Under Attack!

num 13

Cover by Likas

UNDERATTHACK

IN QUESTO NUMERO

Programming

Storie, Etica & Cultura hacker

N.13

Prefazione

Sono sempre i bambini che credono alle favole, però sono solo gli adulti che le scrivono.
La cosa più divertente delle fiabe è che tutto funziona alla perfezione, un bianco e nero completo dove i buoni sono buonissimi, i cattivi cattivissimi, il bene vince sempre e il male sempre soccombe; non si capisce se lo scopo è quello di illudere i giovani oppure dare a noi stessi l'idea che esista qualcosa di diverso dalla cruda realtà.
Per quanto sembri strano anche noi adulti crediamo alle favole che inventiamo noi stessi: della politica, della pubblicità, sulla carriera professionale, sul futuro dei nostri figli e via di seguito; evidentemente il concetto di maturità è molto meno reale di quanto sembri. Mi viene da pensare che esista una qualche analogia tra fiaba e concetto di 'modello' come modelli economici, modelli matematici o i nuovi modelli morali passati di recente alla ribalta; come i modelli anche le fiabe di per sé sono giuste, se analizzate sono funzionanti, se messe nella pratica sono ridicole.
Esiste un pensiero dominante che altro non è che la media aritmetica dei singoli pensieri sommati, non si pensa mai alla verità ragionata ma si pone come vero ciò che viene detto dalla maggioranza delle persone, in altre parole il Leviatano è l'uomo medio che esiste come modello ma non è identificabile con nessuno di noi.
Le implicazioni più divertenti di questo fatto sono sostanzialmente due: la prima è l'impossibilità non solo di realizzare ma nemmeno di concepire qualcosa di diverso, io stesso potrei aver appena esposto un mio modello per spiegare che i modelli non funzionano, oltre al fatto che se non comanda l'uomo medio allora comanda un dittatore o un'oligarchia. La seconda implicazione è più simpatica e più inquietante allo stesso tempo; se una cosa è vera perchè 'la dicono tutti' siamo sicuri di chi sia 'tutti'? 'Mario è un delinquente, lo dicono tutti!', quel 'tutti' più o meno corrisponde a mio cugino più il tabaccaio più il barista, se lo dicessero il sindaco oppure il parroco sarebbe già cosa riconosciuta. A questo punto mi chiedo cosa accade quando 'tutti' corrisponde al maggior quotidiano del paese, così come al telegiornale più seguito o al presentatore più in voga al momento.
Allora io voglio uscire da questo pazzo meccanismo, dirò quindi che il mio pensiero precedente è solo mio, 'tutti' potranno accettarlo o criticarlo in una parte più o meno estesa, 'nessuno' penserà esattamente ciò che io penso.

.

Buona lettura
Floatman

Under_NEWS_AttHack

Pwn2Own 2011: chi trova buchi nei browser vince!

La quinta gara annuale per hacker organizzata a Vancouver dal 9 all'11 marzo dal team Zero Day Initiative dei DVLabs TippingPoint, una divisione di HP che si occupa di sicurezza, è terminata con un sostanziale pareggio fra i contendenti: da un lato i produttori di browser web per sistemi operativi e dispositivi mobili e dall'altro hacker di lunga data ed esperti di sicurezza.
Nel corso della competizione i partecipanti hanno avuto 30 minuti di tempo per scovare e sfruttare falle e "buchi" software nei principali browser web installati e preparati per l'occasione con le ultime versioni disponibili su piattaforme Windows 7, Mac OS X e alcuni smartphone fra cui un iPhone 4, un Dell Venue Pro con Windows Phone 7, un Blackberry Torch 9800 con Blackberry 6 OS e un Nexus con Android. Il premio in palio per chi è riuscito nell'impresa? Un sacco di soldi: 105.000 dollari di montepremi in totale e ben 15.000 dollari e il computer/dispositivo bersaglio usato nel corso della gara per singolo vincitore. Quest'anno Google aveva preteso di essere inserito fra le "cavie" del concorso e aveva scommesso sull'impermeabilità del suo Chrome aggiungendo alla posta altri 20.000 dollari.

Nel corso della gara sono cadute le difese di Safari su Mac OS X, Internet Explorer 8 su Windows 7, iPhone 4 e Blackberry Torch 9800, mentre Firefox, Chrome, Android e Windows Phone 7 sono praticamente rimasti inviolati. Alcuni dei colossi informatici hanno già rilasciato patch e aggiornamenti per i browser che sono stati bucati o anche solo scalfiti dai colpi degli hacker che, per la verità, in qualche caso sono riusciti nell'intento solo grazie a vulnerabilità di plug-in e software accessorio come Acrobat Reader e i vari Flash player. La competizione è stata anche oggetto di critiche da parte di vari esperti del settore, secondo i quali questo tipo di gare non ha alcun senso dal momento che i partecipanti dispongono di mesi di tempo per prepararsi e sfruttare debolezze trovate in precedenza ma mai divulgate fino al momento di concorrere per i premi.
Maggiori dettagli sul blog ufficiale:
http://dvlabs.tippingpoint.com/blog/2011/02/02/pwn2own-2011

cercamon

Piove, governo ladro!

Una pioggia di bit sul sito del Governo.

Lo scorso 6 febbraio alle ore 15:00 Anonymous Italy, un gruppo di attivisti hacker non identificati sparsi per il mondo, ha sferrato un attacco di tipo **Ddos** (Distributed Denial Of Service) con il nome in codice "***Operazione Italia***" contro il sito web del nostro governo. Come noto, un'azione Ddos consiste in un flusso massiccio e coordinato di richieste d'accesso da parte di un enorme numero di workstation su tutta la Rete dirette verso un determinato sito alfine di metterlo fuori uso proprio a causa dell'impossibilità del web server ospitante di fare fronte a tutte le richieste pervenute. Durante l'attacco il livello di traffico diretto contro il server sale fino a saturare tutte le risorse disponibili. Le conseguenze dirette sono state dapprima una lentezza generalizzata della connessione al sito e successivamente il collasso totale del server. Dopo pochi minuti la home page del Governo risultava completamente irragiungibile.
La Polizia Postale, competente in casi del genere, ha spiegato che non è stato possibile contrastare l'azione informatica, anche se questa era stata ampiamente preannunciata, soprattutto perché le richieste di accesso provengono da computer sparsi non solo in Italia ma anche all'estero. Si è potuto soltanto cercare di individuare gli indirizzi IP da cui sono giunte le richieste più numerose e si è tentato di bloccarli mediante servizi di tipo firewall. Ma naturalmente non è stato possibile bloccare l'enorme numero di computer che sono stati azionati per l'attacco. La sigla "Anonymous" è stata di recente utilizzata in molte operazioni di questo tipo dirette contro vari obbiettivi, come ad esempio è accaduto per la cosiddetta "Operation Payback" messa in atto in difesa del sito Wikileaks e diretta contro i siti PayPal, Mastercard e Amazon, colpevoli di aver bloccato le risorse finanziarie dell'organizzazione facente capo a Julian Assange.
A giudicare dal comunicato diffuso degli hacktivist di Operazione Italia il governo tricolore è stato inserito fra i bersagli del gruppo perché reo "di censurare il web, di rendere la giustizia uno strumento iniquo, di favorire la prostituzione (anche minorile), di praticare oscuri rapporti con la mafia e di corrompere e manipolare l'informazione per fini personali". Come dissentire?

cercamon

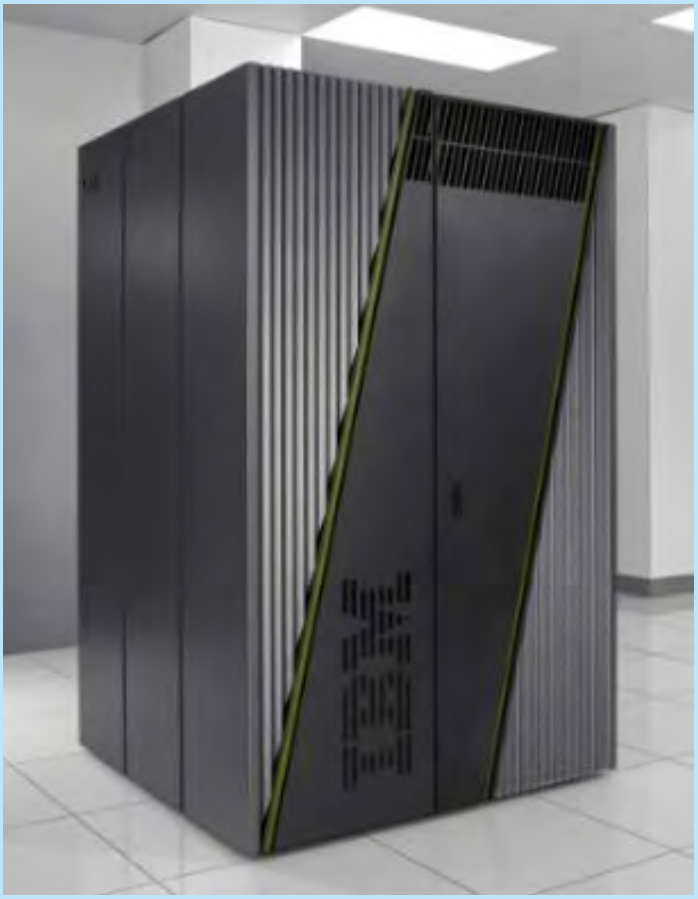
Quando la banda scarseggia,
il file sharing non galleggia!

Dallo scorso primo marzo anche Telecom Italia, primo operatore nazionale di servizi Internet basati su ADSL, rallenterà le connessioni verso la Rete in caso di congestione delle proprie risorse di banda internazionale. A rimetterci saranno soprattutto i servizi di file sharing e peer-to-peer, fra i più ingombranti in termini di occupazione di banda ma da sempre molto gettonati fra gli utenti di servizi di connessione ADSL. Altri operatori come Wind e TeleTu hanno già adottato la misura in questione, ma nel caso di Telecom si tratta del proprietario della rete nazionale ed è conseguenza diretta, secondo molti esperti del settore, della cronica mancanza di infrastrutture e di investimenti adeguati su tutto il territorio nazionale.

La carenza di infrastrutture moderne per rispondere alle esigenze del mercato italiano e per colmare il divario esistente nei confronti degli altri paesi europei quanto a disponibilità di banda è notoriamente cronica. In Italia il 20% delle centrali, che riguarda meno del 10% della popolazione, non è in grado di far fronte alle richieste di traffico ed accade che quando la banda è sovraccarica ulteriori utenti non riescano a connettersi. Il problema quindi è che solo Telecom, in quanto proprietario della rete nazionale, oggi può decidere dove e come chiudere il traffico Internet basato su ADSL. La soluzione sarebbe la fibra ottica ma Telecom dichiara che la sua installazione estesa costerebbe milioni di euro e risulterebbe anti-economico investire e l'importo degli abbonamenti diventerebbe insostenibile. In mancanza di infrastrutture sufficienti l'unica possibilità, secondo Telecom, è limitare l'accesso in modo da consentirlo a tutti. Non dovrebbe però essere il gestore nazionale a decidere chi rallentare, ma, per garantire la cosiddetta net neutrality (l'accesso alla Rete con parità di trattamento fra gli utenti), dovrebbe farlo un'autorità pubblica come un garante. Meglio ancora, l'ideale sarebbe che la rete nazionale fosse di proprietà pubblica e che gli investimenti per potenziarne le risorse venissero non solo dallo Stato ma anche dai principali content provider, dai grandi social network e dai colossi del web.

cercamon

Vi bastano 10 petaflops
per giocare a Quake?

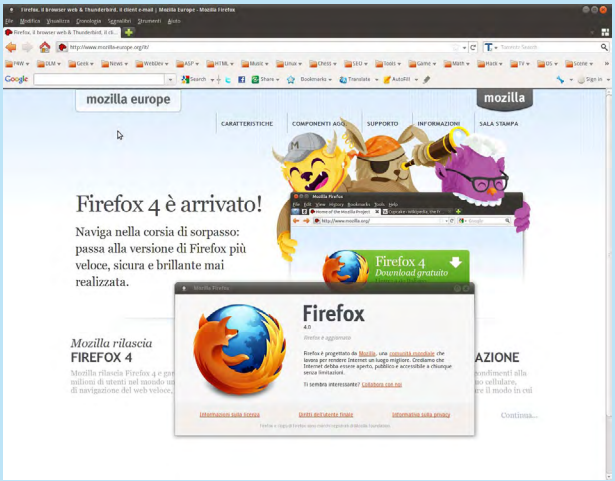


Se la risposta è sì, allora potete recarvi nel più vicino rivenditore di supercomputer IBM e ordinare il vostro prossimo PC: Mira. Questo è il nome del nuovo mostro macina-numeri progettato da Big Blue e basato su architettura Blue Gene/Q, capace di performance dell'ordine di 10 petaflops, cioè all'incirca 10^{15} operazioni (cicli CPU) al secondo!
Questa meraviglia di silicio sarà composto da 750.000 processori PowerPC A2 a 16 core e frequenza di clock da 1.6 Ghz, il tutto installato in rack da 1024 nodi ciascuno. Ciascun nodo sarà in grado di gestire fino a 16 GB di RAM per un totale di circa 750 TeraByte!
Musica per le orecchie degli appassionati di hardware, ma anche i sostenitori del free software potranno gioire: il sistema operativo installato sui nodi di calcolo è un Lightweight Kernel Operating System basato su GNU/Linux, mentre i nodi di I/O e l'interfaccia di gestione del supercomputer funzioneranno grazie a una versione modificata di Red Hat Enterprise Linux.
Per saperne di più: http://top500.org TOP500, la classifica dei supercomputer e l'articolo di PCWorld http://www.pcworld.com/article/218951/us_commissions_beefy_ibm_supercomputer.html

cercamon

Firefox 4 è tra noi!

A poche ore dal rilascio ufficiale della versione stabile di Firefox 4 mi sono precipitato sul sito di Mozilla per scaricarlo e aggiornarlo sul mio pc... la prima cosa che ho detto appena l'ho avviato è stata: "Ma ho avviato chrome?" ebbene si, il look del nuovo nato di casa Mozilla è molto simile al concorrente di casa Google.
Devo però dire che la scelta adottata dal team è stata ottima perchè ad una prima occhiata adesso saltano subito all'occhio (scusate il gioco di parole xD) le cose importanti, ovvero le tab che sono aperte e non più tutti gli addon come succedeva fino alla versione precedente.
Per quanto riguarda la velocità Mozilla ha dichiarato che Firefox 4 è ben 6 volte più veloce rispetto al precedente grazie alla riscrittura del motore javascript che garantisce un notevole incremento delle prestazioni...un'altra nota favorevole è il supporto ai nuovi tag HTML5 e CSS3 anche se ancora non sono uno standard affermato.
Ricordate che ogni volta che aggiornavate o installavate un plugin dovevate riavviare Firefox rischiando che, in rari casi, non venivano riaperte le tab? Beh, il team ha capito che questa era una scocciatura e ha deciso di riscrivere anche il gestore di plugin... adesso non sarà più necessario riavviare il browser per rendere la modifica effettiva... un punto in più a favore per Firefox e una "grana" in meno per l'utente!



Per finire vi voglio parlare di un'altra chicca grafica (che personalmente non trovo particolarmente utile se avete pochi finestre aperte) che è "Gruppi di schede" che vi permette di vedere un'anteprima di tutte le tab aperte e di effettuare delle ricerche tra di esse.
Penso di avervi dato un'idea generale su come sia firefox 4, il mio consiglio è comunque quello di provarlo e di giudicare voi stessi se ne vale la pena o no sceglierlo come browser predefinito!

Ecco a voi il link per scaricare il browser:
http://www.mozilla.com/it/firefox/

Christian "ultimoprofeta"Giupponi

Post - Office

E ancora una volta benvenuti nella rubrica Post-Office, due mesi pesanti sono passati, mesi in cui sono piovute email che ci hanno richiesto: articoli, chiarimenti, l'esilio... Ed ecco pubblicate quelle che abbiamo selezionato per quest'uscita. Vi ricordo ancora una volta che per collaborare con la nostra ezine basta inviarci una email a: **underatthack@gmail.com**

Da: garcetto
<garcetto@yahoo.it>

vi ho scoperti da poco e volevo complimentarmi con voi per la bella iniziativa!
sono un sistemista ormai di lungo corso e devo dire che siete concreti e di sostanza in un panorama di markettari e script-kiddies, quindi rinnovo i miei sinceri complimenti, ogni tanto noi italiani dobbiamo darci delle belle pacche sulle spalle quando ci vuole e ve lo meritate!
ciao e buon lavoro
Marcello

Le pacche sulle spalle, a mio avviso, sono una della miglior forma di complimenti, e anche se virtuale la tua la abbiamo apprezzata ancora di più...dato che sei un esperto affermato nel settore e sentirci fare i complimenti per la concretezza da una eprsona come te è il massimo. Grazie a te e buona lettura!
vikkio88

Da: Valentina Stimpfl
<valentina.stimpfl@gmail.com>

Ciao!
ho perso nome utente e password di typo3 e non riesco piu ad accedere al backend, ho provato ma l'url non sembra vulnerabile.
L'order by nemmeno non funziona ... come posso fare ?
Grazie mille

Beh ommioddio, non credo che per risolvere un problema di questo tipo debba rivolgerti a noi, se hai perso nome utente e password di sicuro ci sarà un modo per recuperarli e non sarà necessario tentare sqlinjection (che fondamentalmente a sentirti specificare credo che tu debba ripassare-studiare meglio)...Se continui ad avere problemi e non trovi come resettare la password, ti consiglio di contattare l'assistenza ufficiale del team di sviluppo di questo CMS...

vikkio88

Da: emanuele barbato
<emanuele92eam@live.it>

Ciao a tutti:) e complimenti per il vostro lavoro!! Mi farebbe piacere avere un parere da voi esperti!) Sono un ragazzo di 18 anni, sto frequentando l'ultimo anno di liceo e mi iscriverò ad ingegneria informatica!! Ho preso da poco la patente europea del computer ma nn mi è servita a un c...o. Erano tutte cose che già sapevo! Ho cercato di approfondire ancora la mia conoscenza del pc, frequentando forum e siti come il vostro ma con scarsi risultati... A volte leggendo quello che scrivete mi sento davvero un ignorante in materia! Non so da dove cominciare per imparare qualcosa sul serio! Lo so che la mia non è una domanda molto intelligente, ma come devo fare??? Cosa mi consigliate? Grazie in anticipo per la risposta che spero mi verrà data!

Ciao Emanuele e grazie dei complimenti.
Esperti? Beh nessuno mi aveva chiamato esperto ed è proprio una bella sensazione anche se fondamentalmente non lo sono. Tornando al discorso, l'ECDL secondo il mio modesto punto di vista, è un pacco inutile dato che nessun corso di laurea ti permette di convalidarti la materia informatica solo con questa, (tanto meno ingegneria informatica)... Conoscere l'informatica trascende completamente da usare il pacchetto office e da saper fare "cartelle" su windows...
Ti senti ignorante e certe volte non capisci quello che c'è su UnderAttHack, beh certo nessuno nasce con tutte le conoscenze del mondo, ma ti posso assicurare che la curiosità aiuta tantissimo. E in questi tempi di in cui tutti (o quasi) abbiamo a disposizione una connessione ad internet, è molto facile raccogliere informazioni utili e riuscire a sodisfare la nostra curiosità.
Che ti consiglio per cominciare ad addentrarti nel mondo geek? Beh, posso dirti cosa ho fatto io:
Programmazione stupida (pascal VB6), tanto smanettamento con windows (dal 95 all'XP), installazione di vari sistemi operativi, formattazione, assemblare pc, linux, c, c++, java, html, css, php, ruby, poi piano piano quando la mente ti diventa più elastica vai capendo da solo la tua strada...non c'è una via univoca per cominciare, ne mi sento di consigliarti proprio lo stesso percorso mio (windows 95 non c'è più)... Ma posso assicurarti che frequentare i forum con spirito critico, e leggere tantissimo aiuta di certo a sapere come si può sempre progredire!
Spero di esserti stato d'aiuto ugualmente, buona lettura!

vikkio88

Basi di Dati & Web Mining
e il loro utilizzo

Introduzione

Con questo paper voglio dare qualche informazione riguardo il Data Mining e il Web Mining, siccome è un argomento tanto snobbato quanto interessante.
L'approccio all'argomento non scenderà nei dettagli degli algoritmi e delle tecniche più avanzate, ma esaminerà il Data Mining nelle sue parti scomponendole e analizzandole cercando di capire come questo processo è interpretato a livello pratico.
Possono esserci imprecisioni, sacrificate al fatto che con questo paper non si vuole dare una lezione, ma solo introdurre e capire l'argomento nel suo insieme.
Detto questo possiamo iniziare.

Il Data Mining

In generale:

> *"Il data mining ha per oggetto l'estrazione di un sapere o di una conoscenza a partire da grandi quantità di dati (attraverso metodi automatici o semi-automatici) e l'utilizzazione industriale o operativa di questo sapere."[1]*

Per la maggior parte di coloro che lo usano il data mining viene applicato principalmente nel marketing dai dati (ad esempio) sui clienti di un'azienda, sul gradimento et cetera.
Ma in ambito web (**Web Mining**) diventa l'estrapolazione di dati da sorgenti web, che possono essere:

- **Pagine online**, ossia la ricerca di pattern o file sulla rete, utilizzando motori di ricerca personalizzati (Web Usage Mining)
- **Sorgenti HTML o XML** dalle quali estrarre link o una porzione di testo corrispondenti al nostro criterio di ricerca (Web Structure Mining).

Un'altra forma particolare di data mining è il Text Mining, che sfrutta la ricerca semantica, con linguaggio naturale, per estrarre testo relativo ad una particolare area di interesse da una grossa quantità di dati.

Il KDD

Un tipo di data mining molto diffuso è il Knowledge Discovery in Databases (KDD).
Il KDD estrae dati per noi interessanti da un database, secondo alcuni criteri ben definiti relativi ovviamente al nostro campo d'interesse.

Partendo dalla sua definizione, diciamo che, fondamentalmente:

> *"il KDD è il processo per identificare nei dati pattern (forme) con caratteristiche di validità, novità, utilità potenziale e facilità di comprensione."[2]*

Andiamo ad analizzare nel dettaglio la definizione:

- i dati sono i record, i campi contenuti nel database che hanno al loro interno le informazioni su cui dobbiamo basare la ricerca
- i pattern sono l'espressione della regola (o regole) che è il criterio della nostra ricerca, quell'elemento che distingue il dato interessante da quello da scartare.
- il processo consiste in più fasi iterative (spiegate meglio successivamente) in cui si ricercano i pattern tra i dati
- la validità consiste nello stabilire il livello di certezza del dato estratto nella ricerca secondo il pattern, ossia quanto e se il dato trovato rispetta il nostro criterio di ricerca
- la novità, in quanto i dati possono essere aggiornati e quindi aggiunti o eliminati dal database
- certamente l'utilità potenziale del dato estratto, in quanto bisogna considerare se e quanto questo è per noi utile per il nostro scopo
- non ultima la facilità di comprensione, requisito fondamentale in quanto il risultato del processo deve essere il più possibile comprensibile relativamente all'uso che se ne fa.

Il processo di mining

Il processo di data mining generalmente si suddivide a grandi linee in:

1] - Information Retrieval (IR) e Information Extraction (IE)

Siccome i dati "grezzi" raccolti sono spesso male o non affatto strutturati c'è bisogno di identificare ed estrarre le porzioni a noi utili. A questo scopo la fase di IR e IE sono incaricate di riconoscere le aree di interesse, raccogliere i dati e renderli quanto più possibile omogenei.
Gli strumenti dedicati a questi compiti sono i "wrapper" il cui lavoro è suddiviso in:

- Analisi lessicale e sintattica, riconoscimento di nomi e altre strutture lessicali e sintattiche
- Individuazione delle aree di interesse secondo i criteri scelti, ossia il Pattern Matching (riconoscimento del pattern dato tra i documenti a disposizione)

2] - Pre-Processing

Consiste , appunto, nel pre-processare i dati in modo da renderli accettabili e facilmente comprensibili durante il processo di mining, sfrondandoli e ottimizzandoli.

Le fasi sono quindi:

- **Pulizia** dai dati doppi, superflui o errati
- **Ottimizzazione** , ossia arricchimento e organizzazione dei dati nel modo più adeguato
- **Codifica**, cioè l'organizzazione delle informazioni in records specifici, in sostanza dare il nome ad ogni area di interesse

3] **Generazione dei template** (o modelli) contenenti i dati: ossia modelli che si adattano al tipo di ricerca e che contengono le porzioni di dati estratte. Questi template sono importantissimi, in quanto sono i "contenitori" dei dati, che li ordinano e che sono usati durante tutto il processo.

In abito pratico i template possono essere anche costruiti da zero a nostro piacimento, però esistono (a seconda del software e del metodo) template predefiniti che vanno solo selezionati nel tipo.
Esiste ad esempio la Data Mining Template Library (DMTL, http://dmtl.sourceforge.net/) che offre ottimi strumenti sia per la creazione che per la scelta dei templates.
La corretta e adeguata costruzione del template dipende infatti strettamente dal tipo di dato con cui abbiamo a che fare ma non meno dall'obiettivo del mining.
I template in ogni caso vanno definiti attraverso o **regole associative** predefinite o attraverso altri algoritmi stand-alone, come reti neurali, alberi decisionali et similia.

4] - **Mining**,
A questo punto i dati sono pronti per essere sottoposti al mining vero e proprio, cioè l'estrazione di conoscenza dal dato.
Il processo segue diverse tappe, vediamo di analizzarle singolarmente:

- **analisi delle sequenze:** si fa una analisi delle interconnessioni tra i dati pre-analizzati. Le sequenze sono i modi in cui il mining interpreta inizialmente i dati che riceve; questa fase dunque analizza le interconnessioni tra le sequenze e le correla tra loro, "preparandole" al clustering

- **clustering:** si osservano eventi che avvengono in gruppo; un cluster (o gruppo) è appunto un raggruppamento degli eventi che hanno elevata "vicinanza", somiglianza. Dati molto simili tra loro apparterranno allo stesso cluster, e dati dissimili a cluster diversi.

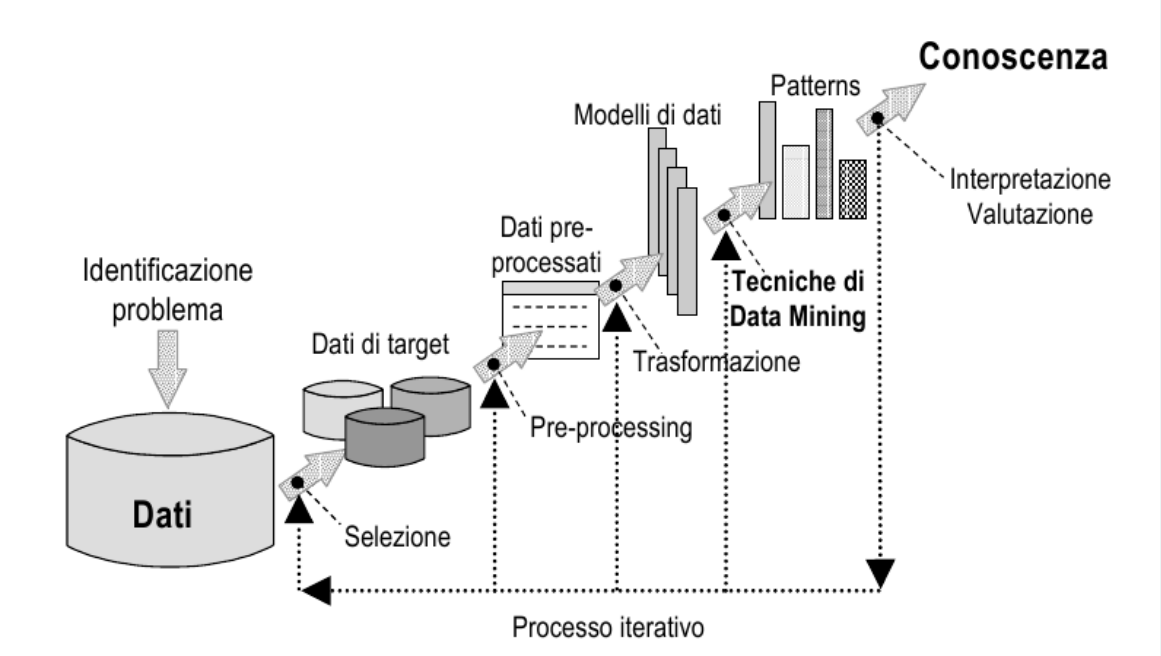
- **associazione:** si ricercano associazioni tra dati in base ad alcuni criteri. Durante questa fase si impostano le regole associative, che sono i criteri di collegamento tra gli attributi di un insieme di dati, ossia coloro che aggregano i cluster secondo la loro somiglianza.

- **classificazione:** lo scopo della classificazione è trovare un profilo descrittivo per i dati (sotto forma di template) e annoverarlo in una classe con un determinato attributo.

- **previsione:** in base ai dati in possesso si cerca di fare una previsione sui possibili eventi futuri. La previsione si basa sui template dati, e si appoggia su analisi statistiche o comunque probabilistiche (sempre appoggiandosi su algoritmi definiti).

5]- Post-processing e reporting, che riguarda una parte fondamentale, ossia la corretta ed esauriente esposizione dei risultati del mining.
Si possono creare infatti a seconda del tipo di dati e di risultato ad esempio istogrammi, tabelle, diagrammi a torta o qualunque altra cosaNel reporting rientra anche la creazione di file che sintetizzino il risultato del mining per poter essere poi salvati e catalogati, come per esempio file XML, CSV o anche una semplice tabella o diagramma.

Uno schema esemplificativo/riassuntivo è il seguente:



Algoritmi

Per la classificazione abbiamo visto che si usano appositi algoritmi, come ad esempio il k-Means [3] , il K Nearest Neighbor (kNN)[4] , la classificazione Naive-Bayes[5] .
Entriamo un po' più nel dettaglio di quest'ultimo:
la classificazione Naive-Bayes si può esprimere a partire dal Teorema di Bayes:

```
P(E | F1,F2,…,Fn) = P(E) * P(F1,F2,…,Fn | E) / P(F1,F2,…,Fn)
```

Detto in termini di classificazione, si supponga che l'evento E sia la categoria da assegnare a un certo documento, e che l'evidenza siano i singoli termini che costituiscono il vocabolario, la regola potrebbe esprimersi come la probabilità che un documento appartenga alla categoria E, data la presenza/assenza delle parole F1,…,Fn è pari al prodotto della probabilità stessa che un documento appartenga a una certa categoria moltiplicato per la probabilità di ciascuna delle parole, presupposto che si tratti della categoria E, diviso la probabilità delle parole stesse.

Anche espresso in questo modo, il concetto non rimane chiarissimo, ma la comprensione di questa legge non è lo scopo di questo articolo. In ogni caso, la legge scritta in questo modo è altamente esauriente ma sarebbe computazionalmente difficile da calcolare e rappresenterebbe il classificatore "ideale" di Bayes.
La classificazione infatti avverrebbe calcolando, dato un documento, la probabilità che questo appartenga a ciascuna delle possibili categorie, scegliendo, ovviamente, quella più alta. Come fare dunque a rendere la legge utilizzabile? Proprio utilizzando il principio di "ingenuità". Infatti, considerato che la probabilità congiunta espressa al numeratore del secondo membro dell'equazione è esprimibile come:

```
P(E, F1, F2, … Fn) = P(E)*P(F1|E)*P(F2|F2,E)*…*P(Fn|Fn-1,…F1)
```

La condizione di indipendenza o ingenuità ci permette di dire che per ogni i diverso da j è valida la seguente relazione:

```
P(Fi | C, Fj) = P(Fi | C)
```

Questo ci permette di ridurre l'equazione precedente (dove l'evidenza sarà espressa come F) nel seguente modo:

```
P(E | F) = P(C)*P(F1|C)*P(F2|C)*…*P(Fn|C)/P(F)
```

Adesso, posto che nella classificazione ingenua, generalmente, le categorie sono considerate equiprobabili, e che l'evidenza è un fattore identico per ogni categoria, il valore che vogliamo calcolare potrà essere calcolato come:

```
P(E | F) = P(F1|C)*P(F2|C)*…*P(Fn|C)
```

Che significa che dato un documento, il valore che ci permette di ipotizzare a quale categoria appartenga, è data dal massimo prodotto di tutti i valori di probabilità di ciascuna parola del vocabolario in relazione a ciascuna categoria.
Il concetto può essere spiegato meglio con un esempio pratico.
Supponiamo di prendere un vocabolario di 5 parole:

W = [cane, soldi, deputato, calciatore, locale]

e di avere 3 categorie:

C = [sport, economia, tempo libero]

La prima cosa da fare sarebbe addestrare il classificatore. Per farlo bisogna calcolare le tabelle di probabilità di ciascun termine per ciascuna categoria e per farlo si passa attraverso l'addestramento. Per esemplificare, si prende un insieme di 3 vettori (documenti) per la categoria "sport":

v1=[1,0,0,1,1]
v2=[1,0,0,1,0]
v3=[0,0,0,1,1]

La riga relativa alla categoria "sport" per la tabella delle probabilità sarebbe:

sport = [2/3, 0/3, 0/3, 3/3, 2/3]

Si supponga di avere anche le altre due categorie e di avere dunque la seguente tabella delle probabilità positive (quelle negative si calcolano come "1 – p"):

[2/3,0/3,0/3,3/3,2/3]
[1/3,2/3,3/3,0/3,0/3]
[2/3,1/3,0/3,1/2,3/3]

Se volessimo classificare il seguente documento: "il calciatore ha giocato come un cane", a cui corrisponde il vettore [1, 0, 0, 1, 1] potremmo calcolare i tre possibili valori come:

v(sport)=2/3(1–0/3)*(1–0/3)*3/3*2/3=4/9=0,44*
v(economia)=1/3(1–2/3)*(1–3/3)*0/3*0/3= 0*
v(tempo libero) = 2/3(1 – 1/3)*(1 – 0/3)*3/3*2/3 = 8/27 = 0,30*

quindi *max(V) = 0,44* cioè: il documento è classificato come "sport".

Emerge subito come il risultato sia fortemente dipendente dai dati con cui viene addestrato il sistema e soprattutto esso non è esente da condizioni limite. Per esempio, l'assenza della parola "deputato", dal momento che è presente nel 100% dei campioni di economia, andrebbe ad azzerare la probabilità nonostante altre parole potrebbero aumentare significativamente lo stesso valore. Questo sarebbe un caso di sotto-stima dell'insieme di addestramento. Al contrario potrebbero esserci significative differenze nel numero dei documenti utilizzati e una categoria potrebbe essere favorita. In questo caso si ha un problema di sovra-dimensionamento. [6]

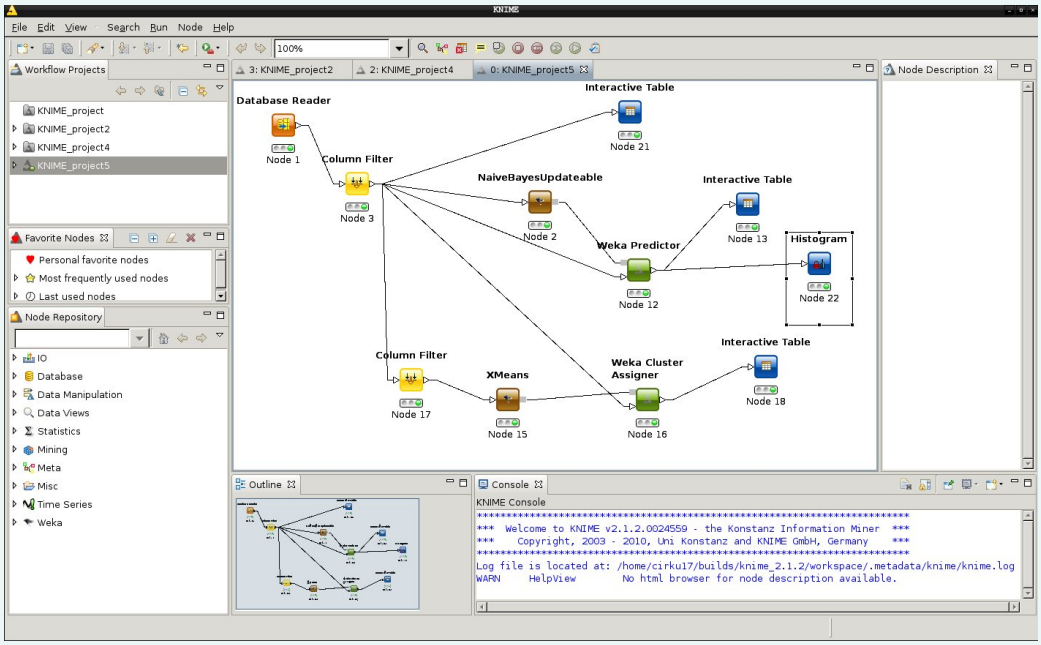
Il software

I software che possiamo usare dipendono dalle nostre esigenze (il più famoso ed utilizzato a livello universitario è il sistema R, oppure tika su server apache).
Ma quello che ho trovato più intuitivo ed efficace è **KNIME**, open-source scaricabile da:

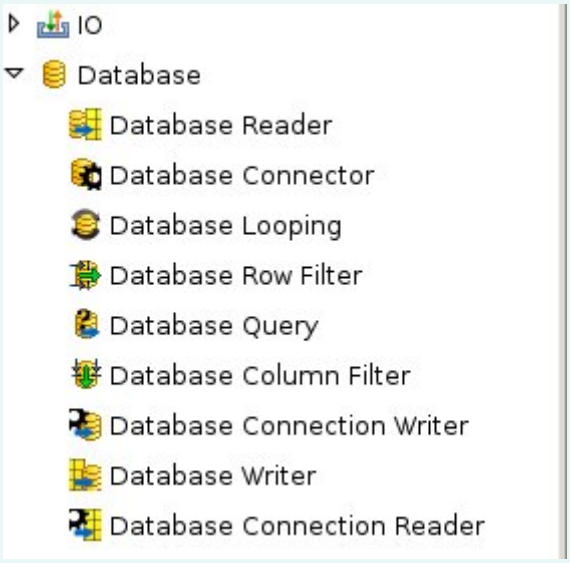
http://knime.org

dove troviamo anche la documentazione necessaria (http://tech.knime.org/documentation) e anche una quickstart guide.
Non mi dilungo troppo su questo software in quanto ritengo che la quickstart guide e il tutorial sul sito sopracitato siano più che esaurienti per padroneggiare il programma.

Analizziamolo solo a grandi linee:
uno screen del programma in funzione può essere questo:

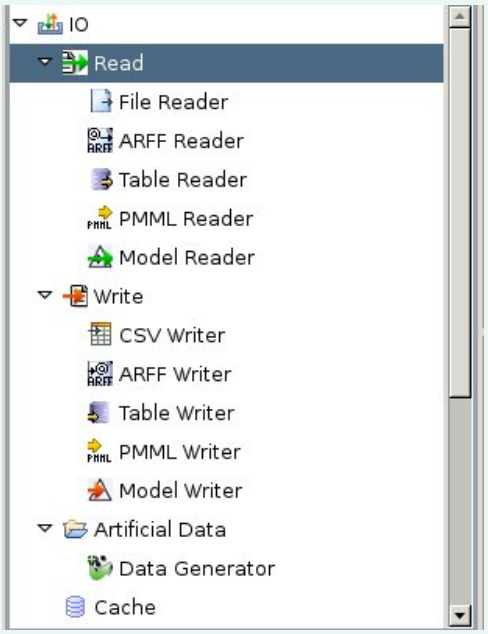


Il programma si gestisce per "nodi" creando una specie di diagramma di flusso, in cui ogni nodo ha una sua funzione, quest'insieme è chiamato appunto flusso, "workflow".
Si parte dal selezionare la sorgente, ossia il database da cui estrarre i dati: lo facciamo scegliendo il nodo desiderato:



Potete vedere che la scelta può ricadere su un database MySQL o PostgreSQL in locale (Database Reader), aggiungendo poi filtri sulle colonne o sulle righe (Database Row e Column Filter). Nella sezione Data Manipulation ci sono poi gli strumenti per operare su di loro.

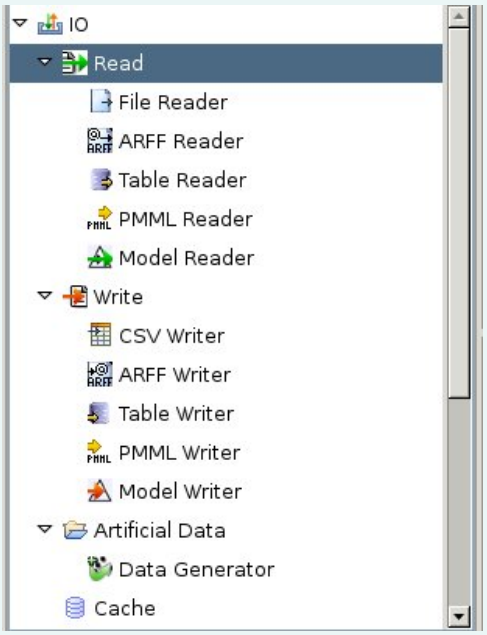
Ci sono anche strumenti per operare su file, in lettura o in scrittura:



i file possono essere dei tipi che vedete, ovviamente db-based, come CSV, ARFF (Attribute-Relation File Format),

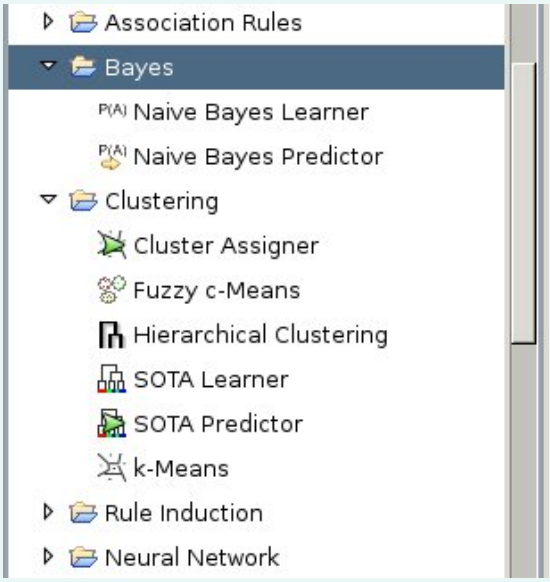
PMML (Predictive Model Markup Language) o anche una tabella di database.

Il mining vero e proprio avviene attraverso i nodi che applicano l'algoritmo vero e proprio:



qui troviamo tutti gli algoritmi di mining disponibili di default in KNIME;possono essere ampliati con plugin.

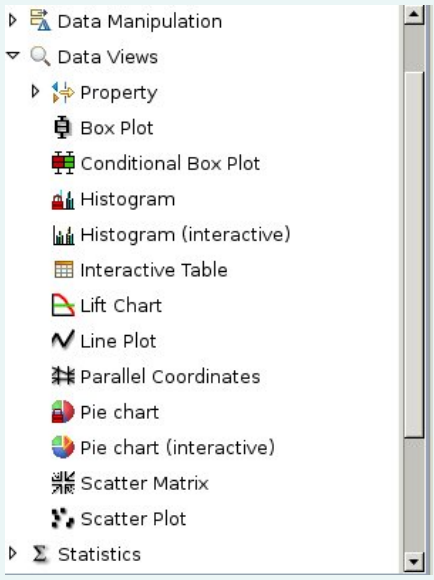
Sono già presenti comunque gli algoritmi più utilizzati,classificazione Naive-Bayesiano, k-means, c-means, X Means, ma c'è anche l'implementazione di Neural Networks e di uno strumento per creare regole personalizzate.



Come abbiamo visto con l'analisi del processo di mining c'è bisogno di più elementi per costituire un percorso di estrazione dei dati.

A questo scopo infatti KNIME usa separatamente i nodi "learner" e "predictor", che rispettivamente "imparano" la regola ("addestramento") e fanno una previsione su di essa.

Ci sono anche i nodi con il compito di "clusterizzare" i dati a loro passati.

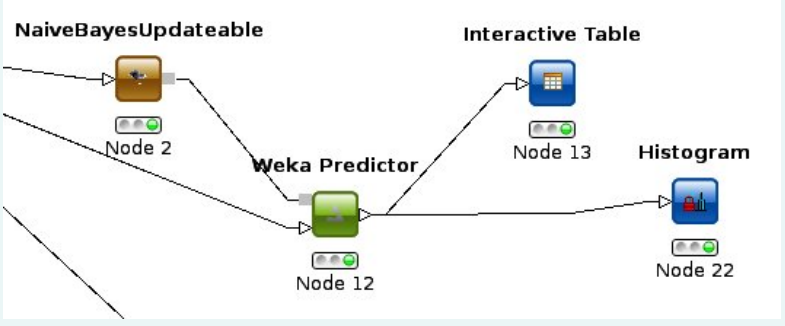


Fondamentale è poi la visualizzazione dei dati:

Ognuno di questi nodi permette una diversa esposizione del dato uscente dal workflow.

Ci sono istogrammi, diagrammi a torta,tabelle, grafici cartesiani etc.

Il workflow, si compone collegando i vari nodi, ovviamente secondo logica, attraverso linee di connessione. Ogni input e output di ogni nodo può collegarsi infatti solo ad un tipo di input e output con lui compatibile.



Il modo di costruire il flow è quindi come potete vedere molto intuitivo.
Comunque, la descrizione di ogni singolo nodo e tutte le operazioni possibili sul workflow e sulla GUI sono disponibili sulla documentazione ufficiale.

L'applicazione

Uno potrebbe chiedersi a questo punto cosa farsene di tutti questi procedimenti complessi se , come me, non dovete fare statistiche su un sito web con milioni di visitatori giornalieri, non dovete fare previsioni di marketing della vostra azienda o nulla di simile.
Le tecniche di web mining sono quotidianamente applicate per fare tutti i tipi di statistiche e i cosidetti "Web Graphs", dove troviamo informazioni sulle abitudini e sugli interessi degli utenti della rete, basandosi sulle loro frequentazioni web.
Un'altro aspetto non indifferente è l'aspetto "social" del web, sempre più emergente, che fa largo uso del data mining nel suo funzionamento.
Personalmente, ho iniziato ad interessarmi al data mining quando mi sono chiesto come facevano gli spider, i crawler, i web-bots a riconoscere le informazioni che gli interessassero in una mole così grande di dati, come può essere il prodotto di un paio d'orette di crawling in rete.
Qualunque spider o crawler solitamente si appoggia ad un database dove vengono salvati i risultati del crawling, organizzati in tabelle, colonne etc.
Volendo, per motivi che solo la nostra nostra fantasia può limitare, far partire il nostro spider/crawler preferito in cerca di chissà quali informazioni, anche se non a livello molto avanzato, ci servirà un metodo di catalogazione di questi dati per poter usufruire efficacemente dei nostri risultati, che altrimenti sarebbe solamente un ammasso enorme di informazioni disordinate.
Allo stesso modo si può analizzare un database di un sito web per estrarne quello che riteniamo più opportuno.

Applicando queste tecniche di data mining al web (web mining quindi), potremo avere padronanza dei dati fornitoci dalla rete, ottenendo un blocco di informazioni perfettamente ordinate e catalogate secondo il criterio da noi deciso.

Detto questo non mi resta altro che augurarvi: enjoy mining!

CirKu17

{Note}

[1]http://it.wikipedia.org/wiki/Data_mining
[2]http://www.microstrategy.it/Contents/Documenti/Capitolo_1_conoscenza_nascosta.pdf?CID=Capitolo_DataMining , http://en.wikipedia.org/wiki/Knowledge_discovery
[3]http://it.wikipedia.org/wiki/K-means
[4]http://it.wikipedia.org/wiki/K-nearest_neighbors
[5]http://it.wikipedia.org/wiki/Classificatore_bayesiano
[6]http://www.joproject.org/?p=273

SQL? NoSQL

Cosa sono i Database NoSQL ?

Un po' di storia

Con il termine **NoSQL** si intende un movimento che promuove l'uso di strumenti di archiviazione di dati non relazionali, quindi che non usano l'**SQL**. Il padre di questo termine è un italiano, Carlo Strozzi, che nel 1998 aveva sviluppato un database opensource relazionale che non aveva bisogno dell'interfaccia SQL.

Secondo Strozzi, è sbagliato usare "*NoSQL*" per indicare il movimento anti database relazionali; sarebbe più corretto chiamarlo "**NoREL**", o qualcosa di simile.

Questo nome fu introdotto nel gergo comune da un dipendente di last.fm, che in un convegno su queste nuove basi di dati chiamò comunemente NoSQL i database più recenti, che non rispettano l'acronimo **ACID** (ovvero: atomicità, consistenza, isolamento e durabilità), alla base del pensiero relazionale.

Quindi, in definitiva, con questo termine si intende indicare tutti i database che non seguono lo schema relazionale, e quindi che non hanno bisogno del suddetto linguaggio per essere interrogati.
Questo movimento contro i database relazionali ha avuto il massimo sviluppo l'anno scorso (sembra che già da quest'anno si cominci a ritornare ai relazionali, ma non sono fonti accertate) e società come Amazon li impiegano come ambiente operativo.

Una delle differenze che c'è tra i database relazionali e quelli non relazionali, è che nel secondo manca l'atomicità dell'informazione: se, ad esempio, un database relazionale avesse in una tabella gli articoli, dovrebbe esserci un'altra tabella "autori" per riconoscere gli autori.

In **XML** ad esempio, l'articolo potrebbe avere nel tag articolo, un attributo autore o categoria che spiega senza per questo aver bisogno di un'altra tabella, oppure se è più di uno, potremmo usare più elementi autore dentro l'elemento articolo.
Da questo esempio possiamo quindi chiarirci l'idea della non unicità dei dati, perché, ad esempio, se lo stesso autore scrive più articoli l'informazione non è più unica ma diventa ridonante.
Il processo di far diventare i dati non ridonanti si chiama normalizzazione e serve per occupare meno spazio possibile su un database stabilendo quali sono i dati, classificandoli per natura e dopo stabilendo le relazioni; per usare questi dati si attua un processo chiamato di "denormalizzazione", infatti i dati prodotti dalle query molte volte sono ridonanti (basti pensare se ci viene chiesto l'indirizzo di ogni lavoratore che produce un prodotto, ci sarà ogni volta l'indirizzo del lavoratore per ogni prodotto).

I dati nei database NoSQL invece sono già denormalizzati perché non seguendo il pensiero delle relazioni non possono seguire questo metodo di memorizzazione dei dati.

Panoramica dei database

I primi database (quando non c'era ancora il modello relazionale) furono di tipo **key/value**, ovvero ad ogni valore veniva data una chiave e per trovare un valore si ricorreva semplicemente ad un "puntatore" alla chiave; chi ha familiarità con la programmazione avrà già usato questo metodo negli array.

Il pregio è sicuramente che se dobbiamo ricevere un valore e conosciamo la chiave questo metodo sarà quello più veloce; il difetto è che non si possono costruire relazioni e che un valore è legato ad una e una sola chiave e soprattutto ad una chiave è associato un solo valore.

Esempio: in un blog noi usiamo un database key/value, il nostro articolo viene chiamato per id e fin qui noi riceviamo il testo tranquillamente.

Se noi volessimo gli articoli per autore, o nel testo scriviamo in qualche modo che poi il nostro codice lo riesca a ritrovare, l'autore, oppure il nostro database, non potrà dare il risultato. In realtà veniva, e viene utilizzata, una sintassi particolare per cui con qualche espressione regolare si riesce a ricavare qualsiasi informazione dal testo, però questo è difficile.

Questi database possono essere in un file oppure in memoria (in php, che è il linguaggio principe del web dove sono maggiormente utilizzati i database http://it.php.net/dba, è una libreria per i database su file e http://php.net/memcached per quelli in memoria), ovviamente non mi soffermo sui pregi e difetti del tenere un database in memoria RAM o in un file.

Un'evoluzione dei primi sono i database **document-oriented**, ovvero quelli che hanno un concetto simile al key/data, ma più evoluto; questi database sono quelli che avrete maggiormente visto e sono quelli su cui mi soffermerò di più.

Molte volte ho sentito parlare di CMS sviluppati con database su file XML per abbattere i costi di un server mysql e per "risparmiare" il tempo di una chiamata tramite socket al database.

Effettivamente il pregio di avere un database in uno o più file rispetto ad un database server è proprio dato dal tempo di chiamata pressoché inesistente. Questo è un ragionamento che si può fare con database di piccole dimensioni; un DB con molti dati (pensiamo ad esempio al database di Youtube) diventa invece inefficiente.

Su php abbiamo una libreria di gestione dei file xml che si chiama **SimpleXML**
http://php.net/manual/en/book.simplexml.php
che trasforma la struttura xml in un oggetto.

Questo approccio è molto intuitivo e rimanendo al nostro caro e amato blog in cui non sappiamo che database utilizzare potremmo effettuare la chiamata per ID semplicemente:

```
$xml = new SimpleXMLElement($xmlstr);
echo $xml->article[$id]->text;
```

L'elenco degli autori si potrà effettuare semplicemente con un ciclo e chiamando al posto del testo l'articolo.

Con poche righe di codice si possono anche effettuare ricerche e/o stampare l'articolo per autore, per categoria ecc.

Un pregio che ha una struttura dinamica come quella dell'xml è certamente la diminuzione dello spazio complessivo del database e la mancanza degli "spazi vuoti" nella tabella che di solito esce dopo query SQL.

La controparte SQL dell'xml usata soprattutto per la gestione di rubriche telefoniche, in qualche schema di alcuni database mysql e in programmi come Excel, è quella dei file **CSV**:

http://tools.ietf.org/html/rfc4180

La struttura del file **CSV** (*comma separated value*) è quella di un file testuale (come l'xml). La prima riga è riservata all'intestazione della tabella; nelle seguenti sono invece inseriti i dati separati da una virgola.

Ogni riga del file corrisponde quindi ad una riga del database, la conversione in memoria è relativamente semplice e in molti linguaggi sono presenti delle funzioni già predisposte, anche in questo tipo di file non ci sono gli "spazi riservati" quindi le celle vuote non occupano molto posto, vengono però segnati comunque con una virgola.

Ecco un esempio di file CSV:

```
Nome, cognome, email, numero di telefono CRLF
Luca, Unsigned, spam@dominio.com, 0123456789 CRLF
Luigi, Rossi, luigi.rossi@dominio.com, 0123456789 CRLF
Mario Rossi,,0123456789 CRLF
```

Ci sono anche dei database server che usano il principio del NoSQL.
Uno interessante è stato sviluppato dalla Apache e si chiama **CouchDB**, questo è stato pensato per le applicazioni in campo **AJAX**,
infatti usa il protocollo http per comunicare e scambia informazioni usando il linguaggio **JSON**.

Le caratteristiche principali sono quelle di un database **document-oriented**, senza schemi, non riserva spazio per i dati e supporta la distribuzione.

I file sono contrassegnati da un *_id* e ogni file ha anche delle subversion con la chiave *_rev*.
Le viste si possono chiamare con o senza autenticazione, questa viene gestita tramite i normali metodi http.

Per interrogarlo il procedimento è molto semplice, avendo installato curl sulla nostra macchina potremmo ad esempio dare un semplice

```
curl -X GET http://localhost:5984/databasename/_design/test/_info
```

Questa chiamata potrebbe produrre il seguente output

```
{
    “name”: “test”,
    “view_index”: {
        “compact_running”: false,
        “disk_size”: 4188,
        “language”: “javascript”,
        “purge_seq”: 0,
        “signature”: “07ca32cf9b0de9c915c5d9ce653cdca3”,
        “update_seq”: 4,
        “updater_running”: false,
        “waiting_clients”: 0,
        “waiting_commit”: false
    }
}
```

Chi ha esperienza con applicazioni AJAX potrebbe guardre con buon occhio un programma del genere perché gli risparmia il tempo di dover effettuare delle “traduzioni” dalla tabella creata dall’sql al JSON da dare in pasto a javascript.

Oltre ad avere una programmazione più rapida e quindi meno costosa, la macchina eseguirà molto meno lavoro, il che significa in grandi progetti un risparmio di risorse.
Possiamo anche creare delle richieste più complesse ma per questo vi rimando alla guida sito ufficiale.

Un altro *JSON-style* database è **MongoDB** che, stando a quello che dice il sito ufficiale, è un database scalabile, ad alte performance, opensource e document-oriented.
La struttura di MongoDB per l’utente è simile a quella di CouchDB, se volete potete visitare il sito ufficiale http://www.mongodb.org/

Un altro tipo di immagazzinamento di dati, simile al modello relazionale, è il database a grafo, basato sulla teoria matematica dei grafi, usa i “nodi” e gli “archi” per rappresentare le informazioni.
Questo tipo di database viene usato ad esempio all’interno di Mac Os X e di iPhone.
I nodi rappresentano delle entità semplici mentre gli archi sono le relazioni (che possono essere o non essere orientate) che ne derivano, nella vita comune troviamo i grafi nei diagrammi di flusso o nelle mappe concettuali.
Questi database usano lo stesso concetto per rappresentare i dati come se fossero quasi “delle grandi mappe concettuali”.
Questo fa risparmiare tempo nelle operazioni di join perché queste vengono a mancare, dovendo “semplicemente” seguire gli archi tra i nodi per riuscire a trovare le relazioni.
Si dice infatti che gli schemi **E/R** di questi tipi di database non sono molto rigidi.
La differenza nell’applicazione è la maggior velocità nel caso di dati con schemi molto mutevoli, invece con relazioni che non mutano in maniera molto veloce i database relazionali sono più veloci.
Come l’sql per interrogare questi tipi di database ci vuole l’**Object Query language** che ha una sintassi simile a quella dell’SQL, per selezionare le persone maggiorenni si usa una query del genere.

```
SELECT persona.eta
FROM Persone persona
WHERE persona.eta > 18
```

Un altro tipo di database poco comune è quello ad oggetti dove le informazioni sono rappresentate esattamente come lo sono gli oggetti in programmazione.
Quasi tutti i linguaggi di programmazione di oggi sono ad oggetti o almeno li supportano visto che questo paradigma assicura uno sviluppo più rapido e una maggior scalabilità dei programmi.
Questo tipo di database invece non ha riscosso molto successo, uno dei pregi dei database ad oggetti è la loro velocità di acquisizione di dati e della molteplicità dei tipi di dati che possono immagazzinare (dai vecchi int ai video e alle fotografie).

Proprio grazie a questa loro capacità lo *Stanford Linear Accelerator Center* ha raggiunto la nomina di più grande database del mondo (essendo stato il primo a superare i 1000TB e il più alto tasso di assorbimento mai registrato per un database commerciale di oltre 1TB per ora *[fonte wikipedia]*.

Ovviamente i database non relazionali non si fermano qui ci sono quelli tabulari, ad oggetti, a tuple quindi ce n'è un po'per tutti i gusti

In quali campi sono migliori i due tipi di database?

Lasciando da parte la soggettività e le idee politiche ci sono dei campi in cui uno è migliore di un altro.
Se i dati non hanno quasi nessuna relazione tra di loro vedo di buon occhio dei database key/value, ad esempio per liste di numeri telefonici oppure per applicazioni come le raimbow table.

Se ci sono poche relazioni e su per giù le query sono sempre le stesse e su siti che usano pesantemente l'AJAX io consiglierei un software come **Couch** o **Mongo**

Per i dati con molte relazioni e con relazioni complesse consiglio di rimanere sui database relazionali perché è più semplice gestirle e magari poi non si hanno le performace che ci si aspettava.
Se le relazioni però cambiano spesso sarebbe più utile e veloce usare un database a grafi, visto che sono nati apposta per contenere entità con relazioni molto mutevoli.

Per programmi come i gestionali e comunque programmi che girano in locale e che non hanno particolari bisogni di dover comunicare con un server esterno consiglio un database localizzato in un file scegliendo tra **xml**, **sqlite**, **csv** valutandoli e anche considerando i propri gusti

I database ad oggetti, come ho già detto, sono molto utili quando si devono tenere tanti record e quando il database ha dimensioni critiche, quindi su dati che si devono immagazzinare molto alla svelta vanno molto bene.

Insomma, alla fine non esiste un database migliore o uno peggiore in assoluto, dipende per cosa bisogna utilizzarlo e da lì se ne usa uno che assolva il meglio possibile alle funzioni.

FAQ

Q: *Se uso solo database noSQL non devo imparare l'SQL?*
A: Sui non relazionali non si usa l'SQL quindi in linea teorica è una capacità non richiesta ma come ho scritto prima secondo me è meglio sapere fare le cose in tutti e due i modi e poi decidere consciamente e non perché "così non imparo l'sql".
Ad esempio i database ad oggetti usano un linguaggio molto simile all'sql quindi una volta imparato in maniera approfondita uno si sa anche l'altro.

Q: *Posso migrare da SQL a un noSQL in un progetto già avviato?*
A: Perché no? Se un progetto potrebbe avere uno sviluppo migliore con un database noSQL e non costa di più la migrazione dei benefici io consiglio di passare all'applicativo che fa più comodo, ad esempio io proporrei a tutti quelli che usano i database su excel (non fatemi scrivere quello che penso a rigurado) di affidarsi ad esperti e passare ad una qualsiasi soluzione, un xml oppure anche un sqlite, dipende dal progetto

Q: *Se adesso studio i database noSQL non rischio di apprendere qualcosa di nicchia o che sarà sorpassato a breve?*
A: Si, potrebbe in alcuni casi succedere, come dice la rivista php|architect e molti altri il picco c'è avvenuto l'anno scorso e come ogni prodotto o ideologia nuova prima di capire se è realmente utilizzabile bisogna che si sgonfi un po'quel pallone dovuto al fatto della "moda", se sono stati di moda i database come mysql potrebbe esserlo anche CouchDB.
L'xml come database è largamente utilizzato e non penso andrà in rovina, più di nicchia sono i server che hanno un database non relazionale ma penso che con la diffusione di AJAX e altre tecnologie basate su JSON si potrebbe considerare conveniente sviluppare con un database che sia già JSON-oriented

Luca "Unsigned" Bruzzone

HTML5
come usare il nuovo tag <*video*>

Il nuovo linguaggio di marcatura **HTML5** definisce alcuni nuovi tag che cercano di definire uno standard innovativo per i browser di ultima generazione.

Fra le novità più interessanti, l'elemento <**video**> riscuote fra webmaster e programmatori una particolare curiosità per via delle recenti scelte industriali di produttori di dispositivi mobili (Apple in primis) che hanno deciso di non supportare la tecnologia Flash di Adobe per l'embedding dei filmati video sulle pagine web.

Le soluzioni più diffuse finora per incorporare filmati ed animazioni multimediali su siti ed applicazioni basate su HTML sono state quasi esclusivamente quelle proposte da Adobe Flash (basti pensare a YouTube), mentre Apple Quicktime e Microsoft Windows Media hanno ricoperto quote di mercato sensibilmente più basse.

Il supporto del nuovo tag <**video**> da parte di tutti i browser utilizzati nei vari dispositivi e piattaforme è ancora in fase di evoluzione, il che in pratica equivale a dire che non funziona dappertutto.
Attualmente il supporto pieno è garantito da Internet Explorer 9, Firefox 3.5+ e 4.0+,
Safari 3 e 4, Chrome e Opera, mentre restano fuori Internet Explorer 7 e 8 e Mozilla Firefox 3.0.
Sui dispositivi mobile il supporto è garantito su tutti i dispositivi Apple (iPhone e iPad in testa) e su quelli basati su Google Android.

Il tag <video> è stato progettato per essere utilizzato senza *plug-in* o script aggiuntivi di riconoscimento e si possono specificare diversi tipi di formati video e audio. Come sempre, i programmi di navigazione che non supportano il nuovo tag lo ignoreranno completamente, ma è possibile utilizzare questa caratteristica a proprio vantaggio per mostrare filmati usando i plug-in tradizionali di terze parti.

Esistono soluzioni come ad esempio "Video for Everybody!" che fa uso di HTML5 solo se questo è disponibile, altrimenti ripiega su QuickTime o Flash. Questa soluzione non utilizza Javascript e funziona praticamente su tutti i browser, compresi quelli progettati per i dispositivi mobili.

Utilizzare il tag <**video**> semplifica enormemente il lavoro del webmaster e del programmatore, sia nella progettazione che nell'implementazione. L'uso di Javascript si rende necessario solo quando si vogliono manipolare i contenuti *multimedia* oppure effettuare elaborazioni o interazioni con gli altri elementi presenti sulla pagina web.

Dal punto di vista del programmatore, verificare il supporto del tag <video> di HTML5 significa utilizzare una semplice funzione come la seguente:

```
function supporta_HTML5_video() {
    return !!document.createElement('video').canPlayType;
}
```

Questa funzione si basa sul fatto che il browser in grado di gestire HTML5 costruisce un oggetto *DOM* (Document Object Model) che supporta l'elemento **<video>** creando automaticamente il metodo *canPlayType*.

I Formati video

Spesso, per motivi di praticità, tendiamo a confondere un filmato video vero e proprio con il file che lo rappresenta dal punto di vista del sistema operativo e che diamo in pasto al nostro player preferito. Ma i "formati video" sono un po' come le lingue straniere. In edicola troviamo quotidiani italiani e stranieri che ci danno le stesse informazioni (le notizie del giorno), ma se si conosce solo l'italiano, allora, è ovvio, soltanto le notizie in italiano risulteranno leggibili. La stessa cosa accade con i video: per poterli visualizzare correttamente un browser deve conoscere la "lingua" (ossia il formato) in cui le informazioni sono state scritte. La "lingua" di un determinato video è il ***codec***, cioè la codifica utilizzata per convogliare i fotogrammi di un video in un flusso ordinato di bit. Naturalmente esistono decine di codec diversi che a seconda delle loro proprietà vengono usati su questo o quel dispositivo, sistema operativo, riproduttore multimediale, ecc. Allora qual è il miglior codec? Qual è quello più comodo da inserire nei propri progetti software?

In realtà non ci sono restrizioni sul codec video, audio o sul formato del contenitore o wrapper che è possibile utilizzare con il tag <video>. Un singolo tag può comprendere collegamenti a più di un file video. Il browser che effettua il rendering della pagina web contenente il tag **<video>** sceglierà di avviare la riproduzione del file video che supporta in maniera più efficiente. Cosa importante: sta al webmaster sapere quali codec sono supportati dai singoli browser disponibili sulle diverse piattaforme.

A parte l'impiego in soluzioni particolari, per garantire la piena compatibilità con il maggior numero di apparecchi o applicazioni, nella pratica di HTML5 si usano due codec principali: il primo è un codec che funziona su Safari e su iPhone/iPad ed è protetto da licenza software proprietaria. Come tale va quindi acquistato regolarmente prima di essere implementato in una qualsiasi soluzione server. Il secondo è invece totalmente *free* e funziona in modo nativo su browser open-source come Chromium e Firefox. Ecco una funzione che è in grado di riconoscere il codec proprietario supportato dai Mac e dagli iPhone:

```
function supporta_h264_baseline_video() {
    if (!supporta_HTML5_video()) {
        return false;
    }
    var the_video = document.createElement('video');
    return the_video.canPlayType('video/mp4'; codecs="avc1.42E01E, mp4a.40.2"');
}
```

La funzione effettua prima il controllo della compatibilità con HTML5 sfruttando la funzione *supporta_HTML5_video()* vista in precedenza. Se il supporto è presente e attivo, viene creato un elemento <video> fittizio per il documento web ma non viene inserito graficamente nella pagina web risultante (di fatto è invisibile all'utente del browser). Subito dopo viene richiamato il metodo *canPlayType()*- sicuramente disponibile dal momento che la funzione *supporta_HTML5_video()* ha restituito esito positivo - il quale s'incarica di controllare se il browser in uso è in grado di visualizzare il formato video H.264 Baseline ed il formato audio AAC LC, all'interno del file wrapper MP4 (da notare l'uso del *mime type* video/mp4 che è associato al formato audio/video MPEG-4).

Il metodo *canPlayType()* in realtà non restituisce un valore booleano, ma una stringa di testo (a dimostrazione del fatto che quando si tratta di formati video vige una certa complessità) i cui possibili valori sono:

- **'probably'** se il browser è abbastanza sicuro di poter visualizzare il formato
- **'maybe'** se il browser crede di poter visualizzare il formato
- **''** (stringa vuota) se il browser è certo di non poter visualizzare il formato

Per verificare invece che il formato video non proprietario sia supportato, basta usare la stessa funzione passando al metodo canPlayType una stringa di formato differente e nello specifico

```
return the_video.canPlayType('video/ogg; codecs="theora, vorbis"');
```

In questo modo stiamo chiedendo al browser (ad es. Mozilla Firefox) se è in grado di riprodurre il formato video Theora e quello audio Vorbis all'interno di un *wrapper* Ogg. Altro formato libero e non gravato da licenze o brevetti è il codec video WebM che probabilmente sarà incluso nella prossima generazione di Chrome, Firefox e Opera. In questo caso la funzione va modificata con la riga:

```
return the_video.canPlayType('video/webm; codecs="vp8, vorbis"');
```

Qui vengono usati il formato video compresso VP8, rilasciato da Google sotto licenza Creative Commons, mentre il *mime type* ed il file contenitore è per l'appunto il nuovo WebM.

Il tag <video> usato sul campo

Nella pratica non c'è una singola combinazione di *wrapper* e codec che funzioni su tutti i browser HTML5. E questo è un trend che non cambierà nel prossimo futuro, vista l'agguerrita competizione cui si assiste nel mercato dei formati video. Quindi per cercare di rendere un video avviabile su tutti i dispositivi e le piattaforme si dovrà effettuare la codifica del video in almeno due o tre formati diversi. La massima compatibilità si può quindi ottenere tramite la realizzazione dello stesso video in vari formati:

1. una versione che usi Theora per il video, Vorbis per l'audio in un file Ogg.
2. una versione H.264 Baseline e AAC LC (audio) in un file wrapper MP4.
3. (per il futuro) una versione WebM che contenga VP8 (video) e Vorbis (audio).
4. una versione "classica" in formato Flash (FLV o MP4) per il fall-back adatto ai browser più datati.

Per l'implementazione vera e propria occorre inserire all'interno del tag **<video>** i riferimenti (link) ai singoli file video, contemplando anche la possibilità che manchi del tutto il supporto ad HTML5.

Lo standard HTML5 offre due diverse modalità per includere filmati video in una pagina web ed entrambe utilizzano l'elemento **<video>**. Se si dispone di un singolo file video basta collegarlo mediante l'attributo src. Ecco un semplice esempio che usa il nuovo formato Webm:

```
<video src="vacanze2010.webm"></video>
```

Tecnicamente questa è l'implementazione minima del tag, ma è sempre meglio specificare anche gli attributi *width* e *height* che indicano le due dimensioni (larghezza e altezza) che sono state usate durante la fase di codifica:

```
<video src="vacanze2010.webm" width="320" height="240"></video>
```

Se una delle dimensioni effettive del video è leggermente minore di quella specificata il browser centrerà la finestra video all'interno del riquadro definito dagli attributi *width* e *height*. In ogni caso HTML5 cercherà di mantenere intatte le proporzioni del video.

Per default l'elemento <video> non mostrerà alcun controllo utente come pulsanti di riproduzione o barre di navigazione. È però possibile usare dei controlli standard forniti da HTML5 oppure costruire un'interfaccia utente usando HTML, CSS e Javascript. Alcuni metodi come *play()* o *pause()* e proprietà come current Time, volume e muted potranno certamente essere d'aiuto. Per utilizzare i controlli standard basta aggiungere l'attributo controls al tag <video>:

```
<video src="vacanze2010.webm" width="320" height="240" controls></video>
```

Altri due attributi opzionali interessanti sono *preload* e *autoplay*. Il primo ordina al browser di effettuare il pre-caricamento del file video appena la pagina viene prelevata dal web server. Questo è importante se il video è la parte centrale di tutta la pagina web. In caso contrario si può specificare il valore "*none*" per l'attributo *preload* così da minimizzare il più possibile il trasferimento dati ed il traffico di rete.

```
<video src="vacanze2010.webm" width="320" height="240" preload="none"></video>
```

L'attributo *autoplay* fa esattamente quello che il suo nome promette: istruisce il browser ad avviare il video non appena ha finito di caricarlo dal server. L'attributo standard *autoplay* è importante soprattutto perché permette al fornitore di contenuti di specificare la partenza automatica dei video nelle pagine web ma al contempo consente al fruitore dei contenuti web di impostare il suo browser in modo da ignorare questa funzione.

```
<video src="vacanze2010.webm" width="320" height="240" autoplay></video>
```

Finora abbiamo visto come gestire un singolo file video mediante il tag <*video*>. Ma cosa succede se vogliamo allargare la compatibilità dei contenuti sulla nostra pagna web utilizzando 3 diversi formati? Nessun problema: HTML5 ci permette di inserire un link per ciascuno di essi attraverso l'elemento <*source*>. Ogni tag <video> può contenere molti elementi <*source*>. Il browser scorrerà la lista delle sorgenti video e avvierà il primo che è in grado di supportare. Ecco un esempio:

```
<video width="320" height="240" controls>
    <source src="vacanze2010.mp4" type='video/mp4; codecs="avc1.42E01E, mp4a.40.2"'>
    <source src="vacanze2010.webm" type='video/webm; codecs="vp8, vorbis"'>
    <source src="vacanze2010.ogv" type='video/ogg; codecs="theora, vorbis"'>
</video>
```

Come si evince dal codice, ciascun elemento <*source*> contiene un link diretto ad un singolo file video tramite l'attributo src e fornisce informazioni sul formato video corrispondente tramite l'attributo type. Quest'ultimo appare un tantino complicato e di fatto lo è. Considerando l'ultimo dei 3 elementi <*source*>, esso specifica tre importanti informazioni sulla codifica del video: prima di tutto il mime type (video/ogg) del file, poi il codec video (Theora) e poi quello audio (Vorbis). Il file sorgente con codifica mp4 è leggermente più complesso poiché il codec video H.264 Baseline ed il codec audio AAC low-complexity (LC) possono avere diversi "profili" ed è quindi necessario specificare la versione completa dei codec utilizzati.

Il vantaggio principale di questa sintassi consiste nel fatto che il browser sarà in grado di controllare l'attributo type degli elementi <*source*> e verificare così se supporta o meno un determinato formato. Se il browser non supporta il video non caricherà il file corrispondente dal server. Ciò si traduce in un notevole risparmio di banda e gli utenti del sito potranno accedere più rapidamente al contenuto video di loro interesse.

Ad oggi la maggior parte dei browser web per i più diffusi sistemi operativi per PC e Mac supportano in un modo o nell'altro il tag <*video*> di HTML5. Anche i browser funzionanti su apparecchi e dispositivi mobile basati su iOS, Android, Windows Phone, ecc. con poche eccezioni sono in grado di far partire i video taggati con HTML5. Ma come supportare le versioni meno recenti dei browser, come Internet Explorer 8 o Firefox 3.0, tuttora largamente utilizzati? Semplice: continuando a sfruttare la tecnologia Flash ed i suoi plug-in disponibili per questi browser. A partire dalla versione 9.0.60.184 Adobe Flash supporta il codec video H.264 ed il codec audio AAC in file wrapper MPEG-4. Per combinare Flash ed il nuovo tag <*video*>, HTML5 ci viene in aiuto dal momento che è possibile inserire un elemento <object> all'interno del tag <*video*>. I browser che non supportano HTML5 ignoreranno l'intero tag <video> ed effettueranno il rendering dell'elemento <*object*>, il quale eseguirà il plug-in Flash attraverso un player come FlowPlayer o JW Player. I browser che invece supportano HTML5 troveranno una sorgente video fra quelle disponibili e ignoreranno del tutto l'elemento <*object*> annidato.

Ecco per concludere un esempio completo di codice HTML5 che utilizza un elemento <video> in presenza di un elemento <object> annidato che sfrutta FlowPlayer per garantire la fruizione del video in Flash ai vecchi browser.

```
<video width="320" height="240" controls>
    <source src="vacanze2010.mp4" type='video/mp4; codecs="avc1.42E01E, mp4a.40.2"' />
    <source src="vacanze2010.webm" type='video/webm; codecs="vp8, vorbis"' />
    <source src="vacanze2010.ogv" type='video/ogg; codecs="theora, vorbis"' />
   <object width="320" height="240" type="application/x-shockwave-flash" data="FlowPlayer.
swf">
        <param name="movie" value="plaver.swf" />
        <param name="allowfullscreen" value="true" />
        <param name="flashvars" value='config={
          "clip": {"url": "http://www.miosito.it/flash/movies/vacanze2010.mp4",
          "autoPlay": false, "autoBuffering": true}
        }' />
        <p>Preleva il video in formato <a href="vacanze2010.mp4">MP4</a>,
          <a href="vacanze2010.webm">WebM</a> oppure
          <a href="vacanze2010.ogv">Ogg</a>.</p>
    </object>
</video>
```

Combinando insieme HTML5 e Flash, il video dovrebbe essere visualizzabile su qualsiasi browser e dispositivo mobile.

cercamon

Cyberization

"Ottenere cento vittorie su cento battaglie non è il massimo dell'abilità: vincere il nemico senza bisogno di combattere, quello è il trionfo massimo."

"L'attacco migliore è quello che non fa capire dove difendersi. La difesa migliore è quella che non fa capire dove attaccare."

Sun Tzu, L'Arte della Guerra

Chi non è troppo giovane ricorderà come gli eserciti poggiassero su fondamenta quali: block/azione, la tartaruga, il juke-box, le saponette sotto la branda e altre tecniche belliche che trasformavano le spine in fantasmi.
Oltre agli aspetti precedenti (le fondamenta della vittoria, anche se la cosa è disdicevole) esisteva ed esiste anche la strategia di guerra; ad esempio la strategia terrestre organizza le forze per far avanzare il fronte e conquistare terreno delle forze nemiche.

Oil

Ho fatto l'esempio della strategia terrestre perché è quello più comprensibile ma non è certo l'unico; pensiamo ad esempio alla strategia navale dove un 'fronte' non esiste ma esistono rotte strategicamente importanti che vanno difese costantemente per punti, ad esempio la storica lotta tra sommergibili tedeschi e convogli alleati dove non avrebbe senso dire che un contendente doveva 'conquistare l'Atlantico'.

Rotte importanti, cioè vie di comunicazione dove transitano materiali strategicamente importanti. Vi viene in mente qualcosa di analogo?
La Rete è un sistema complesso dove circolano le informazioni più disparate possibili, comprese quelle riservate e quelle che implicano lo spostamento di enormi capitali finanziari.
I sistemi telematici interconnessi non sono diversi dal punto di vista strutturale, hanno soltanto diverse forme di protezione più o meno evolute. Tornando all'esempio navale un convoglio è un insieme di navi protetto da una scorta più o meno potente che transita in un passaggio omogeneo (il mare). Allo stesso modo un computer dell'NSA e il nostro PC di casa transitano sullo stesso mezzo (le reti) ma hanno protezioni delle informazioni enormemente diverse.

Se è vero che la fine della guerra fredda ha fatto passare in secondo piano l'adozione della strategia nucleare, è degno di nota il fatto che negli ultimi decenni si parla sempre più spesso della cosiddetta guerra asimmetrica.
I concetti di asimmetrie strategiche sono ben conosciuti da tempi immemorabili: i corsari caraibici, i Boeri, le tribù del deserto contro gli eserciti coloniali, fino ad arrivare alle teorizzazioni della guerriglia di Mao e Guevara e infine al modernissimo terrorismo internazionale. Il concetto di base è il medesimo, cioè il costo insostenibile di uno scontro diretto comporta la modifica completa del teatro di guerra in modo da rendere inutile la superiorità avversaria.
In questa particolare forma di guerra non può non inserirsi anche il controllo del cyberspazio, un nuovo fronte che non può essere compreso pienamente senza inserirlo all'interno dei canoni classici delle strategie asimmetriche.

Salt

Mao Tse Tung individua tre parametri fondamentali per la riuscita di un'azione di guerriglia:

1. L'appoggio attivo di una parte anche ridotta della popolazione locale (i guerriglieri), anche se la netta maggioranza risulta indifferente al conflitto.

2. La costituzione di zone sicure (santuari) interne o adiacenti al teatro delle operazioni; con funzioni logistiche, addestrative e di comando.

3. Un appoggio esterno da parte di un alleato

In questi tre elementi si possono ritrovare le organizzazioni di tutti i fenomeni insurrezionali che la storia ci ha mostrato, allo stesso modo in cui è possibile individuare anche le cause dei successi e delle sconfitte degli stessi.
La cosa che colpisce ancor di più della perfetta individuazione dei fattori è la sostanziale indipendenza da aspetti strettamente militari, la capacità organizzativa e logistica risulta la vera ragione di vittoria anche se il mezzo militare è nettamente inferiore a quello dell'avversario.
Se quindi la guerra cibernetica è una recente forma di lotta asimmetrica allora dovrebbe potersi inquadrare in questi canoni. Un elemento però va aggiunto nella trattazione: è un fattore specifico che più di un carattere particolare sembra essere un superamento epocale delle regole classiche.

La superiorità militare di uno dei due contendenti è sempre stata legata strettamente alla sua superiorità tecnologica. La tecnologia superiore ha sempre significato disparità enorme tra le forze in campo, in cui le truppe favorite erano quelle più avanzate anche a scapito del numero di uomini e mezzi disposti sul teatro di guerra.
Il concetto strategico Napoleonico di 'massa' lanciata in battaglia è già entrato in crisi durante la Grande Guerra con l'avvento delle mitragliatrici e dei gas tossici; oggi con l'avvento dell'arma atomica non ha alcun significato contare il numero dei soldati in campo o dei carri armati presenti al fronte.
La disparità delle forze basata sul gap tecnologico ha il limite di funzionare sempre meno quanto maggiore risulta il livello medio di diffusione della tecnologia stessa.

Scrive l'esercito Americano nel suo documento TC 31-93 (Paramilitary and Non-military Organizations and Tactics):

Cyberterrorism is not only about physically damaging systems, inserting worms or virus but also about facilitating communication and intelligence gathering. Evidence confirms that terrorists are using information technology and the Internet to communicate, formulate plans, recruit members, and raise funds. […]
The primary advantage is that all of these acts can be accomplished undetected (from the terrorist's home or another remote location) and prove to be extremely hard to trace. For the price of a computer and a modem, an extremist or would-be terrorist can become a player in national and world events.

[Il cyberterrorismo non consiste soltanto nel danneggiamento fisico dei sistemi, inserendo worm o virus ma anche nel facilitare la comunicazione e l'acquisizione di informazioni. L'evidenza conferma che i terroristi stanno usando l'information technology e internet per comunicare, formulare piani, reclutare membri, raccogliere fondi. […]
Il vantaggio principale è che tutti questi atti possono essere compiuti in maniera nascosta (dall'abitazione del terrorista o da un'altra località remota) e si dimostrano estremamente difficili da tracciare. Al costo di un computer e un modem, un estremista o un aspirante terrorista può divenire attore degli eventi nazionali e mondiali.]

Sulla base di questo fattore e di queste parole possiamo quindi riformulare quanto detto per adattarlo e identificare il problema con maggiore completezza.

1. L'applicazione delle tecniche di cyber-warfare e cyber-terrorismo necessita di un ristretto gruppo di membri specializzati, dotati di apparati informatici facilmente acquisibili a costi nettamente ridotti. Oggi in ogni paese del mondo è disponibile accesso alla Rete per un numero di persone magari esiguo ma sicuramente ben maggiore delle necessità richieste dall'organizzazione.

2. La necessità di apparati logistici di supporto risulta praticamente nulla in quanto tutta l'azione si può svolgere in rete e senza muoversi dalla postazione, inclusa la comunicazione tra i membri, l'acquisizione di risorse, la pianificazione dell'attacco.

3. Il luogo di residenza di tutti o parte dei membri è ininfluente sulla gestione dell'attività operativa; la stessa origine dell'attacco può essere facilmente mascherata ed è alquanto difficoltoso individuarne l'origine reale.

Qualcuno sicuramente ricorda come il governo degli Stati Uniti abbia dichiarato che nel primo decennio di questo secolo un'intensa attività dello spionaggio cinese abbia sottratto per anni una quantità notevole di informazioni di carattere militare riguardanti la progettazione di sistemi difensivi.
Nel 2007 in un momento di crisi politica nello stato del Kirghizistan l'intero paese risultò per un certo periodo off-line, un attacco massiccio proveniente naturalmente da server esterni al paese e controllati dall'esterno del paese.
Nell'estate del 2010 la scoperta del worm Stuxnet presentò la prima applicazione rootkit pensata espressamente per raccogliere informazioni sui processi industriali e tecnologici probabilmente a danno del programma nucleare iraniano.
Questi sono soltanto alcuni esempi di come la guerra digitale non sia un fenomeno futuribile e teorico ma abbia già raggiunto livelli fattuali decisamente maturi.
D'altra parte è ben più elevata la possibilità di utilizzare la Rete sia per l'acquisizione di capitali tramite operazioni di phishing, sia per espandere l'azione di organizzazioni e governi tramite tecniche di propaganda e disinformazione.

Gli stessi esempi mostrati in precedenza confermano anche come la militarizzazione del cyberspazio sia stata a lungo sottovalutata da alcuni tra i maggiori attori internazionali.
L'approccio di molte nazioni avanzate è infatti stato quello di considerarlo un fattore di contorno non solo rispetto alle forze tradizionali ma addirittura rispetto ad altre forme di lotta asimmetrica; in molti casi il problema è stato visto come fatto di ordine pubblico più che come minaccia di tipo bellico.

Nonostante la nascita di Internet per motivi strettamente legati alla comunicazione all'interno delle forze armate, negli Stati Uniti l'interesse alla sicurezza della Rete ha origine a partire dall'ormai lontano 1998 quando si capì l'importanza strategica di Internet per le sue implicazioni economiche. La protezione di tali sistemi si rivolgeva però principalmente all'analisi e alla prevenzione di atti di *cyber-crime* ben lontani dalla visione militare dell'uso della Rete.
Soltanto l'amministrazione dell'attuale presidente Obama ha definito lo spazio virtuale come risorsa strategica di interesse nazionale avviando studi e test sulla capacità di sorreggere attacchi di tipo cibernetico su larga scala.
La scelta americana di interessarsi al fenomeno in maniera decisa, uscendo dalla visione economica, ha mosso immediatamente anche i paesi europei con varie iniziative a carattere interno e sovranazionale. Sebbene queste mosse sembrino prendere in mano la situazione in modo da tenerla sotto controllo, alla luce di altre iniziative, il cammino da percorrere sembra ben più lungo del previsto.

Già nel 1999 l'Armata Popolare di Liberazione Cinese poneva le basi teoriche di un metodo di conflitto che andasse oltre le barriere nazionali per essere combattuto nelle reti telematiche.
L'ipotesi all'epoca quasi fantascientifica fatta da una potenza minore che iniziava ad acquisire importanti tecnologie moderne per superare il gap con l'occidente, non mosse particolari preoccupazioni internazionali.
Tra il 2003 e il 2005 diffusi attacchi alle reti governative americane ed europee, noti come *Titan Rain*, furono identificati come provenienti da server cinesi.
Già nel 2004 *l'Institute for Security Technology Studies* indicava la Cina come un paese non solo interessato agli sviluppi della guerra digitale ma già notevolmente più avanzato dell'occidente:

Within the framework of an integrated national plan, the People's Liberation Army (PLA) has formulated an official cyber warfare doctrine, implemented appropriate training for its officers, and conducted cyber warfare simulations and military exercises.

[All'interno della struttura di un piano nazionale integrato, l'Esercito Popolare di Liberazione ha formulato una dottrina ufficiale di cyber-warfare, implementando un appropriato addestramento per i propri ufficiali e conducendo simulazioni di cyber-warfare e esercitazioni militari.]

Lo stesso istituto pubblica un nuovo documento riguardante quel paese nel 2008, dove lo identifica come l'unico paese ad aver sviluppato i cinque aspetti della capacità operativa di cyberwarfare:

- elaborazione di una dottrina operativa
- capacità addestrative
- capacità di simulazione
- creazione di unità addestrate alla guerra cibernetica
- sperimentazione di attacchi hacker su larga scala.

Come si può notare facilmente il divario è più che significativo rispetto alle potenze occidentali, cosa che si potrebbe notare anche riferendosi ad altre potenze emergenti in questo settore quali Russia, India e Iran (probabilmente anche la Corea del Nord, i cui dati sono però sempre discutibili).

Dal confronto si nota, come detto in precedenza, il riconoscimento tardivo di una minaccia (o risorsa) insita nella gestione malevola del cyberspazio; soltanto alcuni paesi hanno giocato d'anticipo creando strutture idonee all'impiego militare delle reti.
Particolarmente significativa in proposito è la posizione del nostro paese, che come molti ancora non riesce a trovare un confine preciso tra minacce criminali e strategico-militari.
Il CO.PA.SI.R Italiano nella sua 'Relazione sulle possibili implicazioni e minacce per la sicurezza nazionale derivanti dall'utilizzo dello spazio cibernetico' del 7 Luglio 2010 elenca in questo modo i metodi di attacco di tipo cibernetico:

Cyber-spionaggio:
Il cyber-spionaggio è l'attività di raccolta di informazioni sensibili, proprietarie o classificate, utilizzando strumenti di ricerca telematici e sfruttando internet, reti telematiche, software e computer.

Vandalismo:
Appartengono a questa categoria gli attacchi diretti a compromettere il funzionamento dei siti web, fra cui la diffusa tipologia DoS (Denial of Service). Salvo eccezioni, si tratta di attacchi riparabili rapidamente e non in grado di infliggere danni gravi.

Propaganda:
L'estensione di messaggi politici attraverso internet o qualunque mezzo che riceva trasmissioni digitali dalla rete, quali telefoni cellulari, palmari, ecc.

Distributed Denial of Service Attacks:
Aggressioni nelle quali un numero rilevante di computer, controllati dal medesimo attore, lanciano attacchi DoS coordinati contro un sistema obiettivo, per comprometterne il funzionamento.

Sabotaggio di equipaggiamento e strumentazione:
Rientrano in questa categoria, ad esempio, l'intercettazione o contraffazione di ordini militari trasmessi attraverso strumenti telematici. Inoltre, le attività militari che sfruttano computer e satelliti per il loro coordinamento sono il principale bersaglio di questo tipo di attacchi.

Attacchi a infrastrutture critiche:
Ad esempio, attacchi telematici ai sistemi di controllo di infrastrutture energetiche, idriche, del trasporto e della comunicazione.

Compromised Counterfeit Hardware:
Componenti hardware che presentano una preventiva installazione di software malevolo nascosto, anche all'interno del microprocessore.

A parte qualche critica tecnica alle spiegazioni, che ricordiamo sono dedicate a politici e non ad esperti del settore, ne risulta a parer mio una fusione (o sarebbe meglio dire "calderone"?) tra fenomeni non ben definiti e ancora meno riconducibili a soggetti precisi che ponendoli in atto ne stabiliscono il grado di pericolosità.
Quanto detto risulta ancora più evidente nell'indicazione degli attori degli attacchi: gestori di botnet, gruppi criminali, servizi stranieri, hacker, insider (lavoratori insoddisfatti), phisher, spammers, autori di malware e terroristi.

Per i lettori di UnderAttHack, che si suppongono essere più informatici che politici, è bene tentare qualche ipotesi un po' più pratica e precisa che riesca a distinguere il puro crimine informatico da operazioni che hanno effetti ben più gravi per la sicurezza nazionale.

Il fatto che Internet sia un enorme contenitore dove è possibile recuperare qualunque tipo di informazione è cosa ormai nota a tutti. È evidente che se le varie ricerche sono dedicate a scopi illeciti non esiste (e non può esistere) alcun filtro reale al recupero di materiale 'utile'.

Pepper

Ciò che ho provato a compiere è una piccola simulazione, certo in scala molto minore di quella che può svolgere uno stato sovrano o un gruppo terroristico ben organizzato e finanziato, ponendomi la seguente questione:

Supponendo di voler organizzare un piccolo gruppo di amici per creare un'organizzazione cyber-terroristica, potrei facilmente trovare in rete gli strumenti necessari?

Per non essere accusato di istigazione al crimine non ho voluto inserire alcun link. Se non credete a quanto scriverò vi invito semplicemente a eseguire le ricerche del caso.

La prima questione che mi ha fatto riflettere è stata proprio la base iniziale relativa a quel gruppo di amici di cui parlo. Questa cellula costitutiva sarebbe probabilmente composta da persone con discrete capacità informatiche, che probabilmente si conoscono dalla frequentazione di Internet ma in realtà non si sono mai visti. Il concetto di scalata dell'organizzazione da parte di investigatori anti-terrorismo (diciamo il 'modello lotta alle BR') andrebbe a perdere gran parte del suo significato. D'altra parte anche se io stesso diventassi un collaboratore non sarei probabilmente in grado di identificare gli appartenenti al gruppo direttivo nemmeno trovandomeli di fronte.
A questa teorica non conoscenza va aggiunto il fatto che potrebbero addirittura non essere nemmeno italiani o dichiarare false identità senza che io nemmeno me ne possa accorgere.

Il primo problema organizzativo sarebbe legato alle forme di comunicazione interne al gruppo e in tal senso le possibilità sono infinite: chat a comunicazione sicura, creazione di mail praticamente anonime e chi più ne ha più ne metta.

Risolto il fattore comunicazione si potrebbe passare all'atto organizzativo più pratico, con l'inizio dei lavori per preparare i mezzi di offesa.

In questo caso salta subito all'occhio qualche elemento abbastanza sconvolgente:

1. La semplicità nel trovare in rete codice sorgente di malware, rootkit e botnet su cui lavorare; compresa la possibilità di trovare supporto per eventuali problemi legati allo sviluppo di tali applicazioni anche in community accessibili a tutti.

2. La possibilità di lavorare su codice open-source di ottimo livello per gestire tutta una serie di necessità dell'organizzazione, dalla creazione di software maligno, ai mezzi di comunicazione, a sistemi steganografici e di criptazione.

3. La facile reperibilità di materiale relativo agli aspetti organizzativi per la gestione di gruppi terroristici.

La possibilità di reclutare nuovi membri o stringere alleanze sembrerebbe piuttosto semplice, direi quasi evidente data l'enormità di gruppi decisamente border-line (e parlo solo di siti in lingua inglese...) di matrice semi-razzista, semi-terrorista, militarista e via dicendo.
Mi sono anche reso conto di come esista anche una sorta di fondamentalismo cattolico-cristiano ben più diffuso di quanto ci dicano, però questa è tutta un'altra storia...

Altra cosa che stupisce particolarmente è la semplicità con cui stando seduti davanti al proprio PC si possano realizzare reti di finanziamento basate su phishing, società e conti esteri in maniera anonima e a costi minimi, incluso l'aiuto di una folta letteratura su come muoversi in questo campo.
Società estere che costituiscono società in paesi terzi, che acquistano server in paesi terzi, per svolgere attività in paesi terzi. Una vecchia storia a cui almeno all'apparenza nessuno ha ancora messo la parola fine, ma evidentemente è solo un'impressione e la sicurezza esiste...

Non parliamo poi dell'ampio ventaglio propagandistico che la Rete offre tra siti web, forum, social network, ecc. L'offerta per il pubblico può essere la più varia e flessibile.
La propaganda può anche essere indiretta, cioè attuata in luoghi diversi dalle sedi 'istituzionali' in maniera da sensibilizzare l'opinione della rete in relazione agli aspetti da noi pubblicizzati; un gruppo abbastanza ristretto potrebbe agire con nick diversi in un numero esteso di portali per trasmettere il proprio messaggio Messianico.

Se credete che quanto ho scritto siano chiacchiere fatte su un problema semplificato fino alla banalità avete tutta la possibilità di informarvi; vi assicuro che scoprirete quanto gli argomenti siano delicati e terribilmente ovvi. Fin troppo ovvi...

Conclusioni

Tutte le grandi invenzioni della storia hanno purtroppo creato vantaggi e pericoli; è poco credibile il fatto che l'uso malizioso del cyberspace sia relegato alle truffe sulle carte di credito. Probabilmente sarebbe un caso straordinario se i problemi si fermassero a quel livello.
Esiste sicuramente un problema di controllo legato anche alla difficoltà di uno screening preciso su vie di comunicazione di tale portata, d'altra parte questa nuova frontiera di guerra è agli albori tanto quanto i mezzi per ridurne gli effetti. Probabilmente in futuro si assisterà come in ogni aspetto tecnologico militare a una rincorsa continua tra mezzi di offesa e di difesa.

In primo luogo credo sia necessario rendersi conto del problema fino in fondo, in modo da stabilire i confini della materia individuando cosa è semplicemente illegale da ciò che si prospetta come vera e propria minaccia alla vita degli individui.
È forse il caso che il dibattito politico per l'ennesima volta esca dalle aule di discussione e cerchi di confrontarsi con gli aspetti tecnici della materia; allo stesso modo in cui i tecnici credo abbiano l'obbligo di denunciare il possibile pericolo.
Sto pensando che scrivendo questo articolo nel mio piccolo mi sono mosso in tal senso, però ammetto che sarei presuntuoso, quindi diciamo che ho aggiunto una lista della spesa ad argomentazioni ben più serie.

L'ultimo problema che voglio porre è relativo agli aspetti etici che solleva il caso.
Alla luce di quanto è accaduto e sta accadendo in molti altri ambiti del mondo delle tecnologie informative, credo che si ponga un timore legittimo verso il fatto che le valutazioni politiche si sostituiscano agli aspetti sostanziali della gestione del problema cyber-warfare.
Come visto in precedenza, si tende a mescolare aspetti che poco stanno insieme e che hanno pochi confini e questo pone la questione di come esista un chiaro rischio che nel nome della sicurezza si vada a imbrigliare la libertà dei cittadini, additando come pericolosa ogni posizione scomoda all'attività di coloro che hanno il dovere di controllare. Sicuramente gli interessi in gioco tenderanno a creare un terrorismo buono ed uno cattivo, come anche una propaganda buona ed una cattiva.
Sta a noi aprire bene gli occhi in modo da essere sempre vigili e diventare controllori di chi controlla.

Floatman

Note finali di UnderAttHack

Per informazioni, richieste, critiche, suggerimenti o semplicemente per farci sapere che anche voi esistete, contattateci via e-mail all'indirizzo **underatthack@gmail.com** Siete pregati cortesemente di indicare se non volete essere presenti nella posta dei lettori.

Allo stesso indirizzo e-mail sarà possibile rivolgersi nel caso si desideri collaborare o inviare i propri articoli.

Per chi avesse apprezzato UnderAttHack, si comunica che l'uscita del prossimo numero (il num. 14) è prevista alla data di:

Venerdì 27 Maggio 2011

Come per questo numero, l'e-zine sarà scaricabile nel formato PDF al sito ufficiale del progetto:
http://underatthack.org