Laura Caserta
Giornalista, ha lavorato come Market Intelligence Analyst presso IDC prima di essere assunta dal Ministero del Tesoro dove ora si occupa, presso Consip, di implementazione di soluzioni informatiche e strategie IT. Presso la Link Campus - University of Malta ha seguito corsi di Business Intelligence & Security. Attualmente, collabora con il Centro Studi Intelligence e Security della suddetta Università, in veste di ricercatrice

Un approccio aggressivo alla conoscenza

LAURA CASERTA

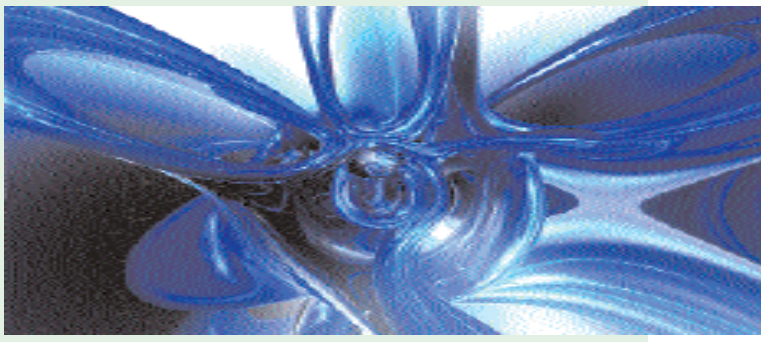
NEL QUADRO DI UN APPROCCIO COMPETITIVO ALLE RELAZIONI UMANE, LA RACCOLTA DI DATI, L'ACQUISIZIONE DELLE INFORMAZIONI E LA FORMAZIONE DELLA CONOSCENZA COSTITUISCONO UNO DEI FONDAMENTI, SE NON "IL" FONDAMENTO. COME ANALIZZARE LA CONOSCENZA?

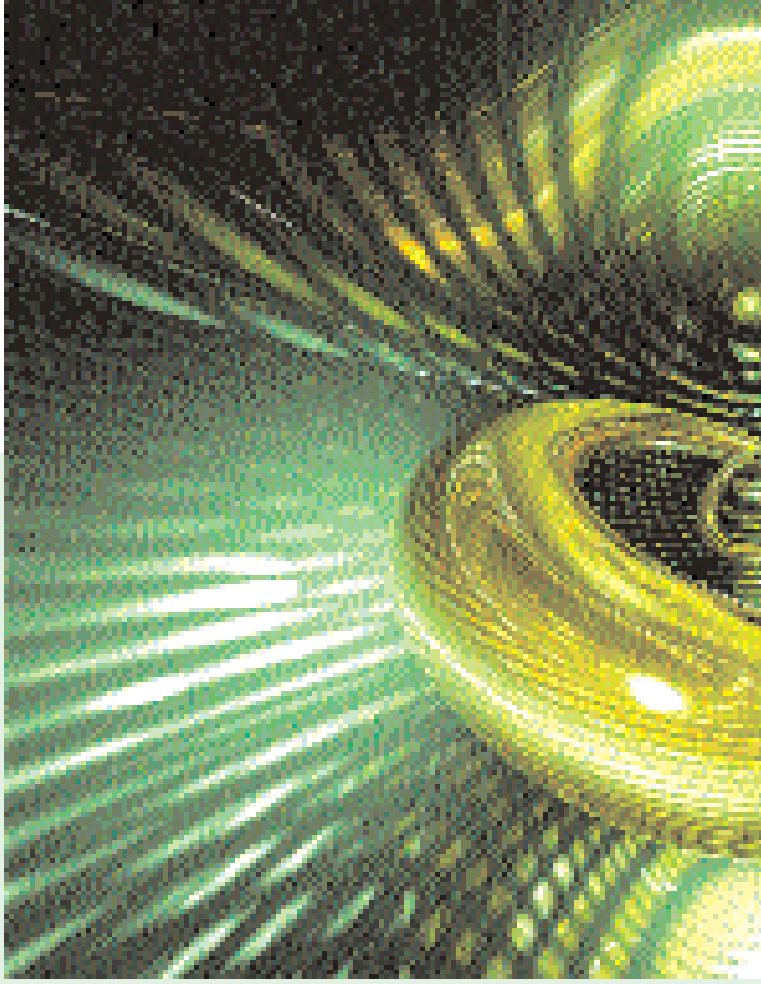
In un mondo dove tutte le organizzazioni, aziendali e/o governative, sono organizzazioni globali, solo quelle con la conoscenza necessaria per rispondere alle necessità globali possono sopravvivere. Questa conoscenza è ormai disponibile, grazie all'information technology (Internet, satelliti, tv), pressoché nella sua totalità, a chiunque.
Il problema è: data la sua mole, come analizzarla per estrarre l'informazione utile?
È questa la sfida che **Temis (Text Mining Solutions www.temis-group.com)** società fondata da un gruppo di esperti in knowledge management italiani, francesi, inglesi e tedeschi, ha raccolto con successo.
All'interno della società umana, l'interazione tra individui e tra organizzazioni assume spesso la forma di competizione. Competizione non è solo azione ostile di uno stato nei confronti di un altro; è competizione anche l'emulazione tra individui, la spinta all'innovazione tecnologica nel confronto tra due multinazionali o blocchi ideologici.
Nel quadro di un approccio competitivo alle relazioni umane, la raccolta di dati, l'acquisizione delle informazioni e la formazione della conoscenza costituiscono uno dei fondamenti, se non "il" fondamento. Per trarre un vantaggio dalle sue azioni, l'attore sociale - qualunque attore - raccoglie prima dei dati.
Senza dati la sua conoscenza della realtà è nulla. Senza dati egli si muove alla cieca, il risultato delle sue mosse è dominato dalla casualità.
Il dato tuttavia è muto. Costituisce il mattone base della conoscenza, ma di per sé non dà valore aggiunto.
È solo dalla relazione tra un insieme di dati che si ottiene informazione. Il valore di un'azione in Borsa non significa nulla. Unito al valore della stessa azione in giorni differenti, otteniamo un trend.
L'ultimo passo, sulla strada della conoscenza che un attore ha della realtà che lo circonda – realtà naturale e realtà sociale – consiste nella comprensione delle cause di un evento. L'informazione "le azioni X stanno salendo" ha valore limitato alla descrizione del fenomeno in un dato arco temporale. Aggiungere a questa informazione la comprensione delle sue cause, ci permette di creare un modello predittivo che apre un intero mondo all'attore che per primo lo realizza.
La conoscenza mostra la strada da seguire, i percorsi da evitare. Guida gli eserciti alla vittoria, li porta al disastro. L'operazione Barbarossa fu pianificata ed eseguita sulla base di una tremenda sottovalutazione delle capacità di mobilitazione – morale e materiale – del popolo russo. Sebbene l'analisi dei dati fosse stata correttamente condotta, troppi dati erano ca-

renti o errati.
Non solo nello scontro tra due sistemi ideologici, ma anche nella vita quotidiana, la disponibilità di informazioni adeguate è fondamentale per il successo. Che noi si voglia collaborare o scontrarsi con altri, che si desideri gestire o pilotare, o solo osservare e conoscere, in ogni circostanza dovremmo attuare una strategia, una chiara direzione di pensiero, e questa non può essere realizzata senza la percezione del mondo che ci circonda. Senza conoscenza non si possono prendere decisioni corrette. Una decisione sbagliata può uccidere.
Il Web regala oggi alla società umana una fonte inesauribile ed autoalimentantesi di informazioni. Non vi è apparentemente limite alla quantità di dati, o alla progressione nella ricchezza informativa di un sistema che nasce dall'anelito a comunicare proprio dell'Uomo. Il problema odierno non è più sapere dove cercare, ma come farlo. Si tratta di riuscire a trovare informazioni utili nel minor tempo possibile; informazioni sparpagliate in un oceano di dati non pertinenti, immerse e nascoste in un assordante *rumor bianco*.
L'estrazione di informazioni utili dal Web ma, soprattutto, la scoperta di nuove conoscenze partendo da dati grezzi, ha rappresentato fino a pochi anni fa una formidabile sfida alle tecnologie informatiche. Il testo scritto nasce con una struttura – la struttura semantica – differente ed alternativa rispetto ad una struttura "informatica". Il linguaggio umano si è del resto evoluto come strumento di comunicazione – orale – tra una persona che parla ed una che ascolta, in presenza di rumori ambientali. In quanto tale, presenta di conseguenza delle caratteristiche "cucite" sullo schema cerebrale e cognitivo dell'uomo, che è profondamente differente da quello di una macchina; qualunque linguaggio umano presenta, solo per fare alcuni esempi:

- concetti astratti
- infinite combinazioni di relazioni tra concetti. Le relazioni possono essere sottili e dipendenti dal contesto
- molteplicità di significati per lo stesso termine
- significati impliciti, decifrabili solo dal contesto
- complessità e flessibilità della lingua scritta e delle sue regole; flessibilità e regole non sono le stesse nei vari linguaggi

Quando parliamo di comprensione automatica del linguaggio naturale, parliamo, in pratica, di intelligenza artificiale ovvero di individuazione dei processi cognitivi e cerebrali umani e della loro formalizzazione ai fini della successiva definizione di regole computazionali. La scuola razionalista della linguistica (Chomski) ha indirettamente supportato questa strada affermando che una porzione significativa della conoscenza umana non deriva dai sensi ma è fissata alla nascita, per derivazione genetica. Seguendo questo approccio, le regole semantiche della lingua umana sono un dato individuabile, fondamentalmente statico, e di conseguenza imitabile.
La strada dell'intelligenza artificiale ha prodotto alcuni interessanti sistemi, che non hanno però superato lo stadio di "giocattoli" [1]. I principali problemi, che hanno ristretto le intelligenze artificiali all'ambito protetto dei laboratori, consistono nel numero e complessità delle regole da un lato, e nella flessibilità del linguaggio naturale dall'altro. Codificare a mano le regole è un compito gigantesco, che produce peraltro un sistema molto rigido e "pesante". Il linguaggio naturale, al contrario, è uno stru-

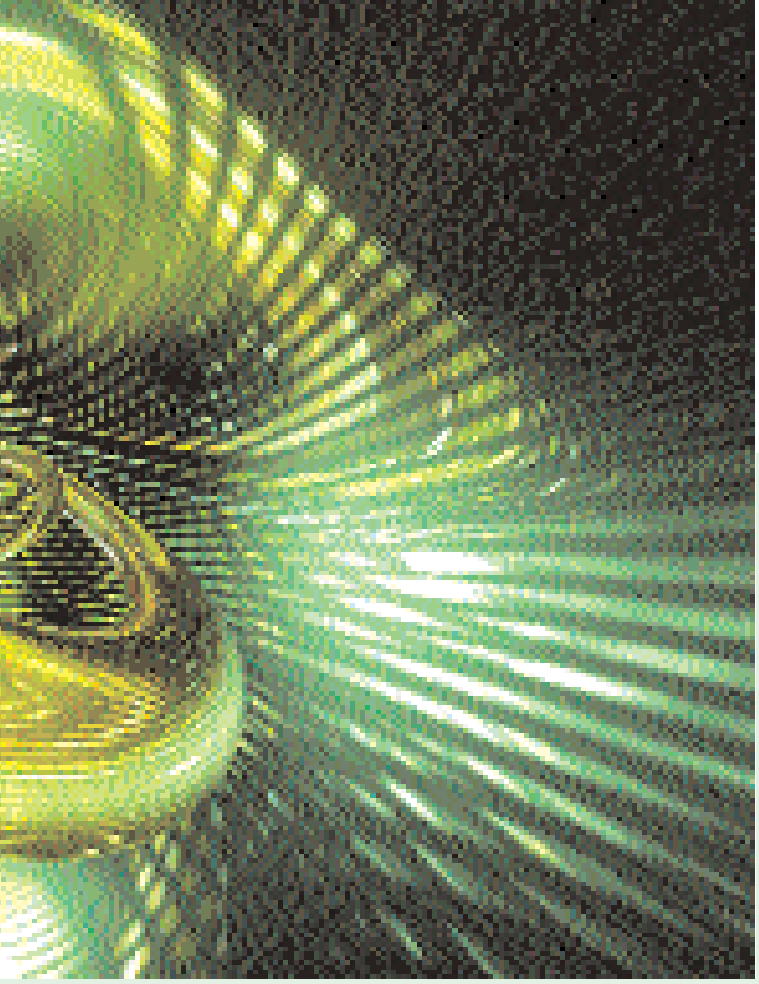




mento flessibile, che viene continuamente piegato e trasformato da chi lo usa.
Un nuovo e migliore approccio si è fatto strada negli anni '90; esso unisce una fase di preprocessing linguistica a tecniche di *data mining*[2]. Non a caso si definisce questo settore di indagini *text mining* composto di:

- un modello linguistico (tarato su lingue specifiche), che contiene le conoscenze grammaticali per scomporre le frasi nelle sue componenti di base (aggettivi, verbi, date)
- un dizionario generico a supporto del modello, eventualmente integrato da dizionari "settoriali" (medico, informatico, etc.)
- un "estrattore di relazioni" che consenta di associare tra loro i termini della frase (soggetto-verbo-oggetto; oggetto - proprietà di – oggetto; etc.)

Il passaggio attraverso la fase di preprocessing linguistico, in sintesi, permette di "destrutturare" il testo e "ristrutturarlo" secondo le necessità del data mining. A questo punto, il testo è pronto per essere trattato secondo collaudate tecniche di data mining. Possiamo menzionare, a titolo di esempio, la categorizzazione (assegnazione di un testo a un gruppo predefinito) e la clusterizzazione (separazione dei testi base alla presenza di elementi comuni, non definiti a priori); ma sono fattibili anche l'evidenziazione di testi che fuoriescano da un modello "standard" o il semplice riassunto di un documento.
L'uso sicuramente più proficuo del text mining concerne comunque la scoperta della conoscenza; ovvero non tanto l'estrazione di informazioni, magari difficili da trovare ma già esistenti, bensì l'individuazione, l'evidenziazione di fenomeni al quale non avremmo altrimenti pensato[3]. È questo del resto l'obiettivo di fondo del text mining: la scoperta della conoscenza[4].
Rispetto ad altre metodologie di estrazione dati dal testo scritto, quella linguistica + data mining è ad oggi la più potente e al tempo stesso la più promettente. Metodi puramente statistici, anch'essi impiegati, non riescono malgrado la notevole sofisticazione a risolvere totalmente l'ambiguità del testo scritto. Né sembrano prossimi significativi sviluppi nel campo dell'Intelligenza Artificiale. Al contrario, l'analisi linguistica applicata al text mining è ormai operativa. Sempre più realtà aziendali stanno traendo beneficio dalla scoperta di nuove conoscenze.
Prima di esse, e più di esse, hanno cominciato a muoversi gli Stati. Echelon, il sistema di intercettazione telefonico / internet gestito dalla NSA[5] può assorbire, pare, milioni di messaggi al minuto. L'FBI gestisce un sistema analogo. L'attenzione delle principali Agenzie di sicurezza statunitensi al problema non è casuale. La vulnerabilità delle democrazie occidentali al terrorismo quasi costringe la nostra civiltà a puntare sulla tecnologia. Il nemico sta usando le nostre risorse e la nostra ricchezza comunicativa per ferirci. Noi possiamo, dobbiamo, usare la stessa tecnologia per difenderci. Su questo fronte, sempre di competizione, il text mining sarà un elemento importante. E forse domani la nuova sfida non consisterà più nell'aumentare ulteriormente le sue potenzialità, quanto nel realizzare un compromesso tra l'efficienza degli strumenti per l'estrazione della conoscenza da un lato e le libertà democratiche dell'individuo dall'altro.
Perché per quanto solidi siano i nostri sistemi politici, davanti alla minaccia di un attacco dall'interno, di una guerra non dichiarata, gli Stati



in quanto collettività sociale hanno sempre subordinato i diritti del singolo alla sopravvivenza del gruppo. E le regole che valgono in tempo di pace non sono le stesse del tempo di guerra.

NOTE

[1] Nel '69 Terry Winograd, come tesi di dottorato al MIT, presentò un sistema capace di spostare oggetti geometrici su un video in risposta a comandi testuali da parte dell'utente, e capace di dare spiegazioni del suo comportamento; nel dialogo seguente U è l'umano, S è la macchina (da SHRDLU, l'elenco delle sei lettere più frequenti della lingua inglese)
U: PICK UP A BIG RED BLOCK
S: OK
U: GRASP THE PYRAMID
S: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN
U: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX
S: BY "IT" I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING
S: OK
U: WHAT DOES THE BOX CONTAIN?
S: THE BLUE PYRAMID AND THE BLUE BLOCK
U: WHAT IS THE PYRAMID SUPPORTED BY?
S: THE BOX
U: WILL YOU PLEASE STACK UP BOTH OF THE RED BLOCKS AND EITHER A GREEN CUBE OR A PYRAMID?
S: OK
.....
U: HAD YOU TOUCHED ANY PYRAMID?
S: YES, THE GREEN ONE
U: WHEN DID YOU PICK IT UP?
S: WHILE I WAS STACKING UP THE RED CUBE, LARGE RED BLOCK AND A LARGE-GREEN CUBE
U: WHY?
S: TO GET RID OF IT
U: WHY DID YOU DO THAT?
S: TO CLEAN OFF THE RED CUBE
U: WHY DID YOU CLEAR OFF THAT CUBE?
S: TO PUT IT ON A LARGE GREEN CUBE
U: WHY DID YOU DO THAT?
S: TO STACK UP THE RED CUBE, A LARGE RED BLOCK AND A LARGE GREEN CUBE
U: WHY DID YOU DO THAT?
S: BECAUSE YOU ASKED ME TO.

[2] Si definisce data mining il processo di estrazione della conoscenza – di nuova conoscenza – partendo dai dati.

[3] A cavallo tra gli anni '80 e '90 Swanson, negli Stati Uniti, estrasse mediante una procedura semiautomatizzata frammenti di articoli da testi medici sparsi; tali frammenti, isolati dal contesto e uniti tra loro, evidenziavano una forte correlazione tra alcuni tipi di emicranie e la carenza di magnesio; l'ipotesi andava provata sperimentalmente, ma l'elemento fondamentale è che scaturì naturalmente, senza intervento umano, da un processo di text mining.

[4] Knowledge Discovery.

[5] National Security Agency; è l'Agenzia statunitense che ha in gestione i più sofisticati sistemi di intelligence americani; per via della segretezza, è stata anche definita come No Such Agency (traducibile all'incirca come "non esiste questa agenzia"). ■