

Il text mining¹ nelle applicazioni per la sicurezza

l'analisi automatica delle informazioni come strumento di lotta alla criminalità

L'osservazione di alcuni recenti eventi criminosi (riconducibili al fenomeno del terrorismo o comunque della criminalità) evidenziano una "escalation tecnologica" nella ideazione, progettazione e attuazione dell'azione criminale: l'uso che in alcune occasioni è stato fatto dei convenzionali servizi di Internet (web, email, newsgroup, forum, chat, ecc.) suggerisce una profonda trasformazione di modelli organizzativi ed operativi all'interno dei gruppi criminali.

Come noto, la diffusione capillare dei servizi di comunicazione basati sul "frame" Internet (web, e-mail, chat, forum, blog, ecc.) ha generato un "ambiente informativo" indipendente - cui spesso ci si riferisce con il termine "cyberspazio" - basato su regole diverse (talvolta opposte) da quelle comunemente accettate in ogni sistema organizzato. A questo ambiente informativo, chiunque può accedere pubblicando e distribuendo i dati e le informazioni che desidera, in piena autonomia, senza che sia richiesta una particolare "autenticazione"² dell'utente, né una specifica "attestazione di autenticità" dell'informazione pubblicata.

Tutto questo avviene con relativa semplicità tecnica, costi tutto sommato contenuti e - volendo - mantenendo l'anonimato sulla propria identità e localizzazione geografica. Ogni singolo utente della Rete (con la propria home page, mailing list, forum, chat, ecc.) contribuisce ad alimentare giornalmente questa "massa informativa", quasi fosse un multiforme organismo vivente in continua crescita ed evoluzione, che fagocita qualsiasi elemento gli venga somministrato senza far troppo caso al "chi", al "cosa" e soprattutto al "perché" lo fa.

Secondo una certa corrente di pensiero, questa peculiare "liberalità" del *Inter-Net(work)* rappresenta la più importante conquista sociale degli ultimi 50 anni e simboleggia valori di libertà di pensiero e uguaglianza da salvaguardare e preservare da qualsiasi tentativo di regolamentazione (anche se effettuato ai fini della sicurezza degli utenti stessi). Come ogni strumento tecnologico però, la rete tende ad assumere la natura di chi lo utilizza: può servire fini leciti e produttivi ma anche scopi illeciti e sovversivi. Recentemente, delle *facilitazioni* intrinseche al modello "open" di Internet hanno giovato anche organizzazioni criminali e terroristiche le quali, servendosi in modo diffuso degli strumenti messi a disposizione dalla tecnologia *consumer*, hanno potuto migrare architetture organizzative e prassi operative verso un modello più "software", caratterizzato da un impatto ben più *a basso profilo* sul tessuto sociale.

A basso profilo perché nel caso dei servizi internet - prima fra tutti la posta elettronica - per lo più non sono previsti particolari adempimenti burocratici per l'attivazione di un "account" ne tanto meno è necessario eleggere un domicilio fisico presso cui riceverla. E' inoltre possibile accedere ad un determinato "account"³ di posta elettronica da qualsiasi computer, in qualsiasi punto del mondo e con un sufficiente livello di anonimato. Se lo si desidera, è anche piuttosto facile implementare tecniche e strumenti crittografici anche complessi per la riservatezza della comunicazione.

La possibilità di anonimato, la capillare accessibilità ai servizi di rete, l'assenza di un'autenticazione dell'utente al momento dell'accesso⁴, unitamente alla velocità di trasmissione⁵ e all'immediatezza del recapito dell'informazione, ecc. sono i fattori all'origine della destrutturazione del paradigma

verticistico-piramidale dell'organizzazione terroristico/criminale convenzionale. Ed è proprio la peculiare "delocalizzazione" (contestuale, spaziale e temporale) dell'utente di Internet, che permette al criminale di superare quel legame fisico che fino a qualche tempo fa lo vincolava ad un territorio ristretto o comunque ad un'area geografica limitata nella quale operava.

Utilizzando i servizi Internet, colui che esercita il comando (o meglio la leadership) lo fa senza avere la necessità di rivelare la propria identità e localizzazione fisica. Ciò rende possibile - in una certa misura - l'elusione delle attività di sorveglianza effettuate tramite i convenzionali metodi di investigazione e scoperta. La disponibilità capillare di punti di accesso "non presidiati" alla rete (i già citati "internet point", ma anche le aule informatiche di scuole, università, come pure le postazioni di lavoro di enti pubblici e aziende⁶) contribuisce ad ampliare la potenziale "dispersione" e "mobilità" geografica delle parti in gioco fino ad un livello transnazionale.

Il convenzionale paradigma "piramidale" dell'organizzazione criminale cede definitivamente il posto ad una più agile ed efficiente architettura reticolare, nella quale l'incontro fisico non è più necessario per la veicolazione delle informazioni, che avviene invece in quel "non luogo" che abbiamo definito convenzionalmente "cyberspazio".

Da un punto di vista più prettamente strategico è da valutare con grande attenzione la possibilità che (così come normalmente accade nel marketing aziendale) le organizzazioni criminali possano essere interessate all'analisi strategica dei contenuti della Rete, al fine di acquisire notizie circa l'orientamento generale della comunità virtuale o il livello di consenso espresso nei confronti di particolari argomenti, tematiche o forme di lotta. Inoltre l'accessibilità a tutta una serie di informazioni pre-strutturate⁷ derivanti da elenchi e risorse pubblicamente accessibili, permette di venire a conoscenza di una vasta gamma di informazioni utili (indirizzi, attività, nomi, numeri di telefono, indirizzi email, ecc.) la cui "scoperta" fino a poco tempo fa necessitava di una lunga e complessa attività "ad-hoc".

Come si è visto, per quelle organizzazioni che perseguono fini sovversivi, eversivi o comunque illeciti, la "Rete" può diventare un alleato. Uno strumento strategico e di gestione operativa di straordinaria rilevanza, altamente tecnologico, potente, veloce, aggiornato. Uno strumento, un'*arma*, immediatamente fruibile e - in più - di libero uso e "detenzione".

La storia ci insegna che l'avvento di un nuovo "armamento" produce sempre modificazioni e adattamenti nella complessità dei conflitti, sia sotto il punto di vista strategico che tattico. A questo fenomeno non è sfuggita la tecnologia informatica. Nella odierna "società dell'informazione", la visibilità sulla rete comincia ad essere considerata dai gruppi sovversivi ed eversivi come un ideale strumento di propaganda. Tale presenza è da intendere sia come attività finalizzata a sensibilizzare e indirizzare l'opinione pubblica nei riguardi di certe tematiche, sia come esistenza celata di "nuclei" specializzati nell'osservazione e valutazione (ed eventualmente *pilotaggio*) delle opinioni degli utenti della "comunità virtuale" ("social engineering").

La liberalità (vera o presunta) di Internet ed il suo continuo "divenire" fa sì che - per la legge dei grandi numeri - essa rappresenti un campione attendibile delle opinioni, delle abitudini, delle convinzioni, delle emozioni dei suoi utenti: nella pratica un sondaggio continuativo di dimensioni planetarie (una panacea per i politici di ogni schieramento!). E' pertanto lecito ipotizzare che la "conoscenza" riposta in questo sterminato agglomerato informativo (soprattutto quella non palesata

o non immediatamente percepibile) sia portatrice di un elevatissimo valore strategico endogeno. A patto però di disporre di metodi analitici adeguati ad estrarre conoscenza "valida" da un agglomerato di informazioni potenzialmente inutili, superflue o comunque non attinenti allo scopo.

L'analisi di questo "patrimonio informativo comune" può costituire un elemento di fondamentale importanza nella lotta al terrorismo, alla criminalità organizzata e per la tutela dell'ordine pubblico. Reti criminali, pericoli di attentati, collegamenti tra insospettabili e nuclei sovversivi, programmi di lotta armata e molte altre informazioni "critiche" potrebbero infatti emergere dall'analisi contemporanea e incrociata di documenti, testi, messaggi apparentemente insignificanti (se presi singolarmente).

Solo relazionando e incrociando tra di loro dati e informazioni anche di natura diversa, provenienti da fonti opposte, anche apparentemente fra loro estranei, è possibile verificare la presenza o meno di collegamenti, legami, relazioni, rapporti che la individuale e disgiunta valutazione dei singoli documenti non evidenziava. Sfortunatamente in un sistema di informazioni così eterogeneo, vasto e complesso, la conoscenza "utile" risulta essere così frammentata e dispersiva e la mole di dati così estesa da rendere improponibile qualsiasi attività di analisi manuale da parte dell'uomo.

Proprio a questi fini è stata sviluppata negli ultimi anni una innovativa disciplina - il "data mining" - che coniugando informatica, matematica, statistica permette di automatizzare molte delle funzioni di analisi delle informazioni che normalmente verrebbero eseguite dall'uomo con grande dispendio di energia in termini di tempo e personale.

L'applicazione specifica del data mining all'analisi di informazioni testuali è il "text-mining". Il text mining permette l'analisi automatica dei testi, interpretandone il linguaggio, la sintassi, la semantica e evidenziandone le varie relazioni logiche, dalle più palesi alle più nascoste.

Uno dei più autorevoli esperti di text-mining l'Ing. Alessandro Zanasi⁸ asserisce che "*...attraverso il text mining si possono analizzare volumi immensi di informazioni, sia in tempo reale che in differita e si possono identificare relazioni e strutture che altrimenti sfuggirebbero alla capacità analitica umana*"⁹.

E' scientificamente provato infatti che il limite del cervello umano è quello di non poter passare agevolmente da una visione ampia ed omnicomprensiva di un problema ad una dettagliata e analitica, mantenendo al contempo la percezione totale del problema stesso; ciò è tanto più vero quanto più esteso è il problema in esame¹⁰. Ebbene, la tecnologia del text mining è stata sviluppata proprio avendo come "goal" il superamento di questo limite, obiettivo reso raggiungibile anche e soprattutto grazie alla continua, vorticoso evoluzione della tecnologia informatica che rende disponibili sistemi sempre più potenti e veloci.

Il text-mining si fonda su diverse fasi di analisi, ognuna delle quali affronta un aspetto in particolare della problematica. La prima fase è chiamata "preparatoria" e viene comunemente identificata come "preprocessing linguistico". In questa fase avviene la preparazione dei documenti attraverso un processo di normalizzazione, vengono risolte le ambiguità semantiche della lingua (es.: "lecca il lecca-lecca"), vengono effettuati il riconoscimento e la lemmatizzazione di espressioni (es.: Fabbrica Italiana Automobili Torino in FIAT) e l'indicizzazione automatica dei documenti.

L'informazione viene inoltre "strutturata" in modo da poter essere elaborata mediante metodi informatici.

In una successiva fase, quella dell'estrazione di conoscenza vera e propria, vengono identificati all'interno dei documenti i termini e le frasi di maggiore rilevanza, si individuano legami, connessioni, analogie presenti nei documenti, estraendo concetti e significati in essi contenuti. L'analisi può proseguire poi con la fase di "clustering" nella quale tutti i documenti vengono raggruppati in base all'argomento trattato (il sistema può riconoscere dal contesto se la frase "straziami ma di baci saziami" è contenuta in un documento relativo ad una campagna pubblicitaria di un noto cioccolatino famoso per celare all'interno un foglietto con delle massime sulla vita in quattro lingue, oppure se è una battuta presente in un romanzo rosa). Oppure con la fase di "categorizzazione" in cui il documento è riconosciuto come appartenente ad una categoria predefinita e a questa assegnato.

Infine, un'interfaccia utente avanzata permette, grazie ad un sistema di rappresentazione iconografica multilivello, di rappresentare visivamente argomenti, documenti, gruppi di documenti, relazioni e livelli di analogie, realizzando in tal modo una "immagine" chiara e definita della conoscenza estratta dalle informazioni sottoposte ad elaborazione. Il "text-mining" non deve essere confuso con l'attività dei motori di ricerca su Internet. Con questi ultimi, l'utente indica un argomento predefinito (la/le keyword) e il motore non fa altro che estrarre dalla base informativa tutti i documenti contenenti quella keyword. Il text mining invece analizza la base informativa (o parte di essa) ed estrae tutti gli argomenti che sono trattati nelle varie classi di documenti presenti, evidenziandone le relazioni semantiche. Il tutto praticamente in tempo reale grazie alle prestazioni dei calcolatori attuali.

In ambito economico questa tecnologia è adottata da grandi e piccole aziende, locali o multinazionali, ai fini della individuazione delle strategie di marketing dei concorrenti (più o meno dichiarate, più o meno segrete) o della valutazione della disponibilità di nuovi mercati, come anche per realizzare politiche di customer satisfaction efficaci. In sostanza in tutte quelle applicazioni dove è necessario recuperare la conoscenza "nascosta" da un insieme esteso di informazioni disponibili.

Quali potrebbero invece essere le potenzialità del text-mining nell'ambito dell'intelligence militare e per la sicurezza? Alcune tra le ipotesi più interessanti riguardano l'ambito della lotta al terrorismo. La verifica sistematica e l'analisi incrociata di grosse moli di informazioni (organizzate in "cluster" semanticamente omogenei) può permettere di risalire a intenzioni, propositi, intenti e strategie non ancora palesate, ma le cui tracce¹¹ sono state inconsapevolmente lasciate in documenti quali testi di discorsi, proclami, rivendicazioni, dichiarazioni di lotta, oppure rinvenute in intercettazioni di colloqui (comunicazioni telefoniche, via fax, email, chat, ecc.).

Lo stesse applicazioni trovano utile impiego nelle attività di intelligence investigativo connesse con la descrizione dei profili criminologici. Non è un caso che la Polizia Postale e delle Telecomunicazioni (che come noto è attiva anche nell'ambito della lotta alla pedopornografia su Internet) abbia implementato uno speciale progetto¹² volto alla produzione di "profili criminologici digitali" (*digital profile*) realizzati analizzando campioni particolarmente significativi di "comportamenti digitali" presenti nelle sessioni di comunicazione via chat. Gli elementi analizzati sono parole, termini e nickname usati, particolarità linguistiche ricorrenti, abitudini di digitazione e

interazione con la controparte, ecc.. Tale attività risulterebbe di fondamentale importanza ai fini dell'incremento dell'efficacia dell'azione investigativa convenzionale, soprattutto per quanto concerne la localizzazione di quei soggetti che abbiano evidenziato una condotta sanzionabile nei termini della normativa specifica della materia.

Il Dott. Marco Strano¹³, della Polizia di Stato, sostiene che *"la peculiarità della ricerca scientifica sul digital profiling... rende di difficile applicazione le tecniche di profiling classiche... Questa situazione necessita quindi della creazione di tecniche analitiche nuove, per certi versi pionieristiche e assolutamente sperimentali e centrate sull'impiego di speciali software..."*¹⁴.

Tecniche pionieristiche nelle quali il text-mining può occupare una posizione di grande rilevanza. Infatti come si è potuto ben notare in questi due esempi, in tutti quei casi in cui è necessario analizzare enormi quantità di informazioni testuali (ed Internet è *ancora* una applicazione dove l'informazione testuale è presente in maggior percentuale) e la dove queste informazioni sono "open source" (e Internet è *ancora* una applicazione tipicamente "open") il Text-mining produce i migliori risultati.

Certamente occorre evitare di spingersi troppo lontano con la fantasia e restare con i piedi ben saldi a terra ma, se pensiamo al web come alla proiezione "on line" delle idee, delle emozioni e dei sentimenti di tutti gli utenti della Rete, e se teniamo conto della diffusione di Internet nel mondo (sempre in continua crescita) ci accorgiamo di avere proprio sotto i nostri occhi una simulazione attendibile della società¹⁵ reale, con tutte le sue componenti (economiche, politiche, culturali, sociali) rappresentate in scala. Per certi versi potremmo stupirci di quanta conoscenza *valida* è disseminata nel "cyberspazio".

Ciò che dobbiamo augurarci e per cui dobbiamo lavorare, è che la Rete non divenga un ulteriore strumento che facilita la vita a chi persegue i propri fini illeciti danneggiando gli altri (terroristi, criminali, pedofili, criminali informatici, ecc.). Soprattutto non deve spaventare l'ipotesi di una *regolamentazione* di Internet o di una sua *osservazione* (ovviamente regolamentato ed effettuato a norma di legge). Ciò non sarebbe affatto una limitazione di quella fondamentale "libertà di espressione" così bene incarnata dalla Rete, semplicemente perché proprio "quella" libertà può essere difesa solo assicurando ad ognuno la possibilità di usufruirne in modo sicuro e egualitario.

In un futuro forse neanche troppo lontano, tecniche "analitiche" quali il text-mining, potranno verosimilmente essere gli unici strumenti a nostra disposizione contro gli abusi e le prevaricazioni di chi usa e userà Internet (o "la rete che sarà"...) come *arma*, anziché come *strumento* di crescita economica, culturale e sociale.

Ringraziamenti:

Ringrazio il Prof. Alessandro Zanasi per la disponibilità e la chiarezza con le quali ha voluto guidarmi nel corso delle mie (breve) incursioni nei meandri del text-mining.

¹ termine preso in prestito dal significato originale di "mining" che indica l'attività e le infrastrutture legate alla estrazione di minerali, metalli preziosi, diamanti dalla terra (Collins, English Language Dictionary, 1991).

² sistema o procedura che permetta di assicurare la possibilità di risalire al un soggetto fisico che ha effettuato l'accesso al sistema.

³ codice identificativo che permette l'accesso ai servizi.

⁴ autenticazione intesa come verifica della corrispondenza tra account e persona fisica. Si veda a tal proposito la possibilità di accedere tramite gli "internet point" a strumenti quali le "web mail" (caselle di posta elettronica on line);

⁵ intesa come quantità di informazioni scambiate nell'unità di tempo.

⁶ L'internet point (qualunque esso sia) scardina la correlazione tra "utente internet" e "linea telefonica di collegamento", elemento di fondamentale importanza ai fini delle indagini tecniche necessarie per risalire con certezza alla localizzazione fisica dell'indagato.

⁷ Ossia organizzate e ordinate attraverso una pluralità di relazioni logiche, utili ad identificare un servizio (es.: elenco telefonico abbonati, aziende, albi professionali, siti internet di emittenti radiotelevisive, quotidiani, ecc.)

⁸ Ufficiale dei Carabinieri in Congedo, ingegnere nucleare, professore universitario e socio fondatore di Temis SA (www.temis-group.com) è uno dei massimi esperti internazionali di text mining, con all'attivo numerose esperienze internazionali nel campo dell'intelligence militare ed economico.

⁹ A. Zanasi "Information Warfare, Business Intelligence, text Mining" - Infosecurity , 12 febbraio 2003

¹⁰ E' scientificamente provato che il cervello umano non riesce a visualizzare mentalmente un numero di elementi/oggetti superiore alle dieci unità.

¹¹ nomi propri, luoghi, obiettivi, bersagli, ecc.

¹² Progetto O.L.D.PE.PSY (*On Line Detected Pedophilia Psychology*)

¹³ Direttore Tecnico Psicologo presso il Servizio di Polizia Postale, Direzione Centrale di Sanità della Polizia di Stato, Centro di Neurologia s Psicologia Medica.

¹⁴ Strano M., "*Uno studio clinico e criminologico dei pedofili on-line*", Relazione al Congresso Internazionale della SOPSI (Società Italiana di Psicopatologia), Roma, Hotel Hilton, 26 febbraio 2003

¹⁵ meglio sarebbe forse dire "emulazione"