglio interessante che possiamo proporre è quello delle visite giornaliere al sito, un dato interessante per capire se ci sono stagionalità o, più in generale, periodi dell'anno in cui l'attività web è particolarmente rilevante.

Grafico 1. Visite giornaliere



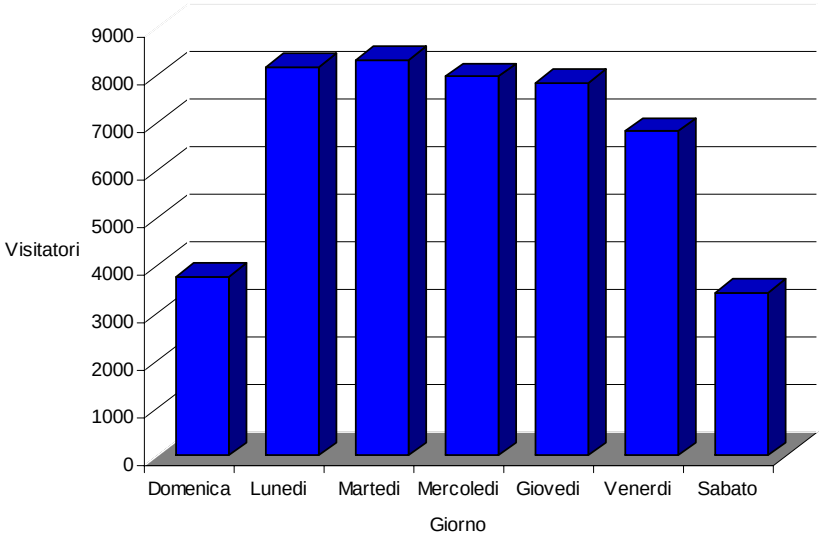
A parte un periodo compreso tra il mese di febbraio e quello di aprile 2004, in cui per problemi tecnici il file di log non ha raccolto i dati in modo corretto, si vede la tendenza alla crescita del numero di visite giornaliere. Tendenza che sembra in leggero calo negli ultimi mesi, anche se ancora a livelli superiori di quelli registrati nello stesso periodo dello scorso anno.

Grafico 2: Visite per ora del giorno.



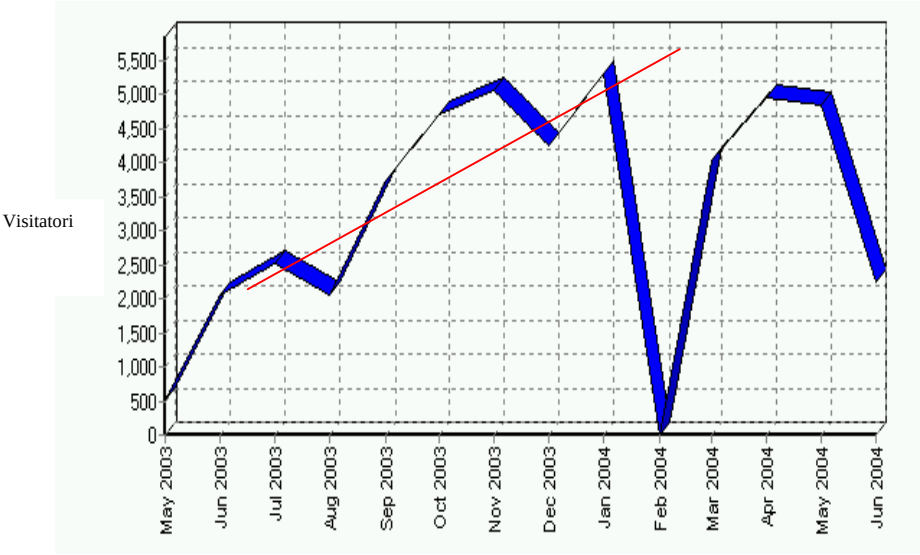
La distribuzione oraria delle visite al sito è interessante per capire quali siano le ore di punta e quanti sono stati i visitatori nelle ore di minor richiesta. Si nota immediatamente che la maggiore attività si riscontra in corrispondenza delle ore *di ufficio* canoniche per l'Italia, ma una attività marginale si riscontra anche in altre ore. Probabilmente si tratta dei contatti di provenienza internazionale oppure di attività dovuta a programmi automatici di ricerca (i famosi spider di cui si servono in genere i motori di ricerca per la raccolta periodica delle informazioni sul mondo web).

Grafico 3: Visite per giorno della settimana



Il grafico 3 mostra come siano più gettonati i giorni feriali a discapito del fine settimana in cui l'attività è più che dimezzata, anche se rimane a livelli di tutto rispetto per il traffico caratteristico del sito in esame.
Ovviamente le visite durante il fine settimana possono essere assoggettate a diversi tipi di vistatori: utenti che preferiscono collegarsi in momenti di minor traffico in Internet e quindi a maggiore velocità oppure programmi automatici dei motori di ricerca.

Grafico 4: Numero mensile di visitatori

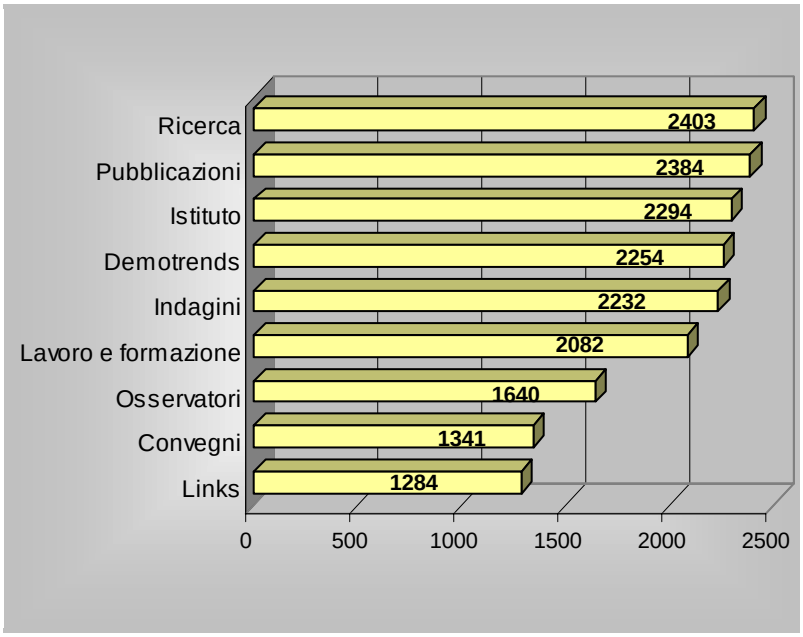


Il numero di visitatori al mese è essenzialmente un dato di sintesi ma è particolarmente interessante per lo studio della tendenza temporale.
Si individua facilmente la tendenza alla crescita nel corso del 2003 e, a parte l'interruzione di febbraio-aprile 2004 (anche il dato di marzo non è completo perché la rilevazione è ricominciata correttamente all'incirca a metà mese) dovuta a problemi tecnici nella rilevazione, sembra che il volume di visite mensili non accenni a diminuire. Ovviamente non bisogna considerare le code di questo grafico: trattandosi di valori assoluti, infatti, il primo e l'ultimo mese di rilevazione non possono essere equivalenti agli altri per il fatto che non sono stati considerati solo quelli completi.

3. Statistiche di accesso

È interessante analizzare le sezioni del sito più visitate.

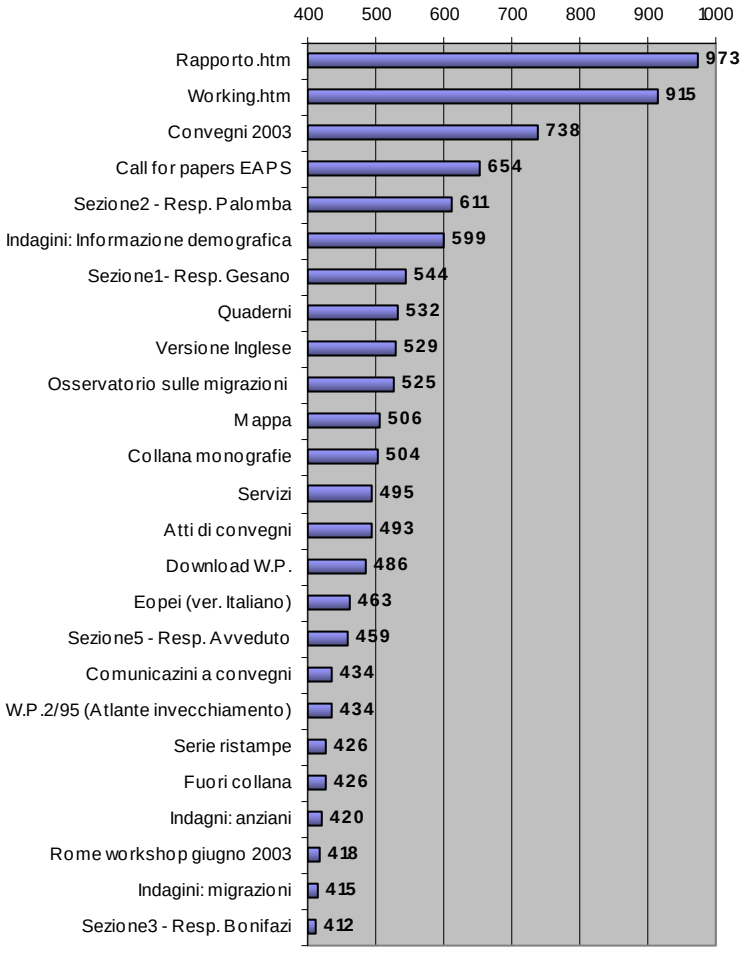
Grafico 5: Le aree più visitate del sito



Il grafico 5 mostra i valori assoluti delle visite alle sezioni del sito. La sezione maggiormente visitata è quella della ricerca con 2403 visite nell'anno esaminato, seguita dalle pubblicazioni, le informazioni sull'istituto, Demotrends, le pagine sulle indagini e quelle sulle opportunità di occupazione che raggiungono livelli superiori ai 2000

contatti nell'anno. In questo caso abbiamo escluso la pagina iniziale perché è ovviamente quella più visitata in assoluto, anche in considerazione del fatto che, come abbiamo già detto, buona parte degli interni all'Istituto hanno come pagina principale predefinita nel browser di navigazione su Internet proprio la pagina home del sito in esame.
Si vede quindi che i visitatori del sito mediamente cercano informazioni che riguardano i possibili prodotti dell'istituto, generalmente pubblicazioni e competenze nei settori della ricerca. Come c'era da attendersi anche le pagine di servizio, come quella delle opportunità lavorative, sono molto visitate a riprova del fatto che il sito è visitato sia da esperti del settore che lo usano come fonte di informazione per la loro attività, sia da persone in formazione o in cerca di occupazione in possesso di qualifiche a loro giudizio compatibili con la nostra attività di ricerca.

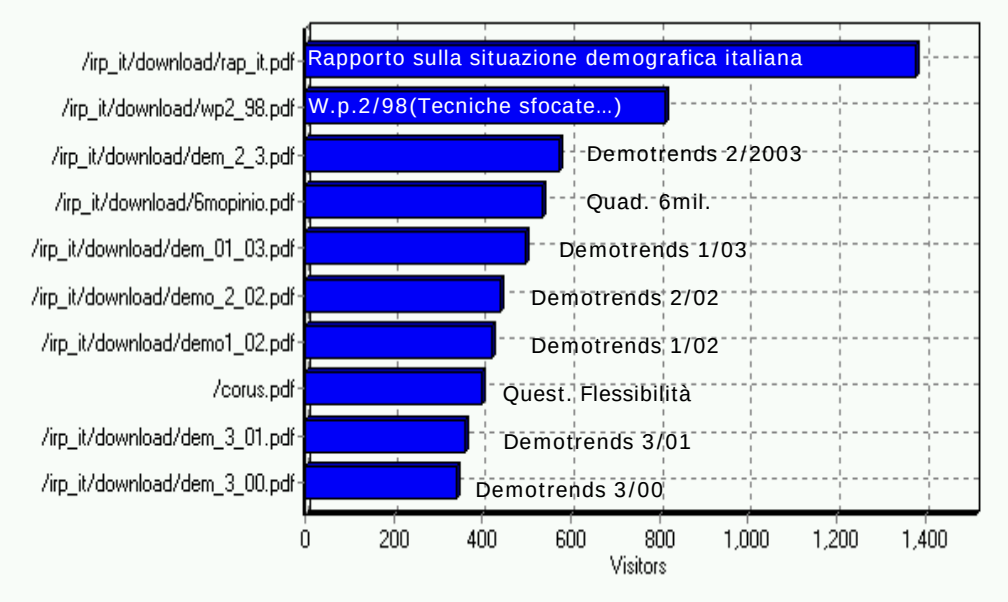
Grafico 6: Le pagine più popolari



Il grafico 6 mostra il dettaglio delle pagine maggiormente visitate. Si tratta di pagine scaricate per l'informazione in esse contenute utilizzabile sia a fini di ricerca (come ad esempio le pagine contenenti pubblicazioni on line e quelle con le informazioni sulle attività di ricerca dell'istituto), sia come servizio per la ricerca (ad esempio le pagine di informazioni su convegni e call for papers). È quanto mai evidente, quindi, che il sito, anche in questa prima versione ancora non completa di tutto il materiale disponibile, rappresenta uno strumento di comunicazione di informazioni apprezzato dagli utenti del web, nonostante la forte specializzazione degli argomenti trattati.

Oltre alle visite alle pagine del sito, siamo in grado di conteggiare i *download* (termine usato per indicare le operazione di 'scarico' di materiale da Internet per l'archiviazione su computer locali) dei files in pdf[4] messi a disposizione degli utenti. Molto spesso questi files sono vere e proprie pubblicazioni on line che, messe nelle pagine del sito, possono essere scaricate gratuitamente. In un futuro molto prossimo sarà possibile un uso differenziato di questo strumento e si potrà avere un'area download riservata solo ad utenti autorizzati all'accesso a materiale messo in rete. In ogni caso, è utile, già a questo stadio dell'analisi, capire che tipo di documenti sono particolarmente apprezzati in rete e, indirettamente, formulare utili ipotesi sul profilo degli utenti del sito.
Nel grafico 7 sono contenuti i valori assoluti delle visite a questi particolari files. Oltre alla presenza di diversi numeri di Demotrends, che riscuote comunque grandi consensi tra il pubblico on line, sono stati scaricati diversi prodotti pubblicati dall'istituto. Tra essi spicca il 'Breve rapporto sulla situazione demografica italiana' che è stato aggiornato nel 2001 e che evidentemente, per la sua formulazione agile e facilmente intelligibile ha ottenuto in assoluto il numero maggiore di visitatori. Sono poi presenti in questa graduatoria un working paper di impostazione tecnico-metodologica (W.P 2/98), un rapporto bilingue sulle indagini dell'Istituto (Quaderno sui 6 miliardi di abitanti) e un questionario on line che è stato visitato su sollecito dei ricercatori responsabili della ricerca.

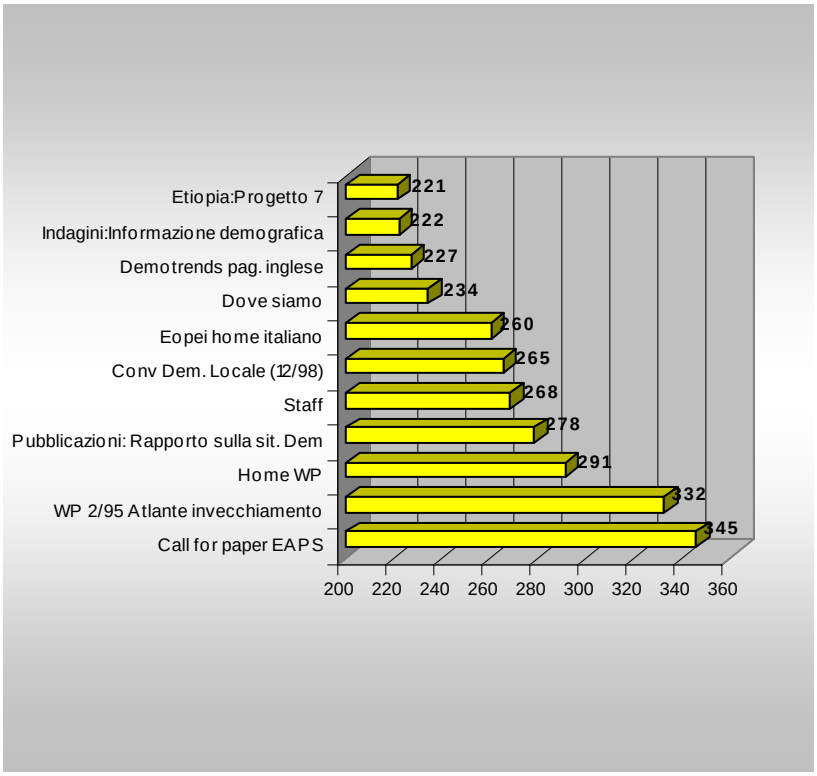
Grafico 7: Download: files pdf più scaricati



[4] Il formato Adobe PDF (Portable Document Format) è di fatto lo standard per la distribuzione e lo scambio di documenti su Internet. È un formato molto diffuso perché i documenti in questo formato possono essere aperti da tutti con l'ausilio di software gratuiti facilmente reperibili.

L'importanza di conoscere quale sia la pagina dalla quale sono entrati nel sito i visitatori sta nel fatto che si mettono in evi-denza le aree del sito che con buona probabilità sono state memorizzate tra le pagine preferite, perché solitamente è da queste che entra in particolari pagine diverse dalla home.

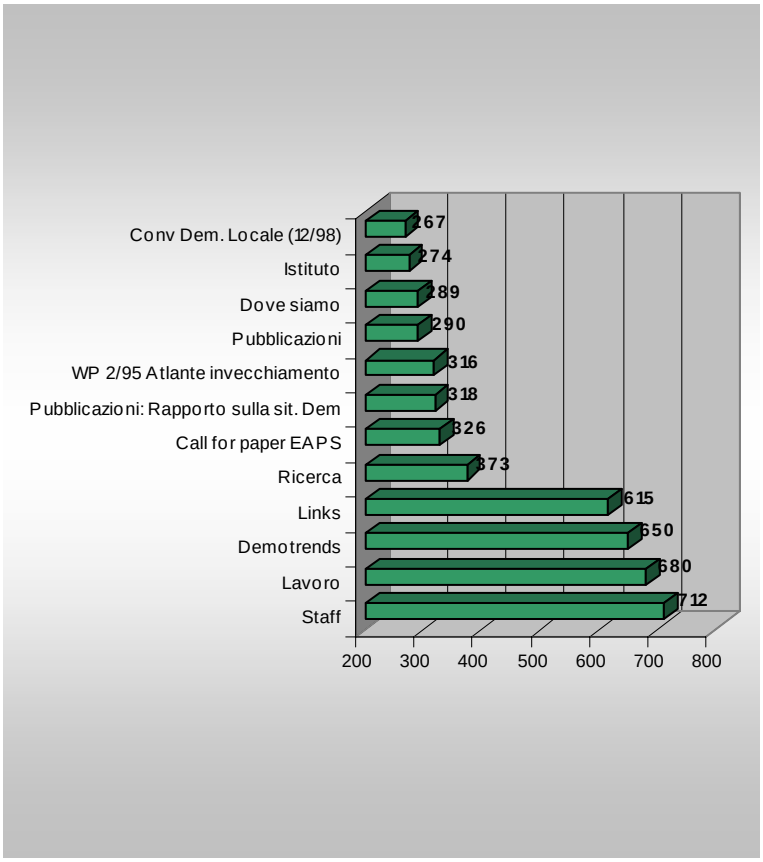
Grafico 8: Le pagine di entrata più frequenti



A volte però si perviene ad una pagina interna di un sito anche attraverso i motori di ricerca disponibili nel web. In entrambi i casi, si tratta di richieste di pagine mirate ad un determinato scopo, per cui è alta la probabilità che si tratti di visitatori realmente interessati all'argo-mento in oggetto. Chiaramente da questa graduatoria va esclusa la pagina iniziale che è stata visitata ben 16956 volte, ma, buona parte di esse, per caso. Invece i valori delle visite contenuti nel grafico 8 probabilmente non sono dovuti al caso, ma ad interessi specifici. Non è dunque una graduatoria basata sul numero di visite, ma si tratta dell'individuazione delle pagine a cui è riconosciuto maggior interesse, indipendentemente dal numero delle visite.

Analogamente, la conoscenza della zona del sito da cui si abbandona la navigazione consente di sapere quanti sono gli utenti che navigano realmente nel sito e non toccano solo per caso la pagina iniziale. Abbiamo rilevato nell'anno di osservazione 12889 uscite dalla pagina principale. Dal momento che non possiamo distinguere gli utenti che sono usciti dalla pagina principale senza visitare prima altre pagine, da quelli che, pur uscendo da questa pagina, hanno intrapreso un percorso di navigazione nel sito, abbiamo deciso di non considerarli affatto.

Grafico 9: Le pagine di uscita più frequenti

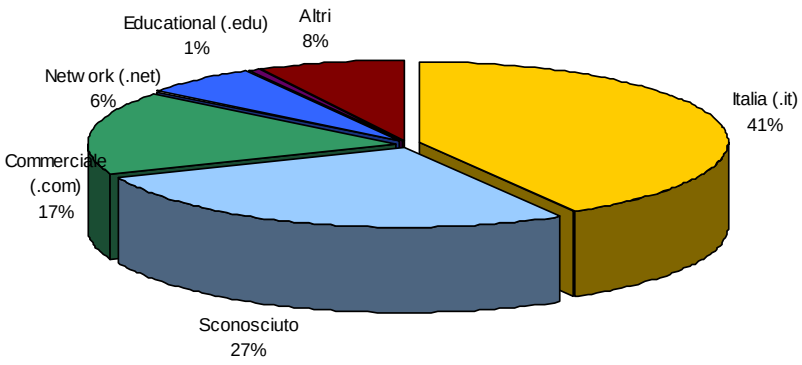


La lista delle pagine visitate come ultime si trova nel grafico 9. Non si può dire che si tratti di pagine indesiderate, ma soltanto di pagine terminali del percorso di navigazione. Esse raccolgono sia gli utenti che esauriscono la loro curiosità con la visita di quella particolare pagina, sia quelli che, giunti ad una certa pagina, non trovano più interessante continuare la navigazione. In ogni caso, il conteggio complessivo di utenti che sono usciti da una pagina diversa da quella iniziale, corrisponde al conteggio di coloro che hanno navigato realmente nel sito e che, di conseguenza, sono stati in qualche maniera interessati a farlo. All'incirca, considerando che abbiamo avuto 46298 visitatori e che di 12889 di essi non sappiamo se abbiano o no navigato nel sito, ne consegue che i rimanenti 33409 hanno in qualche misura navigato nel sito. Essi corrispondono al 72% dei visitatori.

4. Utenti visitatori

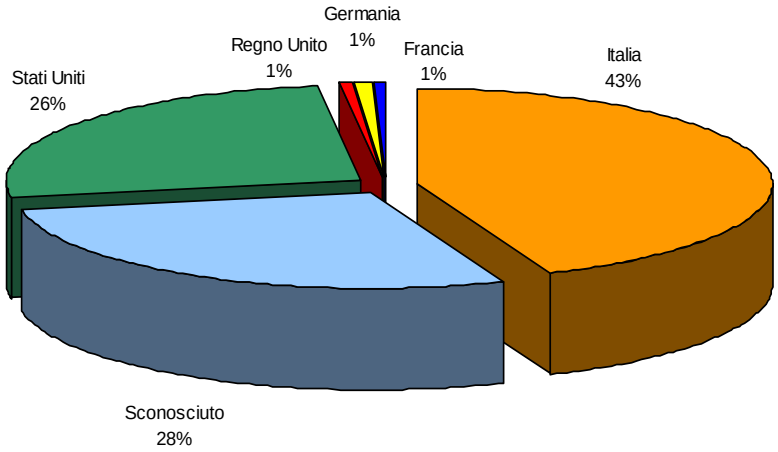
É interessante conteggiare quali siano i tipi di dominio, per quelli per i quali è possibile avere questa informazione, a cui afferiscono i nostri visitatori. Come era lecito attendersi, la maggior parte di essi (42%) proviene dall'Italia (domini .it) e una quota piuttosto consistente (circa il 27%) non è assegnabile ad alcun dominio noto. Il settore di tipo commerciale (domini .com) è presente nel 17% dei casi, mentre i Network (domini.net) compaiono nel 6% dei casi e gli Educational (domini .edu) sono conteggiati in meno dell'1% dei visitatori.

Grafico 10: I domini più frequenti



Tutti gli altri domini sono presenti nel 13% dei casi.
Analogamente si può pensare di conteggiare il paese di provenienza del dominio: nel 42% dei casi si tratta di domini italiani, seguiti da quelli sconosciuti (che sono il 27%) e quelli statunitensi (24%). Tutti gli altri paesi sono presenti in meno dell'1% dei casi, ma si tratta di un numero considerevole (oltre 100) di paesi del mondo.

Grafico 11: I paesi di provenienza dei visitatori



Ovviamente, in questo caso, non distinguiamo tra visitatori in carne ed ossa, e tutti quei visitatori di tipo elettronico che periodicamente visitano i siti per la costruzione di database utili per i motori di ricerca internazionali. Il dato è molto interessante ma non può non preoccupare il fatto che la versione inglese del sito non sia adeguatamente sviluppata, per cui i motori di ricerca internazionali potrebbero aver non considerato le nostre pagine in quanto non hanno trovato in esse i termini di uso comune nel linguaggio internazionale necessari per essere inseriti in un qualunque motore. Ciò nonostante, la numerosità dei visitatori internazionali è tale da far pensare che il sito è comunque presente a livello internazionale, molto probabilmente per il fatto che sono stati inseriti diversi 'meta tags', cioè parole chiave inserite nel codice html che sono ricercate dai programmi di ispezione dei siti da parte dei motori di ricerca.
Ovviamente questa lista rappresenta un elenco che soddisfa una pura curiosità dal momento che non si può determinare con assoluta certezza la provenienza del visitatore; si può invece determinare, anche se non sempre, solamente la localizzazione del server a cui l'utente è collegato. Rimane comunque valido il fatto che ci sono paesi del mondo con frequenze di visita molto elevate (e pertanto è alta la probabilità che gli utenti

appartengano a questi paesi che si trovano in cima alla lista della tabella 2), mentre sono meno attendibili soltanto le provenienze dai paesi con valori minori di tali frequenze.

Tabella 2: Paesi di provenienza dei visitatori più attivi

Paese	Visitatori	% di vistatori sul totale	Paese	Visitatori	% di vistatori sul totale
Italia	19,302	41.69%	Ungheria	30	0.06%
Sconosciuto	12,539	27.08%	Romania	30	0.06%
Stati Uniti	11,363	24.54%	Marocco	29	0.06%
Regno Unito	324	0.70%	Arabia Saudita	27	0.06%
Germania	307	0.66%	Grecia	27	0.06%
Francia	266	0.57%	Nuova Zelanda	25	0.05%
Svizzera	200	0.43%	Federazione Russa	23	0.05%
Australia	176	0.38%	Irlanda	23	0.05%
Olanda	173	0.37%	Sud Africa	23	0.05%
Canada	165	0.36%	Cecoslovacchia	23	0.05%
Belgio	127	0.27%	Turchia	22	0.05%
Spagna	90	0.19%	Croazia	22	0.05%
Austria	89	0.19%	Cile	21	0.05%
Polonia	65	0.14%	Filippine	16	0.03%
Brasile	64	0.14%	Lussemburgo	14	0.03%
Danimarca	59	0.13%	India	14	0.03%
Giappone	56	0.12%	Perù	13	0.03%
Norvegia	54	0.12%	Taiwan, Repubblica Cinese	13	0.03%
Finlandia	52	0.11%	Indonesia	12	0.03%
Svezia	51	0.11%	Malesia	10	0.02%
Messico	51	0.11%	Estonia	10	0.02%
Portogallo	47	0.10%	Slovenia	10	0.02%
Israele	35	0.08%	Lituania	9	0.02%
Argentina	31	0.07%	Colombia	9	0.02%
Singapore	31	0.07%	Emirati Arabi Uniti	8	0.02%

Oltre a ciò, abbiamo pensato di analizzare quali istituzioni hanno fatto uso del materiale messo a disposizione sul sito dell'IRPPS. Il risultato è sorprendente perché abbiamo avuto visite da moltissime università italiane ed enti di ricerca pubblica nazionali ed internazionali, e da diversi enti amministrativi locali. Il motivo di tante visite in molti casi è noto, dal momento la finestra di osservazione che abbiamo deciso di utilizzare comprendeva un ampio periodo in coincidenza del quale è stato messo on line un questionario di rilevazione sulle forme di lavoro a tempo negli enti

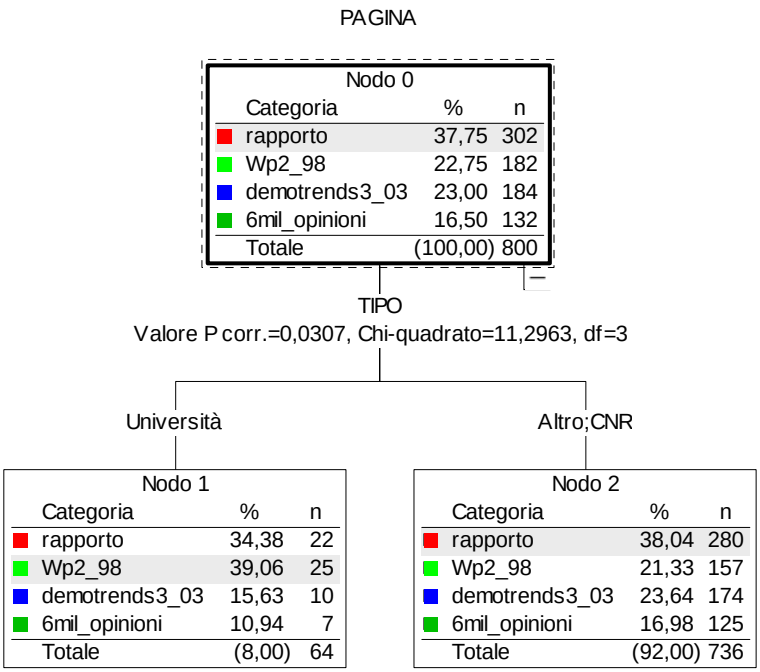
pubblici di ricerca italiani. Ma il volume delle visite in diversi casi è superiore alle aspettative e ciò lascia ben sperare sul fatto che *l'IRPPS on line* abbia potuto fornire i servizi che auspicava.

5. Un'analisi di approfondimento

Facciamo un ultimo cenno ad una piccola analisi di approfondimento di questi dati, il cui risultato può sembrare banale, ma per noi rappresenta la prova che la nostra analisi è congruente con le nostre ipotesi di partenza in quanto fornisce informazioni sulle preferenze dei visitatori e da considerare come un esempio delle ulteriori analisi che si possono eseguire per estrarre altra informazione dai files di log.
In sintesi, sono state selezionate soltanto le visite fatte nei mesi di aprile e maggio 2004 e sono stati individuati i 4 documenti pdf più scaricati in questo periodo. Si tratta di materiale di cui abbiamo parlato in precedenza e che riassume una gran parte delle caratteristiche generali delle pubblicazioni on line del sito in esame: il *Rapporto sulla situazione demografica italiana*, che rappresenta un esempio di materiale altamente divulgativo, tanto breve quanto efficace in termini comunicativi; il *Working Paper 2 1998*, che è un testo di tipo tecnico-statistico, una sorta di manuale per lo studio e l'applicazione di metodi di classificazione sfocata ai dati di popolazione; l'ultimo numero di *Demotrends* del 2003, la newsletter dell'istituto che pur essendo scritta in forma adatta ad un pubblico non propriamente specialista degli studi di popolazione, contiene materiale di un certo livello scientifico dal momento che raccoglie i contributi di diversi studiosi di richiamo internazionale; il volume sulle indagini svolte dal'IRPPS in occasione della data segnalata a livello internazionale come quella del raggiungimento di quota 6 miliardi di abitanti sul pianeta, un testo bilingue che può interessare, oltre a studiosi di popolazione, anche curiosi di altro tipo. Si tratta quindi di materiale dalle caratteristiche diverse, pur essendo comunque tutte opere di argomento generale comune.
Dalle tracce lasciate dai visitatori di queste pagine si sono ricavate in particolare le seguenti informazioni: se il visitatore proveniva da un dominio .it, .com oppure .net; se proveniva da un'università o da un istituto cnr. Si tratta di pochi scarni dettagli che però hanno già dato qualche risultato interessante. È stata creata una nuova matrice di dati contenente l'informazione di quale delle pagine pdf di nostro interesse era stata richiesta dal visitatore e il possesso da parte dello stesso visitatore delle 5 caratteristiche appena descritte.
Applicando a questi dati la tecnica dei 'decision tree', e validando in maniera casuale l'albero che ne scaturisce, si è ottenuto un risultato interessante: quando si circoscrive alle sole università l'analisi del tipo di documento scaricato viene fuori una differenza con il resto dei visitatori che fondamentalmente ri-specchia l'interesse che il materiale pub-blicato può suscitare oltre che l'uso che se ne può fare. La differenza più vistosa che si percepisce è tra materiale più divulgativo rispetto a più spiccatamente tecnico e quindi riservato ad un pubblico selezionato. Cosicché nel ristretto gruppo di visitatori appartenenti alle università la graduatoria tra le 4 opere

considerate subisce delle modifiche portando in testa il Working paper, seguito dal rapporto, poi da demotrends ed in ultimo dal volume bilingue.

Figura 1: Albero decisionale: periodo aprile-maggio 2004 per 4 tipi di pdf



Nell'altro gruppo, per la verità molto più numeroso perché raggiunge i 736 contatti, abbiamo in testa il rapporto seguito da demotrends e solo in terza posizione troviamo il working paper.
Vale a dire: se i visitatori non sono in qualche maniera riconducibili a studiosi di popolazione è più probabile che cerchino materiale di grossa divulgazione, e, come d'altro canto è ovvio, il materiale più specialistico viene richiesto dagli studiosi della materia. Questo non può che confortarci sul fatto che anche in questo caso la scienza può essere divulgata al grande pubblico fermo restando che i livelli di approfondimento necessari differiscono a seconda dei casi.

6. Conclusioni

La versione del sito dell'IRPPS che è stata on line durante la finestra di osservazione qui considerata aveva dei requisiti tecnici e strutturali tali da non risultare del tutto adeguata agli scopi del sito in oggetto. Ciononostante, il numero di utenti che hanno usufruito dei servizi offerti è notevole, considerando campo di interesse molto specifico, ed anche gli utenti internazionali sono presenti in misura considerevole. Questi dati allora sono serviti da stimolo a migliorare le pagine on line sia dal punto di vista dei servizi offerti, sia dal punto di vista della navigabilità e facilità di accesso agli stessi. È stata quindi predisposta una nuova versione del sito dalla struttura completamente diversa e molto più ricca di contenuti riguardo la ricerca effettuata dall'Istituto. L'analisi del materiale scaricato di preferenza dagli utenti visitatori suggerisce che i tempi sono maturi per usare il sito come un portale di scambio e di comunicazione nel settore della ricerca sulla popolazione e le politiche sociali: lo rileviamo dal numero

di contatti all'ultimo call for papers per un convegno internazionale, prassi comunicativa ormai consolidata in eventi di questo tipo nel mondo della ricerca, oppure dalla quantità di numeri di Demotrends (la nostra newsletter) che viene abitualmente scaricata. È indicativo anche il successo di una pubblicazione pdf molto sintetica e schematica, sullo stile dei compendi annuali che ad esempio si trovano sul sito dell'ISTAT o di un lavoro di argomento tecnico/statistico: probabilmente ci sono molti utenti che preferiscono materiale 'pronto all'uso', tipo manualistica che può essere di notevole aiuto nello studio e nella ricerca ed è ancora più gradito quando è facilmente scaricabile e gratuito.
La versione futura del sito avrà anche una parte di interattività utile per avere risposte alle nostre domande direttamente dagli interessati oltre che indirettamente dalle statistiche di accesso. Questi sistemi inoltre consentono il filtraggio dell'offerta di servizi: si potrà pensare, ad esempio, di fornire un certo servizio ad una particolare categoria di utenti abilitati a riceverlo escludendo tutti gli altri. La tecnologia attualmente a disposizione consente l'accesso facilitato a questo genere di attività on line per cui ci pare giunto il momento di metterle a disposizione della ricerca.

Bibliografia

Baldi P., Fiasconi P., Smyth P., (2003) *Modeling the Internet and the Web, Probabilistic Methods and Algoritms*, Wiley.

Giudici P., (2001) *Data mining – metodi statistici per le applicazioni aziendali*, McGraw-Hill, Milano.

Lynch P.J., Horton S., (2001) *Web – Guida di stile – Progettazione dei siti Web*, Apogeo, Milano.

Pazzani, M. J. (2000*), Knowledge discovery from data*?, IEEE Intelligent Systems March/April 2000, 10–13.

Kohavi R., Provost F. (2001) *Applications of Data Mining to Electronic Commerce*, Data Mining and Knowledge Discovery, 5, 5–10, Kluwer Academic Publishers, Manufactured in The Netherlands

Rosenfeld L., Morville P., (1998) *Information architecture for the World Wide Web*, Sebastopol, O'Reilly.

Silani S., Tarantino M., (2001) *Tecniche di Data Mining con SAS Enterprise Miner*, SAS Institute.

Riassunto

La gestione del sito web dell'istituto ha richiesto nel tempo competenze sempre più affinate e mirate ad ottimizzare il servizio offerto con le pagine pubblicate sia per gli utenti interni che per gli esterni. In tale ottica si pone la necessità di sviluppare ed applicare competenze di data mining ai dati disponibili in rete: ciò consente, oltre alla rilevazione delle aree del sito di maggior interesse, anche l'ottimizzazione del sito stesso in termini di fruibilità e facilità di accesso, tenendo ovviamente conto degli scopi e delle peculiarità di un sito afferente ad un istituto di ricerca e dalle caratteristiche assolutamente difformi da un sito di natura commerciale o di altra natura.
Il file di log disponibile sul server è stato analizzato attraverso le tecniche di data web mining attualmente disponibili che hanno consentito di estrarre preziose informazioni sui comportamenti di visita alle pagine pubblicate anche se la particolare struttura del sito ne consente al momento un uso piuttosto limitato. Ciò nonostante è stato possibile reperire informazioni essenziali per l'ottimizzazione del sito stesso e per l'implementazione di nuove funzionalità e servizi per gli utenti visitatori.

Summary

The web site design and management of our institute has required more and more specialized competencies over the time. Our aim is to optimise the service offered by the web pages for both internal and external users. Our approach is to develop and apply data mining methods to web data, this enables the information and data retrieval connected to different topics of the web site according to the user interest. The optimisation is also related to usability and accessibility criteria. Of course it is necessary to take into account aims and characteristics of a research institute web site, which are completely different from a commercial one. The log files available on the server have been analysed using of the available data web mining technique. In this way we have obtained precious information on the behaviours of web visitors, even if the structure of the web site allowed us a limited data use at that moment. Nevertheless, it was possible to find important information for the optimisation of the web site and for carrying out new functionalities and services for the visitors.