



With financial support from the AGIS Programme  
European Commission - Directorate General Justice, Freedom and Security  
Contract nr. JAI/2004/AGIS/113 - December 2004

## ***Raggruppamento Carabinieri Investigazioni Scientifiche Reparto Investigazioni Scientifiche di Roma – Sezione Telematica***

**Seminario Internazionale – Roma, 23 e 24 maggio 2005**

# Spider software di supporto alle indagini sulla pedopornografia sul web

*Bruzzone R., Strano M., De Marco F., Rossi A., Innamorati M.*

### **Introduzione: il nuovo scenario criminale e la rete internet**

L'evoluzione e la modernizzazione di molte forme di crimine sono strettamente correlate all'ingresso dell'informatica e delle comunicazioni telematiche nella loro operatività. Diverse dinamiche emergenti caratterizzano in tal senso lo scenario criminale di questo inizio di secolo. Si sono consolidate attività delittuose legate ai grandi *flussi migratori* internazionali (prostituzione, droga, armi, traffico di organi, turismo sessuale ecc.) che implicano sovente l'esistenza di una leadership criminale localizzata nel Paese di origine e una serie di insediamenti (in altre Nazioni) stabili che gestiscono localmente i traffici illeciti. Tali attività necessitano di riservati scambi di informazioni che sembrano trovare nella rete internet un canale ottimale. La grande *criminalità organizzata* ha già da tempo individuato nella telematica e in particolar modo su internet un efficace strumento per il riciclaggio del denaro sporco proveniente da svariate attività illecite. Gli investimenti nella *net-economy* e i sistemi di trasferimento telematico dei fondi rappresentano così la nuova frontiera di espansione del crimine organizzato. Alcuni gruppi portatori di subculture violente ed illegali utilizzano il web per svolgere attività di proselitismo a bassi costi raggiungendo moltitudini di persone con i loro messaggi. Nel caso delle *sette pseudoreligiose* o di *gruppi razzisti*, ad esempio, si assiste al proliferare di siti web che offrono contenuti illegali talvolta ad alto rischio come pedofilia, violenza, esortazione al suicidio, esortazione all'odio razziale eccetera (Strano M., 2000, Rossi A., Innamorati M., 2004). Una nuova e più sofisticata forma di *terrorismo* minaccia le nazioni: non più quello che mira solo all'eliminazione fisica degli antagonisti politici ma una forma di attività eversiva che punta sulla guerra dell'informazione e individua anche nei gangli informatici della società tecnologica i principali *target*. Per numerosi esperti internazionali si tratta di una nuova generazione di terroristi molto più pericolosa di quelle del passato, definiti tecnicamente *cyberterroristi*. L'utilizzo della rete da parte dei nuovi terroristi trova anche finalità di propaganda e di proselitismo e il nuovo *media* sembra aver portato ad una modificazione della strategia comunicativa. Ma l'ambito dove internet sembra poter offrire maggiori opportunità di sviluppo ai gruppi

terroristici è quello logistico e organizzativo, soprattutto nelle comunicazioni segrete tra cellule distanti tra loro anche migliaia di chilometri (Strano A., Neigre B., Galdieri P., 2002). Con lo sviluppo di internet, gli studiosi e gli investigatori hanno dovuto rilevare la presenza di una nuova dimensione organizzata della *pedofilia*, (centrata prevalente sulla pornografia), che sembra essere in fase di incremento quantitativo (Strano 2000). La rete mette infatti in connessione pedofili di tutto il mondo apparentemente con minori rischi di essere scoperti vista l'enorme quantità di collegamenti che la rete accoglie. I pedofili sul web nella maggior parte dei casi fruiscono di pedopornografia (scambiata con altri pedofili o acquistata sui siti). In alcuni casi però creano siti che contengono elementi di legittimazione della pedofilia e tendono a creare reti di interconnessione internazionale.

### **Tecno-intelligence: le nuove tecniche di indagine sul web**

Un'efficace azione di contrasto alle nuove strategie "digitali" dei criminali deve così necessariamente passare per la rete, attraverso il controllo mirato delle *comunicazioni* (soprattutto chat ed e-mail) nell'ambito della Polizia Giudiziaria e attraverso la localizzazione di *segmenti di informazioni illegali* sui siti web, nell'ambito della Polizia di prevenzione e dell'intelligence. Gli autori del presente saggio, a tal proposito, stanno sperimentando dei software "spider" che sono in grado di "scandagliare" la rete isolando segmenti sensibili di informazione e che rappresentano una utile ed economica tecnologia di base per lo svolgimento di investigazioni telematiche e di *web-intelligence*. Tale tecnica investigativa consiste sostanzialmente nella ricerca di brani di conversazione che presentano caratteristiche simili ad una serie di *stringhe comunicazionali* precedentemente impostate nel programma e che sono tipiche dei pedofili, dei terroristi, dei razzisti ecc. Tali strumenti software di *analisi testuale* potrebbero essere in grado di localizzare ed evidenziare alcune informazioni significative, "pescandole" tra i moltissimi siti e newsgroup presenti sul web. Ad esempio, un software spider appositamente addestrato potrebbe tentare di localizzare sulla rete alcune informazioni illegali o pericolose, ad esempio esortazioni al suicidio, proposte da *sette sataniche* o da altri gruppi pseudoreligiosi. La localizzazione di contenuti valutati come "pericolosi" consente di attivare azioni investigative mirate.

### **Il software spider ICAA<sup>1</sup> per l'intelligence su internet**

Il team di ricerca sulla tecnointelligence dell'ICAA sta sperimentando alcuni software "spider" impiegabili in alcuni specifici ambiti investigativi. In particolare, i software sono in fase di applicazione nella ricerca di contenuti ascrivibili alla cosiddetta "pedofilia culturale", all'esortazione all'odio razziale e ai gruppi pseudoreligiosi distruttivi. Il sistema di web intelligence è composto da varie componenti software che vengono "lanciate" sulla rete e che operano in sinergia tra loro. Descriviamo nel dettaglio le singole componenti:

*Componente web Crawler.* Ha il compito di analizzare le pagine web dei siti target utilizzando le regole sintattiche dei linguaggi di programmazione con lo scopo di estrapolarne la struttura tecnica del sito. Elaborando il linguaggio naturale, con l'ausilio di tecniche d'analisi semantica, ne sintetizza il contenuto. Questa componente del sistema ICAA W.I. è composta da una batteria di programmi che lavorano in parallelo tra di loro, allo scopo d'aumentare la potenza esplorativa dell'intero sistema. Per portare a termine la loro azione cognitiva, estraggono da una base dati le regole e i dizionari che utilizzeranno nella fase di classificazione dei siti target. I risultati prodotti dall'elaborazione dei link (presenti sulla pagina) vengono poi inseriti all'interno di un database (dei link) per essere utilizzati nelle fasi successive dalla

---

<sup>1</sup> International Crime Analysis Association

“batteria” dei Crawler con lo scopo di individuare i prossimi siti target da esplorare. I crawler ICAA utilizzano inoltre alcuni accorgimenti per evitare di caricare il sito target con una mole di richieste, mettendo in allarme chi lo gestisce.

*Componente database dei metodi.* In esso sono contenute le informazioni che indicano ai Crawler, le regole da impiegare per classificare ed esplorare un sito. Come evidenziato al punto precedente, tali regole sono di tipologia diversa (semantica o sintattica), ma concorrono entrambe all’analisi ed all’interpretazione del testo. In questo archivio sono presenti anche i dizionari tematici, che hanno la finalità di rendere più semplice e mirata la classificazione della pagina tramite un’azione comparativa tra il linguaggio presente su di essa e i termini presenti nel glossario.

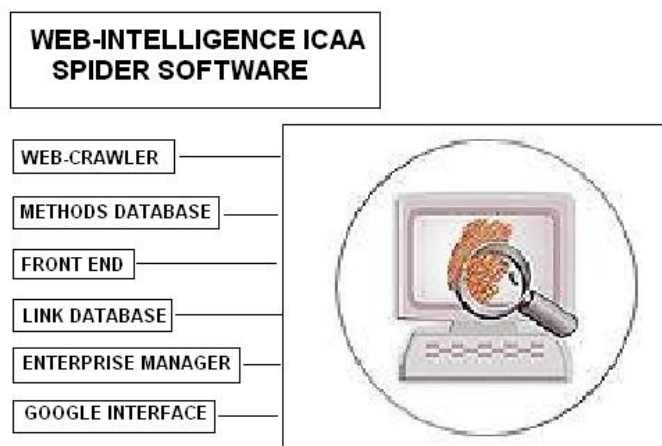
*Componente database dei link.* In questo archivio vengono memorizzati tutti i link che i crawler classificano nella loro fase esplorativa sulla base dei criteri di ricerca utilizzati. L’informazione sul link viene successivamente incrementata con alcuni dati aggiuntivi:

1. data di creazione e di modifica della pagina;
2. se il link analizzato punta verso risorse esterne (siti gestiti da terze parti) oppure verso pagine interne allo stesso sito;
3. il linguaggio di programmazione utilizzato per le pagine dinamiche (asp, php, jsp)
4. se il link punta verso immagini, memorizzando in questo caso il nome, le dimensione e la data del file;
5. se il link si riferisce ad un indirizzo di posta elettronica passandolo, se richiesto, ad un modulo d’interfaccia con Google per determinare se l’e-mail è stata utilizzata in altri contesti sul web.

*Componente di Enterprise Manager.* E’ il modulo d’amministrazione dell’intero sistema, da esso si controllano in maniera indiretta i Crawler grazie all’inserimento o la modifica delle regole nel data base dei metodi. L’enterprise offre, inoltre, le procedure per la gestione dei dizionari, degli elenchi dei siti target e la gestione dei criteri di “pesata”.

*Componente Front End d’Analisi.* Con questo sottosistema viene creata la reportistica di sintesi, grazie alla quale l’operatore di polizia potrà restringere il campo di ricerca, oppure valutare l’indirizzamento del sistema su determinate zone del web per ottenere una maggiore azione di controllo.

*Modulo d’interfaccia con Google.* Al fine di una integrazione dell’informazione alcuni link possono essere utilizzati come criteri di ricerca per Google in modo del tutto automatico, in questo caso il motore di ricerca ci fornirà ulteriori dati che concorreranno alla classificazione del sito web.



### Esempio di applicazione del sistema di web-intelligence ICAA

Volendo analizzare un sito razzista xxx (contenente esplicite esortazioni all'odio razziale) e i siti da esso referenziati (connessi), basterà inserire tale percorso all'interno del *data base dei link*, mentre, tramite il *data base dei metodi* si specificheranno i criteri che i crawler dovranno utilizzare nella fase di classificazione delle pagine rispettando le funzionalità operative che possono essere del tipo:

- ✓ analizza solo questo sito o/e i siti referenziati;
- ✓ individua gli indirizzi di posta elettronica e controllali con Google tramite il modulo d'interfaccia;
- ✓ utilizza il dizionario con le parole relative all'argomento del sito di interesse (es. odio razziale);
- ✓ valorizza il parametro di profondità dell'esplorazione<sup>2</sup>;
- ✓ crea un nuovo dizionario dei termini;
- ✓ analizza la frequenza dei termini utilizzati;
- ✓ crea il cluster<sup>3</sup> di connessione allo scopo di determinare i nodi che hanno più rilevanza all'interno della sottorete da esplorare;
- ✓ analizza le caratteristiche tecniche utilizzate per la creazione del sito;
- ✓ analizza i riferimenti circolari per i siti che si auto- referenziano tra di loro (analisi del banner di raccomandazione);

Tutte queste informazioni saranno fornite al sistema utilizzando *l'Enterprise Manager*. Una volta impostati questi parametri il *web Crawler (o programma spider)* ha a disposizione, una volta attivato, tutte le informazioni per iniziare il suo processo esplorativo. Si posizionerà sulla pagina iniziale cercandone di estrapolarne il contenuto tramite i dizionari, la classificherà e successivamente passerà all'analisi dei link presenti su di essa, memorizzandoli, se necessario, nel *data base dei link*. Una volta terminata l'esame della pagina il crawler ripeterà il ciclo precedente (lettura – analisi – classificazione) leggendo il link successivo (analizzato e classificato nella fase

<sup>2</sup> I siti analizzati vengono raffigurati graficamente dal sistema ICAA W.I. mediante una struttura ad albero dove la radice è la pagina di partenza (sito razzista xxx) e i rami successivi sono le pagine referenziate a partire dalla radice. La profondità è il numero di livelli che intercorrono tra foglie e la radice. Con una profondità pari ad uno i crawler si fermeranno non appena termineranno l'analisi delle pagine referenziate dalla radice. Maggiore sarà il valore maggiore sarà il livello di profondità dell'esplorazione.

<sup>3</sup> Gruppi di siti che si linkano a vicenda;

precedente) e aprodo la risorsa ad esso associata per poterla classificare. Una volta raggiunta la profondità d'analisi richiesta i crawler si arresteranno e, a questo punto, si passerà tramite il *Front End* alla stampa dei reports che evidenzieranno:

1. le interconnessioni statistiche tra siti (cluster)
2. i termini maggiormente utilizzati;
3. la tecnologia utilizzata;
4. le statistiche sulle risorse analizzate;
5. l'elenco degli indirizzi e-mail riscontrati e integrati con Google;

Ad esempio, sulla home page di un sito possono essere presenti dei link dai quali è possibile accedere a siti di stesso orientamento culturale. E' possibile poi che su questi ultimi ci siano dei puntamenti che riconducono alle pagine web presenti sul sito di partenza. Questi *loop* circolari sono utili per poter evidenziare una eventuale stretta relazione/stima/certificazione a garanzia dei contenuti. Un sito con maggiore visibilità che referencia un sito con minori accessi ribalta, in maniera indiretta, parte del proprio prestigio mediatico, assumendosi nello stesso tempo un ruolo di certificatore dei contenuti informativi presenti sul sito referenziato. Queste dinamiche "virtuali" evidenziano la natura dei rapporti tra i responsabili della gestione dei siti.

### **Conclusioni**

Come si evidenzia dalla descrizione applicativa, il sistema ICAA W.I. è in grado di analizzare in maniera approfondita il materiale presente sul web, trovando aree di interesse specifico (siti web a contenuto sensibile), evidenziando interconnessioni (anche non immediatamente apprezzabili) con siti di altri gruppi omologhi contraendo notevolmente i tempi di ricerca. Il suo impiego riduce infatti i tempi di lavoro rispetto a un'analisi di web-intelligence convenzionale (la lettura sistematica dei siti partendo da motori di ricerca commerciali) e fornisce informazioni "nascoste". Il sistema può infine essere utilizzato con successo per azioni di monitoraggio costante di alcuni siti di interesse, mostrando, attraverso report standardizzati, eventuali modificazioni del loro linking che possono evidenziare cambi di strategia comunicativa, attivazione e chiusura di alleanze, evoluzioni tattiche e strategiche.

### **Riferimenti bibliografici**

[www.criminologia.org](http://www.criminologia.org) (Telematic Journal of Clinical Criminology)

Galdieri P., Giustozzi C., Strano M, *Sicurezza e privacy in azienda*, Apogeo editore, Milano, 2001.

Innamorati M., Rossi A., *La Rete dell'Odio*, E. Walter Casini, Roma, 2004.

Strano M., *Computer crime*, Ed. Apogeo, Milano, 2000

Strano M., Neigre B., Galdieri P., *Cyberterrorismo*, Jackson Libri, Milano, 2002