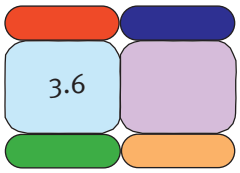
INTELLIGENZA ARTIFICIALE: I PRIMI 50 ANNI

Luigia Carlucci Aiello
Maurizio Dapor

Circa 50 anni fa, durante un famoso seminario al Dartmouth College, veniva ufficialmente introdotto il termine "Intelligenza Artificiale". Vi era allora un'atmosfera di euforia tecnologica indotta dall'avvento del computer, una macchina in grado di manipolare simboli, e l'intelligenza artificiale sembrava a portata di mano. Che cosa è successo dopo? L'articolo presenta alcuni momenti particolarmente significativi della storia dell'IA,le principali aree di ricerca e le prospettive applicative.



1. UN NOME FORSE TROPPO IMPEGNATIVO

Ogni disciplina scientifica è animata da sogni e motivata da grandi progetti. Grazie ad essi si procede, conseguendo risultati forse differenti rispetto a quelli immaginati, ma spesso utili e in grado di conferire all'intero programma un significato che sovente va al di là delle speranze e degli obiettivi originali. Accade anche di conseguire, talora, risultati inaspettati e originariamente imprevedibili.

Qual è il sogno dei ricercatori di intelligenza artificiale? Nella maggior parte dei casi si sogna, cosa piuttosto ambiziosa in verità, di realizzare quello che di certo appare come il più inaccessibile tra i progetti scientifici: capire i principi e i meccanismi del funzionamento della mente umana allo scopo di riprodurre l'intelligenza umana su una macchina.

L'espressione "*Intelligenza Artificiale*" (IA) descrive accuratamente questo sogno traducendo correttamente l'obiettivo finale. Ma, forse, se gravati da un nome meno impegnativo, i ricercatori di IA si sarebbero imbattuti in minori difficoltà lungo il loro cammino. In effetti, l'espressione "Intelligenza Artificiale" ha innescato paure irrazionali, alimentate da certa letteratura e cinematografia fantascientifiche. Anche il sarcasmo proveniente da alcuni ambienti antiscientifici non ha giovato. Forse il campo avrebbe incontrato meno ostilità se per esso fosse stata scelta la dizione britannica – derivata da A. Turing - di "*Intelligenza delle Macchine*" (IM), in quanto essa rammenta costantemente che, per quanto intelligenti, pur sempre di macchine si tratta. Forse, invece, un forte dibattito era ed è inevitabile in quanto costantemente l'IA (o IM) mette in ballo quella che è ritenuta la più esclusiva prerogativa degli esseri umani: l'intelligenza.

Va aggiunto che la scelta di un'espressione tanto forte come "Intelligenza Artificiale" ha di certo generato aspettative eccessive, in particolare considerando le limitazioni della tecnologia con la quale gli artefatti intelligenti andavano via via realizzati. D'altronde, qualcuno ha giustamente osservato che ogni volta che l'IA raggiunge un nuovo

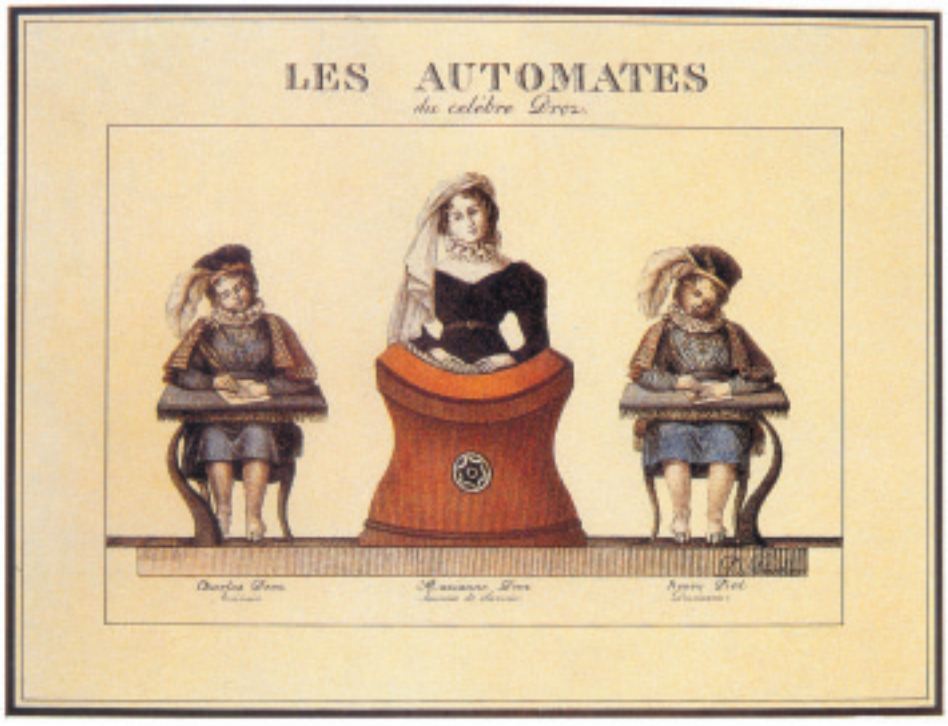
traguardo questo non è più annoverato tra i suoi risultati ma si trasforma in un prodotto dell'informatica tradizionale: un destino ingrato per una disciplina che, in realtà, ha conseguito molti successi e ha facilitato la vita di ognuno di noi. Una spiegazione di tale circostanza risiede nella constatazione che l'IA parte da problemi scientifici che vengono affrontati, risolti e, quindi, ingegnerizzati con gli strumenti dell'informatica. Un esempio abbastanza noto è costituito dai sintetizzatori vocali con cui si interagisce telefonicamente quando si chiama un *call center*, per esempio, se si deve segnalare i valori del contatore alla società fornitrice di elettricità. Oggi si considerano tali sintetizzatori come un utile prodotto dell'informatica. In passato, tuttavia, quello dei sintetizzatori vocali e, più in generale, della elaborazione del segnale vocale, era un problema di interazione uomo-macchina a ragione collocato nella ricerca in IA.
Il programma di fondo dell'IA, quello di costruire macchine in grado di comportarsi come gli esseri umani, non è mai stato abbandonato e il test di Turing, basato sull'idea secondo cui se il comportamento di una macchina è indistinguibile da quello di un essere umano allora quella macchina può essere considerata intelligente, viene oggi quotidianamente superato in molte circostanze, almeno in determinati e ben precisi ambiti cognitivi. A semplificare (o complicare?) le cose oggi interviene Internet. Le informazioni, grazie a Internet, si trasferiscono da un capo all'altro del pianeta in tempo reale: quando si interagisce con uno sconosciuto giocando a scacchi in rete, non si è in grado di stabilire se l'interlocutore sia un programma *software* o un essere umano. Così il test di Turing, nel particolare ambito cognitivo costituito dal gioco degli scacchi, appare perfettamente superato. In effetti, rete o no, per quanto riguarda il gioco degli scacchi ci sono programmi che riescono oggi a battere i campioni del mondo. Nel 1997, Deep Blue, un sistema sviluppato dalla IBM, riuscì a sconfiggere l'allora campione del mondo di scacchi, G. Kasparov, in un torneo regolare di sei incontri. Si trattò di un successo clamoroso dell'IA. L'osservazione dello stile di gioco di Deep Blue già aveva fatto constatare a Kasparov, l'anno precedente, quanto segue: "Posso percepire, ne sento persino l'odore, un nuovo tipo di intelligenza dall'altra parte del tavolo". Una dichiarazione che farebbe credere di essere andati ben oltre il semplice superamento del test di Turing.
Nella situazione specifica, la supremazia dei calcolatori riguarda una data capacità cognitiva: gli scacchi. Ma è chiaro che un computer esperto in quel gioco incontrerebbe difficoltà insormontabili se si pretendesse che si cimentasse in un compito per noi assai semplice come quello di friggere un uovo. O se si tentasse di sostenere con lui un'amabile conversazione. Per limitarsi alle conversazioni con i computer condotte in linguaggio naturale, le difficoltà sono effettivamente notevoli. Si può ricordare in proposito un famoso programma degli anni Sessanta, ELIZA (scritto nel 1965 da J. Weizenbaum), che simulava una seduta dallo psicanalista. Ecco un tipico dialogo tra il programma e una ragazza:

UTENTE: Gli uomini sono tutti uguali.
ELIZA: In che senso?
UTENTE: Ci rompono sempre le scatole per una cosa o per l'altra.
ELIZA: Ti viene in mente un esempio specifico?
UTENTE: Il mio ragazzo mi ha fatto venire qui.
ELIZA: Il tuo ragazzo ti ha fatto venire qui?
UTENTE: Dice che sono sempre depressa.
ELIZA: Mi spiace che tu sia depressa.

E così via, senza fine. La strategia utilizzata da ELIZA per rispondere all'utente è talmente ingenua da non richiedere ulteriori commenti: anche se, va precisato a onor del vero, molti di noi sono inclini a sospettare che talora gli psicanalisti veri si comportino allo stesso modo. A parte gli scherzi, abbiamo certo sorriso con condiscendenza a fronte di dialoghi come questo tra un computer e un utente umano. Ma dalla metà degli anni Sessanta ad oggi molto lavoro è stato fatto e si sono conseguiti risultati assai più interessanti di questo. ELIZA riveste, ormai, un interesse puramente storico, perché oggi l'elaborazione del linguaggio naturale scritto e parlato è diventata un'area matura e con molteplici applicazioni.

2. I PRIMI ANNI DI STORIA: IL PERIODO EROICO

È bene sapere che l'uomo ha tentato di costruire artificialmente forme di vita intelligente assai più spesso di quanto comunemente si creda: già nel XVIII secolo Jacques de Vaucanson costruì un'anitra artificiale che non era solo in grado di nuotare ma anche di ingoiare chicchi di grano (ovviamente non necessari per il suo sostentamento). L'orologiaio svizzero Pierre Jacquet-Droz fabbricò alcuni automi che sono oggi conservati nel Museo di Arti e Storia di Neuchâtel tra cui un piccolo scrivano, un bravo disegnatore, nonché una giovane e promettente musicista i cui movimenti sono controllati da una serie di codici memorizzati su alcuni dischi di metallo. È in grado, questa sorprendente ragazzina, di eseguire cinque melodie differenti utilizzando la tastiera di un vero organo a canne: durante l'esecuzione respira e simula una certa emozione; chiude la sua esibizione con una gentile riverenza e china il capo a seguito degli applausi del pubblico.
Il grande filosofo e matematico G. W. Leibniz era affascinato da un sogno: un giorno, probabilmente non lontano, un *calculus ratiocinator* sarebbe stato in grado di dirimere qualunque controversia intellettuale: si sarebbe trattato di un complesso sistema di meccanismi in grado di affrontare ogni problema e di risolverlo. L'idea di Leibniz godette, nei secoli a seguire, di una notevole fortuna. Attraverso i primi meccanismi in grado di effettuare le operazioni più semplici dell'aritmetica, l'idea di Leibniz condusse alla macchina di J. Von Neumann che, costituisce la base dei moderni calcolatori.
Nel periodo che va dal 1930 al 1955 si assiste a un fiorire di idee e teorie che costituiranno le basi della odierna IA. Sono di questo periodo i lavori sulla logica, sulla calcolabilità (A. Turing e A. Church, in particolare) che sono stati di fondamento alla progettazione dei primi computer e base all'approccio simbolico all'IA. È di questo periodo (1943) il lavoro fondamentale di W. S. McCulloch e W. H. Pitts in cui viene proposto un modello di neurone che ha fatto poi da base a tutta la ricerca sulle reti neurali. Nel 1949, D. O. Hebb propone un meccanismo per aggiornare la "forza" delle connessione tra neuroni che è ancora oggi in uso ed è noto come "apprendimento hebbiano".



È del 1951 la costruzione da parte di M. Minski (insieme a D. Edmonds) del primo computer neurale.
Nel 1950, A. Turing pubblica un articolo, "Computing Machinery and Intelligence", in cui introduce il famoso test che poi prenderà il suo nome (test di Turing), l'apprendimento automatico, gli algoritmi genetici e l'apprendimento per rinforzo. Nell'articolo previde molti degli ostacoli che questi studi avrebbero incontrato negli anni a causa della forte interferenza degli artefatti da essi prodotti con le funzioni cognitive degli esseri umani.
La nascita ufficiale della disciplina è legata a J. McCarthy, che - allora docente al Dartmouth College - organizzò un seminario di due mesi nell'estate del 1956, invitando ricercatori interessati alla teoria degli automi, alle reti neurali e allo studio dell'intelligenza (M. Minski, T. More, A. Newell, N. Rochester, A. Samuel, C. Shannon, O. Selfridge, e H. Simon). I ricercatori presenti al seminario avevano interessi che andavano dallo sviluppo di sistemi di ragionamento automatico (A. Newell e H. Simon) a giochi quali la dama (A. Samuel).
Il problema dell'IA non fu risolto in due mesi, come pare fosse tra gli obiettivi del seminario, ma i protagonisti del campo si conobbero e gettarono le basi per una ricerca che, insieme a quella dei loro studenti, forgiò la disciplina e la caratterizzò per i successivi venti anni.
Dopo il seminario si succedettero molti risultati scientifici e riflessioni filosofiche relativi alla IA: nel 1957, A. Newell e H. Simon realizzarono il *General Problem Solver* (GPS), mentre N. Chomsky pubblicava "Le strutture della sintassi", uno studio della linguistica che diventerà un caposaldo nella progettazione della sintassi dei linguaggi di programmazione.
I primi anni della ricerca in IA, fino alla fine degli anni Sessanta, sono stati caratterizzati da grande entusiasmo ma anche da grande ingenuità. Essi hanno visto un fiorire di "successi", nel senso che sono stati sviluppati e proposti in letteratura molti programmi e sistemi che facevano cose impensabili per l'informatica tradizionale dell'epoca: i nuovi programmi di IA facevano inferenze (come nel GPS di A. Newell e H. Simon), giocavano a dama (A. Samuel)[1], risolvevano problemi di analogie (T. G. Evans) ed elementari problemi di geometria (H. Gelernter), "capivano" semplici frasi in linguaggio naturale (P. Winston) e risolvevano problemi di integrazione simbolica (J. R. Slagle). Si tratta degli anni in cui J. McCarhy ha progettato il LISP[2] (*List Processing*), poi diventato e rimasto linguaggio di elezione per l'IA per i successivi decenni, e l'Advice Taker, un programma che permetteva di rappresentare conoscenza e di fare semplici inferenze. Sono gli anni in cui J. A. Robinson ha introdotto il "principio di risoluzione", che sarà poi alla base della programmazione logica. Sono gli anni in cui all'SRI International viene realizzato Shakey, il primo "robot cognitivo" per il quale sono stati sviluppati sistemi di pianificazione automatica (STRIPS) e algoritmi euristici per la soluzione automatica di problemi tuttora di largo uso (l'algoritmo A*, introdotto da P. E. Hart, N. J. Nilsson e B. Raphael). Ma — come il nome stesso denunciava (da *shaky* = traballante) — Shakey era un sistema tutt'altro che robusto.
In quegli anni sono stati introdotti gli esempi dei "micromondi", diventati poi noti come *toy problem*. Si tratta anche degli anni in cui, alle dichiarazioni entusiastiche e lungimiranti sugli obiettivi grandiosi della disciplina IA, corrispondevano programmi che risolvevano quiz del calibro di quelli dei giornalini di enigmistica: non c'era verso di poterli utilizzare per risolvere problemi di dimensioni realistiche perché spesso le soluzioni proposte erano del tipo — parole di J.McCarthy — "look ma, no hands"[3]

[1] Nel 1962, A. Samuel fece giocare un suo programma contro l'ex campione di dama del Connecticut. Il programma vinse l'incontro.

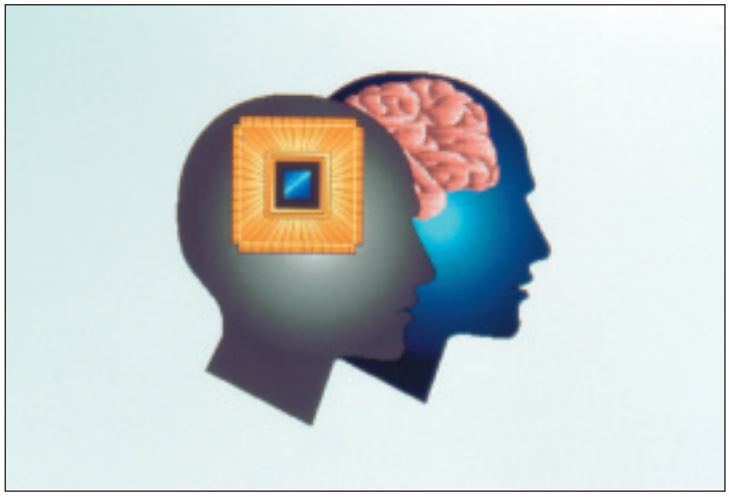
[2] Il LISP nacque come un'estensione del linguaggio FORTRAN. Mentre il FORTRAN era stato progettato per il calcolo numerico, il LISP era dedicato all'elaborazione di informazione simbolica (cioè non numerica) e alla rappresentazione e manipolazione di strutture di dati dinamiche.

[3] "Guarda mamma, senza mani!", gridano i bambini alla mamma per attrarne l'attenzione mentre, pedalando, tolgono le mani dal manubrio della bicicletta.

È di quegli anni lo sviluppo da parte di J. Weizenbaum del programma ELIZA, cui si è accennato nel paragrafo precedente. ELIZA, che si comportava come uno psicoterapeuta, è diventato molto famoso e per anni è rimasto il programma esemplare di ciò che l'IA può fare. Forse gode di minor fama il fatto che Wizenbaum, che della sua creatura ELIZA ben conosceva i limiti, turbato da tanto successo secondo lui immotivato, ha abbandonato la ricerca in IA.
Mentre da una parte fiorivano gli sforzi per sviluppare programmi capaci di esibire comportamenti intelligenti, dall'altra si sviluppava il dibattito sulla fattibilità o meno dell'intelligenza artificiale. Nel 1960, J.R. Lucas asserì che il teorema di incompletezza di Gödel impedirebbe che una macchina possa risolvere problemi che la mente umana, invece, sa come affrontare. Credette in tal modo di poter trarre la conclusione che la mente non è una macchina. Nel 1979, D. Hofstadter pubblicò il celeberrimo "Gödel, Escher e Bach" in cui, tra l'altro, si contrappose alle tesi sostenute da Lucas mostrando che il teorema di incompletezza permetterebbe di prefigurare sistemi autoconsapevoli.
Nonostante i successi conseguiti dalle prime reti neurali artificiali, cui loro stessi avevano contribuito, nel 1969, M. Minsky e S. Papert ne evidenziarono i limiti nel libro "Perceptrons": data l'autorevolezza degli autori, la pubblicazione del libro provocò una consistente contrazione dei finanziamenti per la ricerca nel campo delle reti neurali artificiali. La conseguente riduzione di interesse si protrarrà sino alla prima metà degli anni Ottanta, allorché la ricerca riprenderà vigore anche grazie a un celebre articolo del 1982 di J. J. Hopfield.

3. L'INDUSTRIA DEI SISTEMI ESPERTI E L'INVERNO DELL'INTELLIGENZA ARTIFICIALE

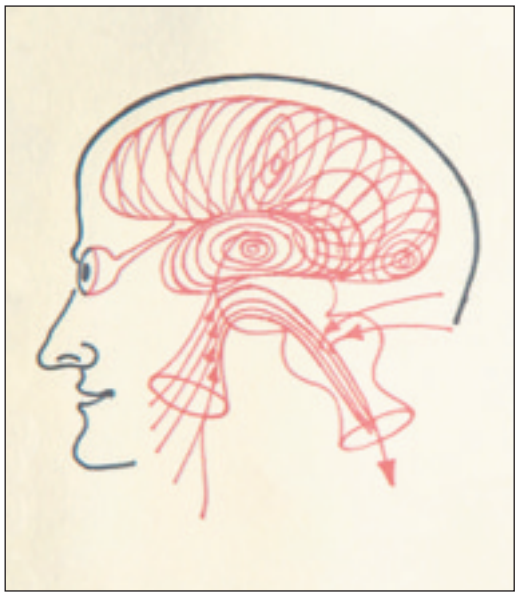
Il sogno originale dei ricercatori di IA era quello di individuare un formalismo per rappresentare conoscenza e un apparato deduttivo con poche regole di tipo del tutto generale che permettessero di riprodurre i meccanismi del ragionamento umano. È in questo senso la proposta del GPS di A. Newell e H. Simon.



Ma presto questo approccio generalista rivelò i suoi limiti, e non solo per la intrinseca limitatezza della potenza di calcolo a disposizione in quegli anni: la costruzione di una macchina che fosse intelligente in ogni campo apparve chiaramente come un progetto di difficile (se non impossibile) realizzazione. Già dalla seconda metà degli anni Sessanta nella comunità di ricerca in IA si faceva strada una preoccupazione: i programmi sviluppati si limitavano a semplici manipolazioni simboliche, non avevano e non utilizzavano, di fatto, nessuna cognizione di quello che stavano facendo e non avevano alcun modo per confrontarsi con la complessità (in genere esponenziale) dei problemi che si incontravano nelle situazioni pratiche. È di questi anni un vero ridimensionamento degli entusiasmi: anziché perseguire l'obiettivo di costruire un sistema capace di intelligenza generale e assoluta, che si manifesterà poi nella soluzione di qualunque problema, ci si accontenta piuttosto di costruire sistemi capaci di risolvere problemi in domini limitati; la conoscenza su un settore limitato viene rappresentata nel sistema e gli permette di comportarsi da esperto nel dominio in questione. In tale periodo, si assiste alla nascita (e al successo) dei "sistemi basati sulla conoscenza" o "sistemi esperti".
"Se sto male non vado al dipartimento di matematica, dove so che ragionano molto bene, ma a quello di medicina, dove so che possiedono e sanno utilizzare conoscenza medica": era lo slogan lanciato da E. Feigenbaum, a sottolineare l'importanza di costruire sistemi che incorporassero conoscenza del dominio

e che poi ne facessero effettivo uso al fine di individuare automaticamente in modo efficiente ed efficace una soluzione per i problemi posti al sistema, tipicamente problemi diagnostici. Più ragionevole e promettente appariva, quindi, l'approccio che, anziché pretendere di catturare l'intelligenza in regole generali in grado di risolvere qualunque problema, si focalizzasse piuttosto sulla capacità di affrontare singole problematiche mediante tutto ciò che si conosceva sull'argomento: utilizzando regole, conoscenze strutturali, principi di ragionamento, trucchi del mestiere, sarebbe certo stato possibile ottenere risultati molto pregnanti. Con i sistemi esperti si mostrò che programmi in grado di conservare e organizzare le conoscenze concernenti ambiti ben precisi e ristretti, e di dedurre quanto non esplicitamente dichiarato, erano in grado di individuare le soluzioni di problemi complessi fino a quel momento non aggredibili con tecnologie informatiche.

In quegli anni sono stati sviluppati sistemi esperti per numerose applicazioni: dalla biochimica, alla prospezione geologica, alla diagnosi medica, alla progettazione di configurazioni complesse, alla finanza. Molti di questi sistemi sono stati usati con notevole successo da parte dei loro realizzatori: tra i più noti si citano R1, sistema sviluppato al CMU per la Digital, che permetteva di automatizzare compiti che avrebbero richiesto molto personale qualificato. Va poi menzionato PROSPECTOR, un sistema esperto in prospezioni geologiche sviluppato all'SRI International che portò, nel 1979, la ricerca in IA per la prima volta sulla prima pagina sul New York Times: un "computer program", segnalava l'autorevole quotidiano, era riuscito a localizzare un giacimento di molibdeno nel nord ovest degli Stati Uniti. Si trattava di un giacimento che i geologi ipotizzavano esistere, ma in 20 anni non erano riusciti a localizzare.



Oggi è possibile asserire che i sistemi esperti costituiscono un indiscutibile successo dell'IA, anche se il nome è stato abbandonato in favore di nomi più "neutri" quale, ad esempio, "sistemi di supporto alle decisioni".

L'entusiasmo per i successi dei sistemi esperti ha portato al fiorire di un'industria di IA, il cui fallimento ha poi determinato la peggiore stagione nell'intera storia della disciplina, definita "l'inverno dell'IA". I motivi di tale entusiasmo sono evidenti: problemi fino a poco tempo prima giudicati irrisolvibili con un sistema di calcolo trovavano adesso soluzione. Era diventato possibile effettuare diagnosi mediche di livello paragonabile a quelle dei medici più qualificati, potendo, inoltre, "spiegare" il ragionamento seguito dal sistema per arrivare a una data conclusione piuttosto che a un'altra. Questo induceva a pensare che — con gli strumenti adatti — chiunque avrebbe potuto in poco tempo costruire il suo sistema esperto. Inoltre, le interfacce "amichevoli" avrebbero reso questi sistemi utilizzabili da chiunque, rendendo inutile la presenza di esperti. Entrambe le aspettative erano, naturalmente, esagerate. Altri motivi di insuccesso, interessanti da analizzare, riguardano la dimensione troppo piccola dei domini di esperienza dei sistemi esperti; la brusca diminuzione di affidabilità qualora un problema non sia "centrato" nel dominio di esperienza del sistema; l'incapacità da parte di un sistema di capire se una domanda rientrava o no nel suo dominio di esperienza.

A fianco dei sistemi esperti, l'industria in quegli anni offriva "ambienti" (talvolta detti "shell") per il loro sviluppo o per "l'ingegneria della conoscenza", ostinandosi a perseguire l'idea che si trattava di compiti affatto differenti dall'ingegneria del software e che richiedevano, pertanto, hardware e softwa-

re specializzati. L'alto prezzo e la bassa affidabilità dei sistemi (hardware e software) rapidamente immessi sul mercato per rispondere alla spasmodica domanda ne decretò il fallimento. Anche il lancio e il successivo fallimento del progetto della cosiddetta "quinta generazione" giapponese, programma di ricerca per la creazione di hardware e software per lo sviluppo di sistemi intelligenti di cui non resta oggi praticamente traccia, fu corresponsabile del sopraggiungere dell'inverno dell'IA.
Di questo periodo è rimasta una connotazione estremamente riduttiva dell'IA: agli occhi di molti infatti l'IA è diventata ed è rimasta la ricerca per la costruzione di sistemi, magari sofisticati, basati su regole di tipo "se...allora". In realtà, la ricerca sui "sistemi a regole" si è esaurita negli anni Settanta, per sopravvivere negli "shell per sistemi esperti" dei primi anni Ottanta.

4. L'INTELLIGENZA ARTIFICIALE ESCE DALL'INVERNO

Si può far risalire alla seconda metà degli anni Ottanta il risveglio dell'IA dal suo inverno. Innanzitutto, la ricerca in IA che, fino a quel momento, aveva cercato di differenziarsi da quella delle discipline limitrofe per trovare una sua identità, ha cominciato a non disdegnarne più gli avanzamenti e i risultati. Ci si riferisce qui agli avanzamenti in matematica, informatica, teoria del controllo, statistica. A partire dalla seconda metà degli anni Ottanta, l'IA ha cercato una maggiore integrazione con queste discipline.
Ormai superata l'era pionieristica, la comunità scientifica ha cominciato a non accontentarsi più di risultati imposti sulla base di *handwaving*, ma a pretendere un maggiore rigore: i risultati dovevano essere fondati su solide teorie (meglio se non inventate: piuttosto modifiche, o estensioni, di teorie esistenti) oppure evinti da sperimentazioni degne del nome.
Risale a questi anni la "riscoperta" delle reti neurali. Come accennato in precedenza, il merito va in particolare ai risultati di J. J. Hopfield (e poi di D. E. Rumelhart e G. E. Hinton): le nuove reti neurali hanno eliminato le limitazioni espressive delle reti proposte in precedenza, allargandone enormemente il campo di applicazione. La ricerca sulle reti neurali ha recuperato risultati di decenni della matematica e della fisica, così come l'adozione degli HMM (*Hidden Markov Model*) ha permesso di portare, nel campo della comprensione del linguaggio parlato, tecniche sviluppate nella ricerca matematica e che hanno fornito al riconoscimento del parlato una qualità che lo ha reso utilizzabile in robuste applicazioni commerciali (per esempio nei sistemi telefonici). Analoga considerazione vale per il calcolo delle probabilità, per lungo tempo disdegnato dai ricercatori di IA (qualcuno certo ricorderà alcuni sistemi esperti di "prima generazione" e del tentativo che essi facevano di re-inventare un calcolo delle probabilità addomesticato, poiché il calcolo delle probabilità non era considerato utilizzabile nel contesto di loro interesse). Sono state proposte le "reti bayesiane", e sono ora molto utilizzate per rappresentare e ragionare in modo rigoroso con conoscenza incerta sfruttando il teorema di Bayes, caposaldo del calcolo delle probabilità.
Nel 1991, il filosofo D. Dennett dà alle stampe il libro *Consciousness Explained*, nel quale sostiene che l'auto-consapevolezza umana non è niente di più che l'effetto di processi biochimici. La coscienza umana, vale a dire la frontiera che, secondo molti filosofi, separa il pensiero umano dal calcolo meccanico, non è, per Dennett, che il prodotto di macchine seriali "implementate in modo inefficiente sull'hardware parallelo fornitoci dall'evoluzione".
Forse perché la soluzione dei vari sottoproblemi dell'IA ha cominciato a dare i suoi frutti, a partire dalla seconda metà degli anni Novanta del novecento i ricercatori si concentrano di nuovo sul progetto dell'"agente intelligente" come entità: fiorisce, quindi, la ricerca sugli agenti software intelligenti e degli agenti intelligenti "incorporati" in un sistema fisico, cioè i robot dotati di cognizione.
Gli agenti intelligenti sono "situati" nel mondo fisico con il quale interagiscono: la percezione dà all'agente informazione sullo stato del mondo e le sue azioni producono cambiamenti in questo mondo. Nel caso dei robot, il mondo fisico è un ambiente solitamente "non strutturato", cioè non completamente descrivibile a priori, quindi non completa-

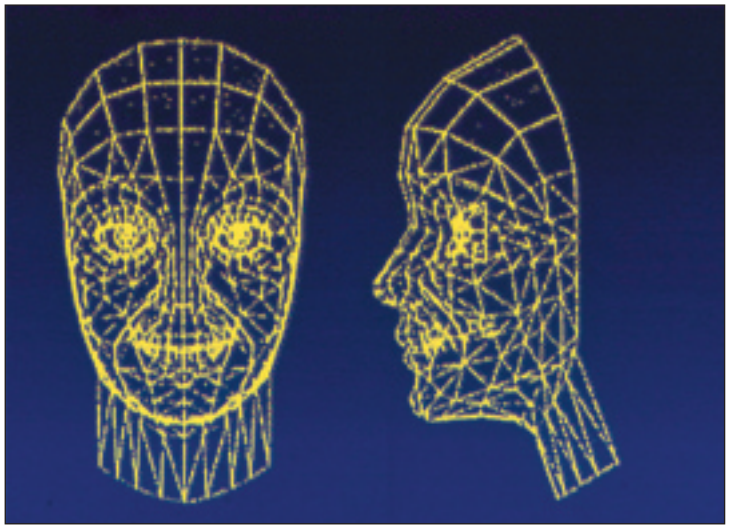
mente prevedibile (mentre, per esempio, è strutturato l'ambiente costituito da una catena di montaggio in una applicazione di automazione industriale). Nel caso degli agenti intelligenti software, l'ambiente fisico per eccellenza è il web e, in questo caso, si parla anche di "softbot".
Trattare in modo sufficientemente dettagliato tutti i settori in cui si articola la ricerca in intelligenza artificiale è al di fuori della portata di questa presentazione. Un quadro sintetico viene presentato a pagina 18, alcuni dei settori di ricerca in IA sono stati già portati all'attenzione dei lettori di Mondo Digitale in articoli recenti (elencati in Bibliografia tra le letture consigliate). Qui di seguito si accennerà soltanto a due filoni che hanno caratterizzato il campo e che sono stati considerati come antitetici e alternativi per la realizzazione di un artefatto intelligente: l'approccio logico e quello connessionista (cioè basato su reti neurali artificiali) alla rappresentazione della conoscenza. Il primo è anche comunemente detto approccio "simbolico", e — per contrasto — il secondo viene detto approccio "sub-simbolico".

5. RAPPRESENTAZIONE DELLA CONOSCENZA: L'APPROCCIO SIMBOLICO

Già si è accennato a J. McCarthy il quale, già nel 1958, sviluppò il linguaggio di programmazione per eccellenza dell'IA, il celeberrimo LISP, e l'Advice Taker. Già in quegli anni egli delineava quello che è diventato l'approccio simbolico (o logicista) alla IA. Esso consisteva nel rappresentare — utilizzando un opportuno linguaggio formale — in una "base di conoscenza" (BC) tutto quello che un agente conosce su un dato mondo (rappresentazione esplicita della conoscenza) e da questa base di conoscenza trarre le conclusioni necessarie a far agire l'agente in modo "intelligente". I meccanismi inferenziali permettono di rendere esplicita anche conoscenza che nella BC non è esplicitamente rappresentata, simulando ragionamenti del tipo del classico sillogismo aristotelico per cui se in BC ho che "Socrate è un uomo" e che "Tutti gli uomini sono mortali", posso dedurre sì che Socrate è un uomo (direttamente contenuto in BC) ma anche che Socrate è mortale (implicitamente contenuto in BC e dedotto con la regola di inferenza del *modus ponens*). Questo tipo di rappresentazione e uso della conoscenza ricorda molto da vicino le teorie assiomatiche in matematica: la conoscenza su un dato pezzo di mondo (per esempio, i gruppi) viene rappresentata in forma di assiomi (gli assiomi sui gruppi: esistenza di una operazione associativa, esistenza di una identità, di un inverso) da cui si traggono conclusioni (i teoremi, cioè le proprietà che valgono nei gruppi).
La rappresentazione della conoscenza in IA si avvicina molto alla logica matematica, ma al tempo stesso ne differisce grandemente: mentre, infatti, in matematica la conoscenza è di tipo statico (per esempio, se una struttura matematica è un gruppo commutativo lo è per sempre, e le sue proprietà per sempre rimarranno vere), in IA si vuole rappresentare una realtà che cambia dinamicamente, in cui quindi proprietà possono passare dal vero al falso, semplicemente perché è cambiato il quadro di riferimento e sono state eseguite azioni che hanno falsificato cose che prima erano vere (o viceversa).
La specificità della conoscenza da trattare nei sistemi di IA ha portato McCarthy a proporre lo sviluppo di una "teoria della conoscenza" cioè lo sviluppo di una nuova logica (o meglio, di nuove logiche) che permettesse di descrivere rappresentazioni del mondo e dei suoi cambiamenti, e che giocasse un ruolo analogo a quello svolto dal calcolo nella fisica e dalla logica classica nella matematica. Come già detto, infatti, la logica matematica si occupa di enti teorici indipendenti dal tempo. L'IA, invece, non può permettersi di trascurare la conoscenza e le credenze che un agente ha su un dato mondo (che possono differire da ciò che in quel mondo è vero, semplicemente perché l'agente ha una conoscenza parziale o errata), le azioni, gli effetti delle azioni, quindi i cambiamenti. In altre parole, le logiche per l'intelligenza artificiale devono permettere di gestire il tempo, lo spazio, l'evoluzione dei contesti di riferimento, le variazioni dell'ambiente e la conoscenza che un agente ha sul mondo, i cambiamenti ad essa apportati mediante percezione ecc..
Lo studio di logiche (dette anche formalismi) per la rappresentazione della conoscenza è

stato ed è molto attivo e si concentra su due aspetti: da una parte la forma sintattica che le formule debbono avere affinché la conoscenza che esse "incorporano" sia riconoscibile e utilizzabile dal sistema, dall'altra il meccanismo inferenziale che dalla conoscenza espressa in questo formalismo permette di trarre conclusioni. Mentre del formalismo si vuole che sia espressivo, conciso e facile da usare, dell'apparato deduttivo si vuole che permetta solo di trarre conclusioni valide (correttezza) e possibilmente sia in grado di trarle tutte (completezza). Trattandosi di rappresentazioni di realtà complesse è subito evidente che ci si scontra con un grosso problema di complessità computazionale: complessità spaziale, cioè lo spazio di memoria occupato dalla BC e lo spazio di memoria necessario al meccanismo inferenziale per fare le sue deduzioni; complessità temporale, cioè il tempo necessario a raggiungere le conclusioni. Mentre ci si può accontentare di un apparato inferenziale incompleto, se questo favorisce un incremento dell'efficienza, certo non si vuole per un agente intelligente un apparato inferenziale non corretto. Proprio per incrementare l'efficienza delle inferenze, una tecnica che è attualmente oggetto di ricerca è quella di operare opportune trasformazioni della BC, dette compilazioni. Queste possono essere eseguite *off-line*, ed essere, quindi, molto utili se su di esse si può scaricare la complessità del problema e semplificare di conseguenza l'elaborazione da effettuare *on-line* quando cioè all'agente è effettivamente richiesto di agire e/o rispondere a specifiche domande.

Un problema cruciale in rappresentazione della conoscenza in IA è stato (ed è) quello della rappresentazione della conoscenza incerta. Ci si trova di fronte a due tipi di incertezza: una frase di cui non si sa se è vera o falsa (per esempio, "il Presidente Bush in questo momento dorme") o una frase di cui non si sa stabilire con esattezza la verità, semplicemente perché non ha senso (per esempio, "Mario è giovane"). Nel primo caso si ha bisogno di formalismi che permettano di trarre conclusioni anche se non si ha una conoscenza completa della situazione (per esempio, è possibile trarre tutte le conclusioni che non riguardino lo stato di vigilanza di



Bush), e che permettano di aggiornare le nostre conclusioni in situazioni in cui la conoscenza viene aggiornata (per esempio, un atto di percezione può aggiornare la BC con un nuovo fatto che dice che Bush, prima sveglio, ora dorme). Per trattare questo tipo di conoscenza sono state fatte molte proposte. La più famosa forse è la "logica dei default" proposta da R. Reiter nel 1980. Questa permette di trarre conclusioni in situazioni in cui si sa che una proprietà vale generalmente, ma non necessariamente per tutti gli elementi di una certa classe. La logica dei default non richiede nessuna modifica alla forma delle formule in BC, ma consiste nel modificare le regole di inferenza facendole diventare regole "di default": ogni volta che si sa che per gli oggetti di tipo A tipicamente vale la proprietà B, si introduce una regola del tipo: "Se vale A e in assenza di informazione che me lo impedisca[4], concludo B, allora B prende il nome di "*conclusione di default*". L'esempio più noto di ragionamento default è quello per cui "se qualcosa è un uccellino, in assenza di informazione che lo neghi, possiamo concludere che esso vola". In base a questa regola di default si può concludere che l'uccellino Titti vola in tutti i casi in cui non si sappia che ha un'ala rotta, che se lo è mangiato Gatto Silvestro ecc.. Ma è evidente che questa conclusione può essere smentita non appena si ac-

[4] In questo caso "impedire" vuol dire essere contraddittorio con B.

quisisca maggiore conoscenza sul mondo, o il mondo abbia subito un'evoluzione che invalida la proprietà dedotta per default.
Nel secondo caso, si ha bisogno di rappresentare verità "sfumate". Nel mondo reale oltre a fatti che sono veri o falsi in modo binario, come in matematica, ci sono fatti (molti di più?) la cui verità è parziale, o da valutare relativamente a un dato contesto. C'è tutta una gamma di "valori di verità" che vanno dal vero al falso, quali ad esempio quelli che servono a descrivere la proprietà "essere alto", "essere giovane" ecc.. Per rappresentare questo tipo di conoscenza "incerta" sono state fatte varie proposte, la più nota delle quali si deve a L. Zadeh che, già dagli anni Sessanta, ha introdotto la logica *fuzzy*, o sfumata. Studiata e sviluppata, a partire dagli anni Novanta ha trovato ampie applicazioni in robotica e nello sviluppo di apparati fisici complessi.

6. L'APPROCCIO SUB-SIMBOLICO: LE RETI NEURALI ARTIFICIALI

A partire dalla seconda metà del secolo scorso, attraverso l'introduzione di reti di unità logiche elementari (drastiche semplificazioni del neurone biologico proposte dal neurofisiologo W. S. McCulloch e dal logico W. H. Pitts nel 1943), si è tentato di realizzare artefatti intelligenti aggirando l'ostacolo rappresentato dalla necessità che un programmatore umano ne stabilisse a priori il comportamento, pianificandolo nella stesura dei suoi programmi.
L'architettura parallela delle reti neurali artificiali si fonda su un grande numero di unità elementari connesse tra loro su cui si distribuisce l'apprendimento. Esse hanno la capacità di apprendere da esempi senza alcun bisogno di un meccanismo che ne determini a priori il comportamento. Eseguono i loro compiti in modi completamente diversi da quelli delle macchine di Von Neumann: stabilendo autonomamente le rappresentazioni interne e modificandole sulla base della presentazione ripetuta di esempi, esse apprendono dai dati d'esperienza. Addestrare una rete neurale artificiale corrisponde a effettuare un certo numero di cicli di apprendimento grazie ai quali la rete genera una rappresentazione interna del problema. Una volta che, ad esempio, ha imparato ad associare tra loro alcune immagini, la rete effettuerà le associazioni corrette anche se le immagini in ingresso sono parzialmente differenti da quelle utilizzate nelle fasi di apprendimento.
Il neurone di McCulloch e Pitts è un'idealizzazione assai semplificata del neurone biologico. Si tratta, in sostanza, di un'unità logica in grado di fornire un'uscita binaria (zero oppure uno) in base al risultato di un semplice calcolo effettuato sui valori assunti da un certo numero di dati che si trovano sui suoi canali d'ingresso. A ognuno dei canali di ingresso viene assegnato un valore numerico, o peso. L'unità logica è caratterizzata da un valore numerico, un valore di soglia. Ogni singola unità logica confronta continuamente la somma pesata dei dati che si presentano ai suoi canali di ingresso con il valore della soglia. Se questo viene superato allora sul canale di uscita si troverà il valore uno. Altrimenti l'*output* sarà zero. L'analogia con il neurone biologico è evidente, così come l'estrema semplificazione rispetto ai reali processi che avvengono a livello cerebrale nelle cellule neurali. I canali di ingresso corrispondono, in questo modello, ai dendriti biologici mentre il canale di uscita rappresenta l'assone. I pesi delle connessioni sono in relazione con le intensità delle sinapsi mentre il calcolo effettuato dalle unità logiche a soglia simula approssimativamente quello eseguito dai neuroni biologici: i quali sulla base dell'integrazione dei segnali post-sinaptici, "decidono" se inviare o no un impulso nervoso lungo i loro assoni.
Una rete di neuroni artificiali di McCulloch e Pitts può apprendere dall'esperienza se si consente che i pesi delle connessioni possano essere modificati, sulla base dei dati di esperienza, con l'obiettivo di riuscire a generare, dopo un periodo di apprendimento, le risposte desiderate a date sollecitazioni.
L'istruttore fornirà, durante la fase di apprendimento, l'insieme delle risposte attese alle varie sollecitazioni, permettendo alla rete di modificare progressivamente i pesi delle connessioni con l'obiettivo di avvicinarsi alle risposte desiderate. La rete adeguerà i pesi secondo qualche specifico algoritmo in grado di

ridurre il più possibile la distanza tra la risposta reale e quella assegnata dall'istruttore. Il meccanismo noto come "retro-propagazione dell'errore" consiste nel calcolo delle derivate (effettuato rispetto ai pesi delle connessioni) della somma dei quadrati degli scarti tra uscite effettive e uscite attese. L'operazione viene ripetuta fino al raggiungimento del livello di apprendimento desiderato.

7. IL DIBATTO SULL'INTELLIGENZA DELLE MACCHINE

Il dibattito tra approccio logico-simbolico e sub-simbolico (reti neurali) è stato molto forte fino al passato recente e ha assunto toni piuttosto accesi. La tendenza odierna è, invece, basata su una ragionevole integrazione tra le varie proposte. Più in generale, l'approccio contemporaneo consiste in un atteggiamento di apertura nei confronti dello sviluppo di qualunque tipo di sistema in grado di esibire comportamenti intelligenti; sfruttando la logica (che si preoccupa della descrizione strutturale), la statistica (che cerca relazioni e associazioni basandosi sull'osservazione di grandi quantità di dati) e le reti neurali (che apprendono dall'esperienza), l'IA oggi si presenta come una disciplina matura e libera da condizionamenti "ideologici". I ricercatori di IA sanno bene che le intelligenze naturali sono in grado di prendere decisioni, risolvere problemi e friggere uova grazie ad una sorta di approccio integrato.
"Mi propongo di affrontare il problema se sia possibile per ciò che è meccanico manifestare un comportamento intelligente". In questo modo, introduceva l'argomento delle macchine intelligenti A. M. Turing nel 1948.
La scandalosa domanda di Turing non ha mai smesso di far riflettere scienziati e filosofi: a prescindere dai metodi scelti dai ricercatori di IA, molti studiosi continuano a interrogarsi attorno alla questione dell'intelligenza delle macchine. Pur non essendo particolarmente colpiti dal dibattito in corso sull'argomento, si ritiene che non si possa trascurarlo. Si farà, pertanto, qualche cenno alla discussione relativa alla possibilità che l'uomo possa costruire o meno macchine intelligenti quanto lui o, forse, più intelligenti di lui.

Il parere di Turing era presumibilmente favorevole a questa eventualità allorché egli formulò la sua domanda. Una domanda retorica, dunque, pensata per stabilire, una volta giunti al cospetto di una macchina, come testarne l'eventuale intelligenza. Turing era interessato al comportamento della macchina. Se, in buona sostanza, il comportamento di una macchina fosse indistinguibile da quello di un essere umano allora, per Turing, la macchina andrebbe definita come un agente intelligente.
Il superamento del test di Turing da parte di una macchina può realmente essere condizione sufficiente per stabilirne l'intelligenza? Non era certamente di questo parere J. R. Searle allorché propose il suo celebre esperimento della stanza cinese. Si immagini di rinchiudere in una stanza una persona che non conosca il cinese e che la stanza contenga un archivio ben organizzato che raccoglie tutti gli ideogrammi cinesi e un manuale, scritto in una lingua nota al protagonista dell'esperimento, che descriva senza alcuna ambiguità tutte le regole di associazione degli ideogrammi cinesi. Fuori dalla stanza cinese si trovano alcuni turisti cinesi che introducono nella stanza ideogrammi corrispondenti a una conversazione in cinese. Consultando archivio e manuale, il nostro amico chiuso

nella stanza manipolerà gli ideogrammi inseriti associandoli ad altri e risponderà ai cinesi attenendosi rigorosamente alle regole. L'uomo, che non sa il cinese, è pertanto in grado di conversare con il gruppo di turisti cinesi pur non capendo nulla della discussione in corso. La stanza cinese, nel suo complesso, supera così il test di Turing. Naturalmente l'uomo nella stanza rappresenta il calcolatore, il manuale corrisponde al programma mentre gli estensori del manuale sono i programmatori. La critica di Searle al test di Turing si può riassumere affermando che mentre un programma manipola simboli, cioè è formale o sintattico, il nostro cervello annette significati ai simboli: è semantico.

Ma si è proprio certi della validità di questi argomenti? Forse la pura manipolazione formale di simboli è sufficiente, dato un livello adeguato di complessità, a far emergere significati. Inoltre: perché è tanto fastidioso ammettere che il signore nella stanza manipola i simboli senza capirne il significato? Non è esattamente quanto accade ai singoli neuroni del nostro cervello? Ogni unità di elaborazione di quella fitta rete manipola flussi di segnali elettro-chimici senza annettere a essi alcun significato: il quale presumibilmente (ma chi lo può asserire con certezza?) emerge dalla complessità dell'organizzazione.

Il dibattito filosofico è molto acceso e, va detto, non è privo di un certo fascino. Tuttavia, si desidera enfatizzare qui che esso, forse, non è molto importante: i risultati della ricerca in IA sono stati e saranno utili indipendentemente dal fatto che sia possibile produrre artificialmente l'intelligenza. Inoltre, *en passant*, va di certo riconosciuto all'informatica avanzata di aver saputo produrre risultati che permettono simulazioni di piccoli frammenti di intelligenza, fornendo talora strumenti utilizzabili per una migliore comprensione dei meccanismi soggiacenti al funzionamento del cervello.

I ricercatori di IA, generalmente abbastanza indifferenti al dibattito filosofico, hanno proceduto con decisione verso la realizzazione di sistemi intelligenti dimostrando che, esattamente come nel caso dell'ozioso dibattito sulle macchine volanti, l'ultima parola spetta più agli ingegneri che ai filosofi.

8. IL PUNTO SULLE APPLICAZIONI

Come già osservato in precedenza, negli anni recenti sia i contenuti che le metodologie dell'IA sono radicalmente mutati rispetto al passato: l'intuizione ha lasciato spazio, nell'odierna IA, ai teoremi e alle evidenze sperimentali. Si può certo asserire che l'IA ha acquisito tutte le caratteristiche di una vera e propria scienza. Dal punto di vista applicativo, l'IA in passato è stata criticata perché troppo spesso si è dedicata a rozze semplificazioni e/o grossolane simulazioni della realtà. Esiste, infatti, una pletora di esempi minimali, che - come precedentemente detto - sono noti col nome di toy problem, il più famoso dei quali è il "mondo dei blocchi" (un tavolo — infinito — con sopra appoggiati alcuni cubetti colorati, oggetto delle azioni di un braccio robotico che li sposta da una posizione all'altra, obbedendo alle più elementari leggi della fisica). La raggiunta maturità della ricerca in IA fa sì che essa sia oggi interessata a fornire soluzioni utili nel mondo reale. Le attività e i campi applicativi sono talmente numerosi da rendere difficile realizzare un elenco esaustivo. Qui di seguito si riportano alcune aree dell'IA che hanno raggiunto la capacità di sviluppare applicazioni che meritano di essere evidenziate per il loro notevole impatto, in alcuni casi, e per le promettenti potenzialità, in altri.

- ***Pianificazione e schedulazione autonome*** - Un esempio è costituito dagli agenti remoti della NASA che, a milioni di chilometri dalla Terra, controllano e pianificano le operazioni di un veicolo spaziale. Tali agenti remoti sono in grado di generare piani sofisticati e dettagliati per il raggiungimento di obiettivi decisi (e specificati ad alto livello) da Terra, e di monitorare le operazioni del veicolo spaziale man mano che esegue i compiti assegnati.
- ***Controllo autonomo*** - In questo caso, l'esempio forse più impressionante è costituito dal sistema di visione Alvinn, addestrato a guidare un veicolo. Piazzato su un'automobile che ha attraversato gli Stati Uniti per 2850 miglia, il computer ha controllato la guida del veicolo per il 98% del tempo.
- ***Pianificazione logistica*** - Un caso in cui questo tipo di applicazione è stato utilizzato nel mondo reale è costituito dalla guerra del

Golfo del 1991, durante la quale la pianificazione e la schedulazione del trasporto di oltre 50 000 veicoli fu affidato dall'esercito americano a un sistema di IA in grado di tenere conto di destinazioni, strade, conflitti e di risolvere problemi che, con i metodi tradizionali, avrebbero richiesto tempi assai superiori.

Robotica - In questo campo le applicazioni al mondo reale sono numerose e sempre più affidabili e consolidate. Esse vanno da quelle nello spazio e nei fondali marini, a quelle in ambienti ostili e/o insicuri per operatori umani, alla robotica di servizio, ai primi robot di impiego domestico: un esempio di questa ultima applicazione è il robot autonomo aspirapolvere o tagliaerba. Per un'applicazione molto importante e ormai (relativamente) diffusa della robotica si pensi che oggi molti chirurghi utilizzano assistenti robot per delicate operazioni di (micro)chirurgia e che sono già state eseguite con successo operazioni chirurgiche in cui l'assistente robot e il paziente, da una parte, e il chirurgo, dall'altra, erano fisicamente distanti, addirittura in due diversi continenti.

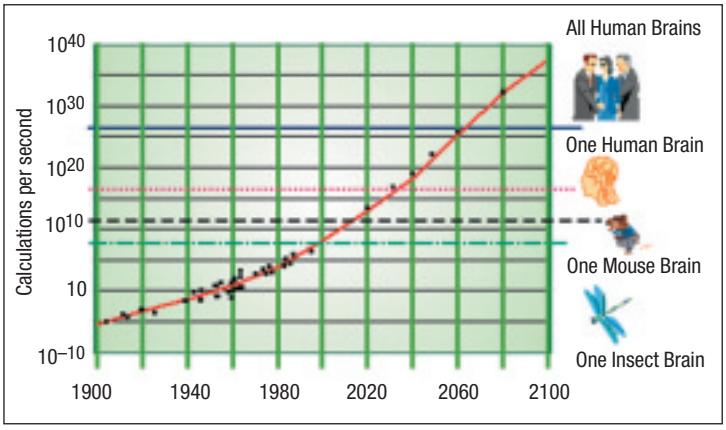
Diagnosi - In questo campo ci sono programmi in grado di effettuare diagnosi in specifiche aree della medicina, oppure di individuare guasti in sistemi fisici complessi. Il livello delle diagnosi di tali programmi è paragonabile a quello delle analisi che potrebbero essere realizzate da specialisti e professionisti.

Comprensione del linguaggio - Qui le applicazioni sono ormai innumerevoli e il livello di sofisticazione dei programmi è tale da essere, in certi settori, superiore a quello degli esseri umani: il programma Proverb, sviluppato da Littmann nel 1999, può risolvere complicati cruciverba (che richiedono competenze e conoscenze di alto livello) che molti esseri umani non sanno affrontare. La ricerca di informazioni sul web costituisce un settore applicativo molto importante per la ricerca sul linguaggio scritto (comprensione, reperimento di informazioni sulla base dei contenuti di documenti e non di semplici confronti testuali, estrazione di riassunti, traduzione automatica), mentre la diffusione di servizi telefonici e dei call center ha dato una notevole spinta alla diffusione di sistemi di comprensione e generazione di linguaggio parlato.

Gioco - Anche la ricerca sui giochi, presente già dagli albori dell'IA e palestra per le prime rudimentali tecniche di ricerca automatica di soluzioni e di apprendimento automatico, ha raggiunto una sua maturità applicativa. L'esempio più evidente è costituito dal gioco degli scacchi: come osservato in precedenza, un programma — Deep Blue — già sette anni fa ha battuto un campione mondiale. La ricerca sui programmi che giocano a scacchi prosegue, anche seguendo tecniche diverse da quelle impiegate in Deep Blue, e — cosa decisamente interessante — influenza il comportamento dei giocatori umani: i giocatori di scacchi, infatti, ormai si allenano specificamente per giocare contro i computer. Non sembra troppo azzardato prevedere che non sia lontano il momento in cui gli esseri umani non avranno più *chance* di battere un computer nel gioco degli scacchi.

9. UN FRAMMENTO DI FUTURO: INTELLIGENZA NELL'AMBIENTE E DOMOTICA

L'IA è, attualmente, forte del raggiungimento di risultati importanti, sia nello sviluppo di sistemi software, che nel campo molto più "visibile" della robotica. Di particolare impatto sul grande pubblico sono: l'impiego di robot per usi chirurgici, per l'esplorazione spaziale interplanetaria e nei campi di battaglia (in particolare gli *Unmanned Aerial Vehicles*, UAV, i veicoli senza piloti). La ricerca in IA sta ricevendo una grande spinta dai recenti avvenimenti internazionali, in



particolare dalle applicazioni riguardanti la sicurezza e da quelle militari[5].

Non è possibile qui passare in rassegna tutti i campi in cui ci si attendono importanti sviluppi portati dalla intelligenza artificiale nel prossimo futuro. Tra tanti è opportuno sceglierne uno, che sta diventando sempre più importante, anche per le significative implicazioni economiche, visto che prevede lo sviluppo di prodotti che potranno trovare impieghi di massa. Il settore è quello delle applicazioni dell'IA nella vita quotidiana, in particolare nella realizzazione della casa intelligente.

La realizzazione della casa intelligente, o "domotica", rappresenta una palestra molto importante per varie discipline e tecnologie, in particolare per applicazioni di robotica cognitiva. La chirurgia o l'esplorazione dei pianeti, quali domini per applicazioni robotiche, costituiscono certamente ambiti affascinanti, utili, e assolutamente imprescindibili per lo sviluppo della disciplina e per il progresso scientifico e tecnologico. Ma solo quando si avrà realizzato un'IA per applicazioni domestiche e per attività di vita quotidiana si potrà asserire di aver raggiunto un obiettivo applicativo che entrerà nella vita di tutti.

Cosa si intende con l'espressione "casa intelligente"? Si tratta di una applicazione ingegneristica alla portata di una tecnologia che se, forse, non è ancora completamente a punto è di certo assai prossima a divenirlo. È una tecnologia che utilizza sensori per monitorare apparecchi, impianti, fughe di gas; ma, ovviamente, non solo. Essa consentirà di azionare elettrodomestici da fuori casa e di controllarne lo stato. Si può facilmente immaginare che essa potrà occuparsi in modo automatico di chiamare la manutenzione del frigorifero, della lavastoviglie, della caldaia e della lavatrice se e quando necessario, o di fare la spesa al supermercato.

Nel passato, oltre che affidare i problemi della sicurezza alla sensoristica, si pensava che il controllo e l'utilizzo di elettrodomestici dall'esterno fosse utile per consentire agli utenti di poter svolgere funzioni in casa pur essendo fuori per lavoro o per diletto.

Oggi, questo punto di vista è considerevolmente mutato: si pensa a rendere più sicura, accogliente e vivibile la casa per chi la abita. In particolar modo, per coloro che hanno qualche forma di debolezza, dovuta a inabilità o all'età avanzata. Si pensi, per esempio, al caso dei malati di Alzheimer: queste persone, quando la malattia si trova in uno stadio avanzato, possono fare azioni in modo inconsapevole, per esempio uscire da casa, rischiando di arrecare danno a sé stesse. Si tratta di eventualità che potrebbero essere tenute sotto controllo da un ambiente intelligente che venisse programmato per evitare tali circostanze.

Nella domotica confluiranno i risultati della ricerca in intelligenza artificiale, robotica – in particolare robotica cognitiva – sensoristica, telecomunicazioni e del *pervasive computing*. Sensori e monitoraggio potranno essere proficuamente utilizzati per la sicurezza degli ambienti. Per quanto attiene alla sicurezza personale da intrusioni, questa sarà garantita da video citofoni, video telefoni nonché da applicazioni di biometria.

Il tele-monitoraggio di malattie croniche, realizzato mediante sistemi che ricordano all'utente di prendere certe medicine di cui necessita, o di fare certe operazioni e svolgere determinati compiti, consentirà di garantirne la sicurezza personale e di seguire da vicino i suoi eventuali problemi di salute.

Ma si può anche facilmente prevedere che servizi quali la pulizia di casa siano delegati a robot mobili e autonomi. Già oggi sono in vendita robot autonomi aspirapolvere o tosaerba. Non solo: chi non desidererebbe un robot-as-

[5] Dopo l'inizio della guerra in Iraq, i finanziamenti da parte del Dipartimento della Difesa USA ai gruppi che fanno ricerca in IA nelle più prestigiose università statunitensi sono aumentati del 43%.

sistente personale che si occupasse, solo per fornire qualche esempio, di ricordargli scadenze, suggerirgli azioni, fargli piccoli servigi o aiutarlo a riordinare le foto di famiglia?
Infine, non vanno dimenticati il tempo libero e gli aspetti ludici. Tele-conferenze, realtà virtuale e altro dovrebbero favorire la socializzazione e promuovere attività divertenti e rilassanti: incoraggiando a usare la fantasia e a trarre giovamento dagli aspetti giocosi dell'informatica, suggerendo, tra l'altro, nuovi mezzi e strumenti per intrattenere rapporti interpersonali e/o realizzare amicizie.
D. Norman in *The invisibile computer*, un libro che Business Week ha definito "la bibbia del pensiero post PC", esprime l'idea che la casa del prossimo futuro sarà intelligente, ma senza che tale circostanza comporti la presenza visibile di cavi e macchinari ingombranti. L'ambiente sarà pervaso di dispositivi intelligenti che coopereranno con gli abitanti per accrescere il loro comfort e la loro sicurezza.
Molta di questa tecnologia è già a disposizione dell'uomo. Questa considerazione è importante perché ci rassicura: non si sta parlando di idee di autori visionari che si realizzeranno in un non precisato futuro. Si sta trattando di tecnologie già ben consolidate che è possibile controllare e che si conoscono assai approfonditamente. Ci si muove su un territorio noto e ci si riferisce a un futuro ragionevolmente prossimo.
Si tratta di un futuro realistico nel quale si potrà contare su un rapporto uomo-macchina sofisticato e adattivo. Si potrà contare sulla costruzione di macchine dotate di (un certo livello di) cognizione e di emozioni e che comunicano in linguaggio naturale. Si tratta di un linguaggio carico di informazioni implicite, di dati d'esperienza e di conoscenze che permettono di comprendere le sottigliezze, risolvere le ambiguità, cogliere le arguzie e capire i sottintesi in una qualunque conversazione.
Ogni volta che si parla di supporto ad attività umane e cognitive ci si deve rendere conto della delicatezza del problema: da una parte, il controllo deve rimanere all'essere umano, perché non lo si vuole rendere prigioniero della sua casa e/o del suo robot assistente-personale. Al tempo stesso è necessario dotare di una certa autonomia il robot il quale deve saper reagire a imprevisti nonché essere in grado di sopperire a difetti di memoria e/o a debolezze cognitive dell'essere umano. Per questo sono ancora necessarie tanta ricerca e sperimentazione.

10. CONCLUDENDO

Il programma di costruire macchine in grado di comportarsi come noi non è mai stato abbandonato: ma oggi i ricercatori di intelligenza artificiale si occupano di una disciplina scientifica matura che, pur ancorata al sogno originale, fornisce soluzioni a problemi importanti e produce sistemi utili all'uomo e alla società.
Le sfide dell'IA sono innumerevoli e promettenti. La ricerca continua, nessuno dei problemi aperti dell'IA è veramente e completamente chiuso, su tutti si continua a migliorare, in un circolo virtuoso che fa corrispondere al progredire delle soluzioni teoriche un progresso notevole della tecnologia, e dal progresso della tecnologia, vede emergere nuove sfide teoriche.
La palestra dell'IA del prossimo futuro si chiama Robocup (www.robocup.org): il progetto lanciato nel 1997 con il primo campionato mondiale tenuto proprio nell'anno in cui si chiudeva la sfida degli scacchi, vuole arrivare, entro il 2050, a realizzare una squadra di robot calciatori autonomi (che giocano rispettando le regole internazionali) in grado di sconfiggere la squadra di calciatori umani campione del mondo.
Robocup sta diventando una grande competizione internazionale che vorrebbe coinvolgere tutta la ricerca in robotica e in intelligenza artificiale; l'idea è che lo stimolo della competizione induca a integrare al meglio le diverse tecnologie che, una volta ottimizzate sul gioco del calcio, potranno poi essere trasferite a problemi concreti e significativi per l'industria e per i servizi.
Con Robocup la comunità dei ricercatori di IA sembra richiudersi intorno a un gioco, piuttosto che affrontare problemi di ricerca "seria". Il gioco è però di una difficoltà non affrontabile agli attuali livelli di conoscenza e di sviluppo tecnologico, quindi, affascina anche chi non entrerebbe mai in uno stadio per vedere una partita di pallone perché ritiene che il calcio sia un gioco ... che non richiede molta intelligenza.

I filoni di ricerca dell'intelligenza artificiale

La moderna visione di un sistema di intelligenza artificiale consiste nel considerarlo come un agente (o una molteplicità di agenti) in grado di espletare compiti di alto livello. I filoni di ricerca verranno, quindi, qui di seguito illustrati in termini delle capacità di cui vogliamo che l'"agente intelligente" sia dotato. Elencheremo dapprima i filoni di ricerca più classici, poi quelli più moderni. Tralasciamo completamente i filoni applicativi (per esempio, "IA e medicina" oppure "IA e formazione"), che, pur richiedendo soluzioni specifiche per problemi peculiari del campo, necessiterebbero di troppo spazio.

Risoluzione automatica di problemi (Problem solving)

Gli agenti per il problem solving sono sistemi che identificano sequenze di azioni per il raggiungimento di stati desiderati. Gli algoritmi di ricerca automatica di soluzione si dividono in "non informati", cioè sprovvisti di informazioni diverse dalla semplice definizione del problema da risolvere, e che, quindi, eseguono una ricerca uniforme nello spazio degli stati e algoritmi di ricerca "informati", cioè provvisti di informazioni utili per individuare la soluzione. Mentre i primi diventano inutilizzabili appena la complessità del problema aumenta, i secondi possono essere molto efficaci se dotati di una buona "euristica", cioè di una funzione che – utilizzando informazioni sul dominio del problema – permetta di guidare la ricerca più rapidamente verso una soluzione.

Rappresentazione della conoscenza e ragionamento automatico

Si tratta di concetti fondamentali per l'intera intelligenza artificiale. La conoscenza e il ragionamento aiutano sia noi esseri umani che gli agenti artificiali ad avere comportamenti adeguati in ambienti complessi e/o parzialmente osservabili. Un agente basato sulla conoscenza può combinare le conoscenze di carattere generale con i dati percepiti per dedurre aspetti non espliciti e/o nascosti che possono essere utili per selezionare e scegliere tra comportamenti alternativi. La ricerca in questo campo riguarda il progetto di linguaggi, metodi e tecniche per rappresentare conoscenza su domini applicativi, e algoritmi e metodi per fare inferenze e trarre quindi conclusioni valide nel dominio di interesse. Per quanto riguarda i linguaggi, il punto cruciale è il potere espressivo (Posso rappresentare il tempo? Posso rappresentare relazioni spaziali? Posso rappresentare la conoscenza di diversi agenti in modo che essa sia "privata" per ciascuno di essi e non accessibile ad altri?). Per quanto riguarda i meccanismi inferenziali, gli aspetti cruciali sono la correttezza e la completezza, e — non meno importante — la complessità di calcolo.

Pianificazione automatica

Dati un dominio applicativo, la descrizione di uno stato di tale dominio in cui l'agente intelligente si trova (detto stato iniziale), la descrizione di un obiettivo (o stato finale) che l'agente vuole raggiungere e, infine, la descrizione dell'insieme delle azioni che esso può compiere, la pianificazione corrisponde all'identificazione di una sequenza di azioni necessarie per raggiungere un obiettivo. Si parla di "pianificazione classica" nel caso in cui gli ambienti siano completamente osservabili, deterministici, statici (tali cioè per cui qualunque variazione è causata solo dall'agente) e discreti dal punto di vista temporale, delle azioni, degli oggetti, degli effetti. È facile convincersi che raramente i domini di interesse per applicazioni pratiche godono di queste proprietà. La ricerca di algoritmi efficienti (e completi) per la pianificazione automatica in ambienti non deterministici e parzialmente osservabili è oggetto di intensa ricerca.

Ragionamento probabilistico

Spesso un agente non ha accesso a tutte le informazioni relative all'ambiente con cui deve interagire e deve, pertanto, decidere in condizioni di incertezza, basandosi, quindi, su ragionamenti di tipo probabilistico. L'incertezza modifica i modi con cui gli agenti prendono decisioni: in condizioni di incertezza la decisione più razionale è quella che seleziona le azioni che corrispondono alla più elevata utilità attesa, ottenuta mediando su tutti i possibili risultati dell'azione.

Apprendimento automatico e data mining

L'idea fondamentale relativa all'apprendimento è quella secondo cui quanto viene percepito da un agente dovrebbe essere utilizzato non solo per il suo comportamento nell'ambiente percepito al momento presente, ma anche immagazzinato ed elaborato per accrescere la sua abilità nelle azioni future. L'apprendimento riguarda la capacità di un agente di osservare i risultati dei processi delle sue stesse interazioni con l'ambiente. Si parla di tecniche di apprendimento induttivo intendendo tecniche di tipo simbolico (per esempio, l'apprendimento di alberi di decisione a partire da esempi). Si parla poi di apprendimento connessionistico (quello che riguarda l'addestramento di reti neurali) e di apprendimento statistico (quello che si basa su tecniche di tipo statistico). Con l'espressione "data mining" si identificano le tecniche di apprendimento applicate a grandi quantità di dati (per esempio i tabulati delle telefonate immagazzinati da una grande compagnia telefonica, i tabulati dei consumi di un grande distributore) alla ricerca di regolarità, regole di comportamento ecc..

Comunicazione

Con il termine comunicazione si intende identificare lo scambio intenzionale di informazioni mediante la realizzazione e la percezione di un sistema condiviso di segni convenzionali. La problematica della comunicazione coinvolge il linguaggio, le ambiguità dello stesso (sia strutturali che semantiche), la disambiguazione, la comprensione del discorso, i modelli probabilistici ecc.. La ricerca nella comunicazione si occupa dell'analisi e della generazione di linguaggio sia scritto che parlato, della traduzione automatica, della generazione di riassunti ecc.. Si occupa, inoltre, della interazione tra essere umani e sistemi artificiali in forma amichevole e adattiva e anche non verbale.

Percezione
La percezione, ottenuta mediante sensori, è necessaria per fornire all'agente artificiale informazioni sull'ambiente con cui interagisce. Tutti i tipi di percezione sono importanti e oggetto di studio. Sono stati particolarmente esplorati la percezione acustica e visiva. Per quanto riguarda la percezione acustica, la ricerca si focalizza sulla comprensione del linguaggio parlato e su quella della scena acustica (per esempio, la localizzazione di sorgenti sonore); per quanto riguarda, invece, la visione artificiale, il riconoscimento di oggetti (e non solo in particolari scene e condizioni di luce) e la comprensione di scene in movimento restano problemi aperti perché particolarmente complessi.

Robotica
I robot sono agenti fisici artificiali che svolgono funzioni e compiti manipolando gli oggetti del mondo fisico. Per questo essi devono essere provvisti di sensori, per acquisire le informazioni necessarie *dall'ambiente*, e di attuatori per agire *sull'ambiente* (ruote, braccia meccaniche ecc.). La ricerca in robotica, che per molto tempo aveva seguito strade indipendenti dall'intelligenza artificiale, sta ottenendo una rinnovata attenzione da parte dei ricercatori di IA, soprattutto per quel che riguarda la progettazione di sistemi robotici cognitivi, cioè robot che esibiscono capacità di decisioni autonome di alto livello.

Sistemi multiagente
Un agente artificiale non agisce in situazioni di isolamento: esso deve interagire con esseri umani ma anche con altri sistemi artificiali. La ricerca su sistemi multiagente si occupa della rappresentazione della conoscenza e del ragionamento in situazioni in cui molti agenti sono presenti in un sistema, ciascuno dotato della sua conoscenza e del suo "punto di vista" sul mondo. Aspetti particolarmente importanti sono la comunicazione, il coordinamento e la pianificazione cooperativa: sono necessari affinché il sistema multiagente possa perseguire efficientemente (distribuendo il carico di lavoro o sfruttando le capacità specifiche di alcuni agenti presenti nel sistema) un obiettivo comune anche in ambienti eventualmente resi ostili dalla presenza di un'altra squadra di agenti avversari.

Programmazione naturale
La programmazione naturale è quella branca della ricerca automatica di soluzioni che si occupa di riprodurre meccanismi che si osservano in natura. Rientra in questo ambito la ricerca sugli algoritmi genetici, cioè algoritmi di ricerca che simulano i meccanismi della evoluzione darwiniana di popolazioni di individui. Sono di particolare attualità gli algoritmi che riproducono il comportamento di colonie di animali quali, per esempio, colonie di formiche che — sfruttando opportuni meccanismi di comunicazione — trovano velocemente il cammino più breve verso l'obiettivo (per esempio, il cibo).

Modellazione cognitiva
La ricerca in intelligenza artificiale ha costantemente tratto lo spunto dai risultati della ricerca in psicologia cognitiva e sperimentale, e ha fornito ad essa modelli artificiali (anche se molto grossolani) del cervello. L'evoluzione della ricerca e della tecnologia porta a un raffinamento di questi modelli: una sfida per il prossimo futuro — posta dai cognitivisti inglesi — consiste nell'uso del *grid computing* come modello del cervello, e in questo caso il modello può essere più realistico, viste le dimensioni del sistema di calcolo costituito da una rete ad estensione mondiale.

Rappresentazione delle emozioni
Una direzione recente della ricerca sulla comunicazione tra persone e sistemi artificiali è quella della comprensione e della generazione di particolari stati emotivi: in particolare quella della comprensione e generazione di situazioni umoristiche. Lo studio del cosiddetto umorismo computazionale dovrebbe, tra l'altro, aiutare a svelare i meccanismi psicologici, ad oggi non ben compresi, che consentono di cogliere gli aspetti impliciti e riposti del linguaggio naturale che danno origine all'ilarità (caratteristica, a quanto sembra, esclusivamente umana). È probabile che tali ricerche possano rivelarsi di grande aiuto nella simulazione di stati emotivi e nel miglioramento delle relazioni uomo-macchina.

Bibliografia

[1] Bonarini A.: Sistemi fuzzy. *Mondo Digitale*, Anno II, n. 5, 2003, p. 3-14.

[2] Burattini E., Cordeschi R.: *L'intelligenza artificiale*. Carocci, Roma 2003.

[3] Buttazzo G.: Coscienza artificiale: missione impossibile? *Mondo Digitale*, Anno I, n.1, 2002, p. 16-25.

[4] Carlucci Aiello L., Cialdea Mayer M.: *Invito all'intelligenza artificiale*. Franco Angeli, Milano 1995.

[5] Castelfranchi Y., Stock O.: *Macchine come noi*. Laterza, Bari 2000.

[6] Dapor M.: *L'intelligenza della vita. Dal caos all'uomo*. Springer Italia, Milano 2002.

[7] Dennett D.: *Consciousness Explained*. Little, Brown and Company, Boston, Toronto, London 1991 [trad. it. della prima edizione: *Coscienza. Che cosa è*. RCS Rizzoli Libri, Milano 1993].

[8] Gori M.: Introduzione alle reti neurali artificiali. *Mondo Digitale*, Anno II, n. 8, 2003, p. 4-20.

[9] Norman D.A.: *The Invisible Computer: Why Good Products Can Fail, the Personal Computer Is So Complex, and Information Appliances Are the Solution*. MIT Press, New York 1998.

[10] Russel S., Norvig P.: *Artificial Intelligence: A Modern Approach*. Second Edition, Prentice Hall, 2003 [trad. it. della prima edizione: I*ntelligenza Artificiale. Un approccio moderno*. UTET-Libreria, Torino 1998].

[11] Williams S.: *Storia dell'intelligenza artificiale*. Garzanti, Milano 2003.

[12] Zaccaria R.: Aspettando Robot. *Mondo Digitale*, Anno II, n. 7, 2003, p. 3-18.

Bibliografia delle Immagini

Pagina 5 – Losano M.G.: Storie di automi dalla Grecia classica alla bella époque. Einaudi, Torino, 1991. Capitolo V, figura 34, p. 102.

Pagina 16 – Luvison A.: Net economy, istruzione, tecnologie digitali, AEI, Vol. 87, n. 11, 2000, figura 3, p. 36.

Pagina 13 – Bozzo M.: La grande storia del Computer, Ed. Dedalo, Bari, 2001, p. 147.

Pagina 15 – Giuseppe O. Longo: Uomo, Computer: partita a scacchi tra esseri alieni. Corriere della Sera.

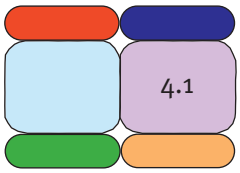
Luigia Carlucci Aiello è professore ordinario di Intelligenza Artificiale all'Università di Roma "La Sapienza". Ha fornito contributi alla ricerca nella rappresentazione della conoscenza e ragionamento automatico ed esplorato vari settori applicativi dell'IA. Autore di circa 200 pubblicazioni, ha diretto progetti di ricerca italiani ed europei. Fondatrice e socia onoraria dell'AI*IA, l'Associazione Italiana per l'Intelligenza Artificiale, è Fellow della AAAI e dell'ECCAI, le associazioni americana ed europea di IA.
aiello@dis.uniroma1.it

Maurizio Dapor, fisico e giornalista scientifico, è il responsabile del Laboratorio Servizi alle Imprese dell'Istituto per la Ricerca Scientifica e Tecnologica (IRST) di Trento. È autore di molte pubblicazioni di fisica computazionale, di alcuni libri scientifici e di numerosi articoli a carattere divulgativo i cui argomenti spaziano dalla fisica teorica all'intelligenza artificiale. I suoi libri più recenti sono: *L'intelligenza della vita. Dal caos all'uomo* (Springer, Milano 2002) e *Electron-Beam Interactions with Solids. Application of the Monte Carlo Method to Electron Scattering Problems* (Springer, Berlino 2003).
dapor@itc.it

IL VALORE DEL CRM NEL BUSINESS DI RETI DI PMI GOVERNATE

Roberto Bellini

Il CRM è rapidamente diventata una tecnologia matura che oggi è possibile modulare in una ampia gamma di soluzioni a costi accessibili per grandi, medie e piccole aziende, soprattutto se parte di reti distributive, governate dalle esigenze del cliente finale che hanno in comune un distributore e il suo fornitore. Per avere successo però, deve essere disegnata, sviluppata e messa in opera insieme a processi di Marketing & Vendita aziendali e alle competenze di ruoli professionali adeguati.



1. INTRODUZIONE

Proviamo a dimenticarci per un attimo che, nell'ambito degli esperti di tecnologia, l'acronimo CRM (*Customer Relationship Management*) non voglia solo identificare una piattaforma tecnologica, ma un Sistema strutturato e operante di Management della Relazione con il Cliente (SMaRC).
Adottiamo questa ipotesi di lavoro per "illuminare" alcuni "nostri" comportamenti da consumatori.
Entriamo in un supermercato alimentare, dove il prezzo medio di un prodotto è di qualche euro, la spesa totale per un singolo accesso è intorno ai 25-30 euro e la frequenza di accesso è mediamente una volta alla settimana: percorriamo con il nostro carrello l'itinerario che ci porta all'acquisto dei vari tipi di prodotti, arriviamo alla cassa e al momento di pagare il cassiere ci chiede se abbiamo la tessera (di socio, di fedeltà ecc.); in caso positivo abbiamo diritto a qualche forma di sconto o di promozione. Nel momento in cui la transazione è completata o al termine della giornata, avvengono due operazioni: la prima di tipo amministrativo-logistico, che scatena l'aggiornamento del livello del magazzino per ciascun tipo di prodotto e la seconda di tipo commerciale, che determina l'aggiornamento dell'archivio dei clienti con tessera, con il saldo del loro acquisto. Se abbiamo bisogno di approvvigionamenti ulteriori e rientriamo dal lavoro tardi, possiamo riempire il nostro carrello virtuale accedendo al portale della catena di supermercati e farci mandare la spesa a casa la mattina dopo, pagando con la carta di credito. Sia nel primo caso che nel secondo, siamo entrati a far parte, dal punto di vista logico, dello SMaRC del supermercato.
Entriamo in una concessionaria auto, dove il prezzo medio del prodotto è di 2 o 3 decine di migliaia di euro e la frequenza di accesso è mediamente una volta ogni 2-3 anni: al momento dell'acquisto tutti i nostri dati anagrafici vengono registrati per la formalizzazione dell'atto di vendita e la produzione della documentazione amministrativa necessaria alla circolazione dell'auto acquistata e alla copertura assicurativa; nel giro del primo anno

ritorniamo almeno 1 o 2 volte presso la concessionaria per le verifiche e le messe a punto previste dalla garanzia. Ogni volta vengono aggiornati, sul *file* attivato sul nostro nome, i dati relativi alle operazioni eseguite; dopo i 24 mesi dall'acquisto cominciano ad arrivarci piccole promozioni dalla concessionaria per la revisione dell'auto e le segnalazioni relative ai nuovi modelli in uscita. Siamo entrati a fare parte del sistema SMARC della concessionaria; se il sistema SMaRC è sofisticato potremmo essere entrati a fare parte anche dello SMaRC dell'azienda Fornitrice dell'auto alla concessionaria.
Entriamo in banca, dove abbiamo il nostro conto corrente e appoggiato il mutuo della casa e/o il *leasing* dell'auto; utilizziamo il bancomat come mezzo di pagamento abbastanza frequentemente; il prezzo medio di ogni operazione è intorno a qualche euro e accediamo ai servizi della banca mediamente una volta alla settimana; ogni nostra operazione non solo aggiorna il nostro conto, ma lascia una traccia nel file associato al nostro nome. Siamo parte del sistema SMaRC della banca.
Possediamo un cellulare. Abbiamo cominciato a usarlo qualche anno fa ed è diventato uno strumento indispensabile nella vita quotidiana, tanto da usarlo molte volte al giorno, dimenticandoci o non facendo caso a quanto spendiamo al mese, anche se la cifra per l'insieme dei servizi pagati mensilmente continua ad aumentare pur riducendosi i costi sia dei terminali che di ogni singolo servizio. Siamo bombardati, letteralmente, dalla pubblicità televisiva di 4-5 grandi operatori che ci contendono l'uno all'altro con servizi sempre nuovi e da un certo numero di *Short Message Service*, o SMS, (per fortuna non tanti) di promozione di alcuni servizi (operazione di cosiddetto *Direct Marketing*). Possiamo sapere tutto sui servizi del nostro operatore facendo un numero di servizio e chiedendo all'operatore che risponde, che forse si trova a parlare da casa, oppure andando sul portale dell'operatore, oppure entrando in uno delle migliaia di negozi sparsi sul territorio nazionale; dovunque accediamo, se ci chiedono il numero di telefono, l'operatore con cui siamo in contatto o il commesso del negozio in cui siamo entrati, possono richiamare, almeno in teoria, il nostro codice (che coincide con il nostro numero di telefono) e vedere il nostro profilo aggiornato in termini di tariffe, consumi, tipo di servizi attivi ecc.. Siamo parte del sistema SMaRC dell'operatore telefonico.
La stessa cosa avviene, con modalità naturalmente diverse in funzione del prodotto/servizio che acquistiamo, per molti altri tipi di prodotti/servizi, indipendentemente dal prezzo medio, dalla frequenza di acquisto o di accesso; è "normale" che il tracciamento dei dati di acquisto e di comportamento del cliente, nell'uso del prodotto/servizio acquistato, venga sviluppato nei mercati maturi e saturi, in cui i vari fornitori si fanno iper-concorrenza cercando di strapparsi i clienti l'uno all'altro.

2. COME NASCE IL CUSTOMER RELATIONSHIP MANAGEMENT

Alla fine degli anni '80 la Harvard Business School condusse un'analisi sui costi e i ricavi derivanti dal modo in cui "servire" i clienti e successivamente pubblicò i risultati della ricerca nel numero di settembre-ottobre del 1990 della *Harvard Business Review*. Il risultato più importante a cui giunse fu che, a causa dell'alto costo di acquisizione di un cliente, un costo aggiuntivo del 5% per la "fidelizzazione del cliente" può portare a un incremento del "profitto per cliente" compreso fra il 25% e il 90%!
Successivamente, molte altre ricerche hanno messo a fuoco che per ogni azienda, la produzione di una "scala" o di una "piramide" dei clienti sia un modo per segmentare la clientela acquisita rispetto a due fattori in contraddizione fra di loro:

- da una parte, la segmentazione della clientela acquisita per livello di fatturato e/o di primo margine di contribuzione permette di ottimizzare le risorse impegnate nella attività commerciale rispetto all'obiettivo di migliorare la propria quota di mercato (inteso come quota di fatturato sul totale del fatturato di tutti i concorrenti attivi);
- dall'altra, nei mercati saturi e ipercompetitivi, l'incremento della propria penetrazione nel mercato (inteso come quota della numerosità dei clienti acquisiti nel mercato disponibile, costituito da tutti i clienti che hanno acquistato almeno un prodotto di quel tipo) costitui-

sce una necessità, costosissima, soprattutto se l'azienda vuole mantenere o migliorare nel medio e lungo termine la propria posizione competitiva anche con nuovi prodotti

Infine, ma non meno importante, sono cominciati a uscire i risultati di ricerche sulla importanza della "Soddisfazione del Cliente" per poterlo considerare come "una risorsa" da "sfruttare" ulteriormente con il tentativo di vendergli nuovi prodotti.

Tutte queste considerazioni hanno portato molti *top manager* a considerare che l'obiettivo del rafforzamento della relazione con il cliente, una volta conquistato, sia centrale per ogni iniziativa di sviluppo della "fedeltà" del cliente acquisito.

A metà degli anni '90 venne coniato per il mercato americano il termine *Customer Relatioship Management* (CRM); contemporaneamente nacque Siebel, società specializzata nella gestione della relazione con il cliente, che si affermò come fornitore capace di accreditarsi presso i suoi clienti garantendo loro una "soddisfazione" del 100%, che tradotto dal gergo commerciale significa "soddisfatto o rimborsato".

In Italia, il CRM emerse nel settore informatico nel 1998, con alcune prime applicazioni che vedono oggi in prima linea, fra altre, la società di consulenza Inno (poi trasformatasi in Allaxia) che citiamo per avere prodotto la metodologia COSMO-*Customer Oriented Sales&Marketing Optimisation*, che utilizzeremo come riferimento nel seguito di questo articolo.

In quali aziende si trovano le prime applicazioni di CRM? Le tecnologie di CRM, disegnate per supportare la rete dei venditori nella conquista di nuovi clienti e la gestione dei reclami dei clienti insoddisfatti, cominciarono a essere introdotte o presso quelle aziende in cui lo sviluppo del *business* nasceva centrato sulla gestione del cliente, come per gli operatori di telefonia mobile, oppure più gradualmente presso quelle grandi imprese in cui si erano già sviluppati interventi orientati a migliorare l'efficienza del sistema produttivo e l'ottimizzazione del sistema di gestione delle forniture (*Supply Chain Management*, SCM).

L'introduzione del concetto di "relazione con il cliente" si amplia per comprendere tutti i rapporti dell'azienda fornitrice con il proprio cliente ovvero: a partire dal momento in cui lo identifica come potenziale cliente e cerca di "accreditarsi come suo fornitore" o anche al momento in cui riesce a conquistarlo strappandolo ai suoi concorrenti, cioè a fargli pagare l'acquisto del prodotto disponibile a magazzino o nel Punto Vendita che porta il suo marchio oppure al momento in cui eventualmente lo assiste in caso di guasto del prodotto o di supporto nell'uso del servizio acquistato o al momento, infine, in cui il cliente, quando avverte un nuovo bisogno, riprende in considerazione l'idea di comprare dalla stessa azienda perché è rimasto soddisfatto del primo acquisto.

Nulla di particolarmente nuovo per l'impresa, salvo il piccolo particolare che il cliente si trasforma, in un ambiente di mercato a competitività crescente e saturato da una offerta che si amplia continuamente e sviluppa la sua aggressività, in obiettivo strategico: la relazione con il "cliente" deve venire ottimizzata per mantenere o migliorare la quota di mercato e il cliente deve essere posto "al centro" dell'attenzione di tutta l'azienda.

La "fidelizzazione del cliente" si ottiene approfondendone la conoscenza in termini di mappatura delle sue caratteristiche e di analisi delle modalità con cui utilizza il prodotto/servizio acquistato, analizzando poi la nascita di nuovi bisogni e della corrispondenza fra bisogni e funzioni d'uso del nuovo prodotto/servizio che può soddisfare tale bisogno (Profilo del Cliente), e ancora in termini di studio del suo processo di acquisto in presenza di proposte competitive (Comportamento di Acquisto).

Lo svilupparsi di Internet e della possibilità di vendere servizi e prodotti anche attraverso il "media" elettronico (Commercio Elettronico) enfatizza ancora di più la possibilità di conoscere il cliente sia nella fase di acquisto che successivamente nella fase di utilizzo del prodotto/servizio acquistato, permettendo, inoltre, di avere risposte molto più rapide e a costi molto inferiori rispetto a quelle che si possono ottenere con acquisti tradizionali.

La tecnologia CRM assume un'importanza crescente in questo contesto perché permette di mantenere un migliore controllo di tutte le fasi del processo di vendita e di fidelizzazione, permettendo, inoltre, di realizzare la "relazione uno a uno" fra il sistema organizzativo dell'impresa e il singolo consumatore

o, nel caso del *Business-to-Business* (B2B), il sistema organizzativo dell'impresa cliente.

3. I PRINCIPALI COMPONENTI DI UN SMaRC

La tecnologia CRM è attualmente una delle tecnologie considerate più importanti nel contesto del business. È tuttavia necessario porre molta attenzione a cosa definiamo come CRM e a valutare le implicazioni che l'introduzione di questo tipo di tecnologia può avere sul piano strategico e organizzativo, soprattutto se il progetto di CRM ha un obiettivo di alto livello di integrazione fra più funzioni aziendali.
Spesso, e in Italia ancora di più, è stato considerato come CRM il *Call Center* (CC), strumento nato per rendere più efficiente il servizio di risposta a clienti insoddisfatti o che hanno bisogno di aiuto nell'utilizzo del prodotto/servizio che hanno acquistato.
In realtà, per ragionare meglio sul CRM è necessario partire da una definizione molto ampia, del tipo: intendiamo per Sistema strutturato e operante di Management della Relazione con il Cliente (SMaRC), un sistema aziendale che si pone l'obiettivo di mettere il cliente al centro della operatività di tutte le funzioni aziendali con cui entra in contatto, da quelle commerciali a quelle di erogazione e di assistenza tecnica, a quelle amministrative per la gestione degli ordini, del fatturato e dell'incassato, a quelle di gestione delle campagne di promozione, di pubblicità e comunicazione, a quelle sostenute dalla funzione Sistemi Informativi per la realizzazione e gestione dei sistemi di interfaccia telefonica e con il *web*, a quella dei servizi generali per quanto riguarda la posta tradizionale, a quelle, infine, del marketing in relazione alle ricerche di mercato e alle indagini di *customer satisfaction*; lo SMaRC, così definito, non indica più, quindi, una soluzione tecnologica di supporto alla relazione con il cliente, ma piuttosto una vera e propria strategia di business disegnata per ottimizzare la redditività, i ricavi e la soddisfazione del cliente.
Per realizzare un sistema complesso di questo tipo può essere molto utile disegnare un modello del "Sistema strutturato e operante di Management della Relazione con il Cliente" in cui siano messi in evidenza non solo tutte le articolazioni dei vari componenti necessari, ma anche i vari livelli di integrazione fra queste componenti nel sistema, che rappresentano una delle maggiori difficoltà per la realizzazione di uno SMaRC di successo: si tratta, infatti, per poter gestire in modo unitario le "relazioni" con il cliente, di riportare nella scheda cliente l'aggiornamento di tutti i vari tipi di relazione che l'azienda intrattiene con il cliente stesso, dal momento in cui il sistema entra in funzione.
L'obiettivo di progettare e costruire un Sistema SMaRC adeguato prevede la necessità di ingegnerizzare tre tipi di componenti fondamentali:

1. il portafoglio dei processi di marketing&vendita strutturati in fasi e attività tali da permettere:
   - ***a.*** la qualificazione e conquista dei clienti, e una volta conquistati, la loro fidelizzazione per poter sviluppare *cross-selling* e *up-selling*;
   - ***b.*** lo sviluppo delle trattative fino alla firma dell'ordine o del contratto o comunque dell'acquisto, con tutte le relative implicazioni di tipo amministrativo;
   - ***c.*** il servizio alla clientela conquistata sia in termini di assistenza tecnica che di risposta a eventuali reclami o altri bisogni informativi; tutte le attività sviluppate nell'ambito dei processi di marketing&vendita sono centrate sull'archivio clienti;
2. il sistema dei ruoli professionali e manageriali dei vari attori coinvolti nei processi, identificando per ogni ruolo i dati e le informazioni necessarie per la gestione dei vari tipi di relazione con il cliente;
3. una piattaforma tecnologica che permetta di sostenere tre tipologie fondamentali di funzioni d'uso;
   - ***a.*** le funzioni di CRM Operazionale, che includono sia quelle del *Contact Center* (CC) che quelle della *Sales Force Automation* (SFA) e hanno lo scopo di sostenere le attività di interazione diretta con i clienti;
   - ***b.*** le funzioni di CRM Collaborativo, che includono la gestione dei vari canali di contatto fra clientela e azienda (via Web, telefono fisso e mobile, SMS ecc.) e hanno lo scopo di regolare e controllare il traffico dei messaggi per i vari tipi di contatto;

c. le funzioni di CRM Analitico, che includono sia le funzioni di *Business Intelligence* (BI) che quelle di *Customer Intelligence* (CI) e hanno lo scopo di sostenere le attività di analisi e comprensione delle caratteristiche e dei comportamenti del singolo cliente e del loro insieme.

La piattaforma tecnologica sopraindicata ha la funzione di supportare i processi di marketing&vendita e alimentare in termini di dati, informazioni e indicatori di prestazione tutti i ruoli coinvolti nel rapporto con i clienti e il mercato.

Questi tre tipi di componenti vanno gestiti secondo regole aziendali indicate dalla direzione aziendale e mirate a facilitare, ma anche controllare, il sistema operante messo in campo. La catena del valore aziendale si modifica con la introduzione della piattaforma di Contact Center e di Sales Force Automation (Figura 1) e con l'inserimento di funzioni di Customer/Business Intelligence, rendendo possibile all'azienda non più solo di "guidare" i suoi clienti all'acquisto dei propri prodotti, ma anche di "farsi guidare" dai propri clienti, ascoltandone e interpretandone i bisogni in modo da anticipare lo sviluppo di prodotti/servizi costruiti su misura ed erogati attraverso i canali più efficienti, in modo che siano pronti quando la richiesta del mercato emerge.

Utilizzando come riferimento, per descrivere nelle grandi linee i principali componenti di un sistema SMaRC, il modello Cosmo, iniziamo l'analisi dei vari componenti indicati a partire dai processi di marketing&vendite.

L'architettura considerata si articola in 8 processi che sono:

1. **Analisi e Pianificazione di Business**: analisi del contesto ambientale esterno di riferimento (mercato, concorrenza e prodotti) e pianificazione delle strategie aziendali e di business; l'*output* di questo processo è costituito dal Piano Strategico di Business che costituisce l'*input* al Piano Operativo di Marketing.
2. **Pianificazione Operativa di marketing&vendite per cliente**: definizione degli obiettivi commerciali per cliente e pianificazione delle campagne di marketing&vendite, di Assistenza Tecnica e Customer Service; l'output di questo processo è costituito dal Piano Operativo di Marketing che costituisce l'input per tutte le campagne di cattura di nuovi *prospect*, di fidelizzazione dei clienti acquisiti, di gestione delle trattative per le offerte più significative, e ancora per le campagne di Customer Care e Customer Service e per le campagne di promozione, comunicazione e pubblicità.
3. **Cattura nuovi clienti**: implementazione delle campagne di marketing&vendite per la cattura di nuovi prospect; l'output di questo processo è costituito dall'incremento del numero dei clienti per i segmenti di offerta obiettivo.
4. **Fidelizzazione clienti**: implementazione delle azioni di marketing&vendite per la fidelizzazione dei clienti acquisiti; l'output di questo processo è costituito dall'aumento degli ordini e del fatturato da parte del portafoglio dei clienti acquisiti, per i segmenti di clientela obiettivo.
5. **Gestione Trattative**: implementazione delle azioni commerciali orientate a gestire il sistema cliente per la definizione, la preparazione e la consegna delle offerte più significative; l'output di questo processo è costituito dall'incremento dei contratti significativi in volume, fatturato e redditività.

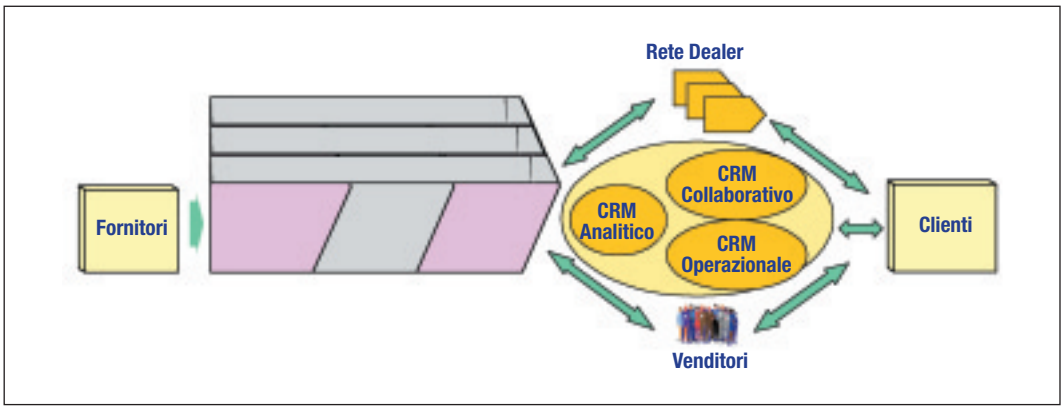


FIGURA 1
Le varie tipologie di piattaforme tecnologiche per il CRM

6. Customer Care/Customer Service: implementazione delle azioni di assistenza tecnica alla clientela sia in fase di installazione del prodotto/servizio che in fase di assistenza all'utilizzo, di sviluppo degli interventi di ripristino delle funzionalità d'uso in caso di guasto tecnico, di risposta alle richieste di informazione e ai reclami; l'output di questo processo è costituito dal miglioramento del livello di soddisfazione della clientela che compra per la prima volta, o ricompra, un prodotto sostitutivo e dalla riduzione dei clienti insoddisfatti in generale.

7. Pubblicità e Comunicazione di Marketing: implementazione delle campagne di pubblicità, comunicazione e promozione previste dal Piano Operativo di marketing&vendita; l'output di questo processo è costituito dall'ampliamento della notorietà presso la comunità degli *stakeholder* (clienti, fornitori, partner e dipendenti).

8. Misura e Controllo: verifica dell'andamento delle campagne di marketing, delle azioni commerciali e di assistenza tecnica e servizio, in termini di efficacia, efficienza, qualità e produttività; fra le misure più significative, viene considerato il livello di "soddisfazione del cliente", che costituisce la base per il processo di fidelizzazione e, quindi, di migliore "sfruttamento" del portafoglio clienti; l'output di questo processo è costituito dall'aggiornamento degli indicatori di prestazione delle varie campagne sopra indicate, che permettono di monitorare lo sviluppo degli obiettivi indicati nel piano operativo di marketing&vendita.

Il portafoglio dei processi ruota intorno al *Customer DataBase*, cioè all'archivio dei dati e delle informazioni che registrano sia i dati identificativi del profilo del cliente (Figura 2) utili per la sua analisi (*Customer Intelligence*) sia i risultati dei vari tipi di relazione fra azienda e cliente. Il Customer Data Base si articola in 5 sezioni:

- *Dati anagrafici*: nome e cognome, numero cellulare, numero telefono fisso, indirizzo postale, *e-mail* ecc..
- *Dati sociodemografici*: data di nascita, sesso, professione, componenti della famiglia ecc..
- *Dati economici e comportamentali*: prodotti e servizi della marca posseduti, ammontare di ogni acquisto, data ultimo acquisto e frequenza di acquisto nell'ultimo anno, spesa media con la marca negli ultimi due anni, tipo e numero di richieste pervenute al Contact Center.
- *Dati motivazionali*: modalità di utilizzo del prodotto/servizio, propensione alla sostituzione del prodotto/servizio, eventuali prodotti concorrenti considerati, soddisfazione del cliente/utente ecc..
- *Dati relazionali*: fonti di conoscenza della marca e del prodotto/servizio, chi decide sull'acquisto, chi paga, chi usa ecc..

Vediamo in che modo possiamo sostenere i processi di marketing&vendita e il Customer Data Base con una piattaforma di CRM adeguata.

Possiamo identificare due grandi momenti nella relazione con il cliente:

1. Customer Care e Customer Service: l'insieme delle relazioni legate alla "manutenzione

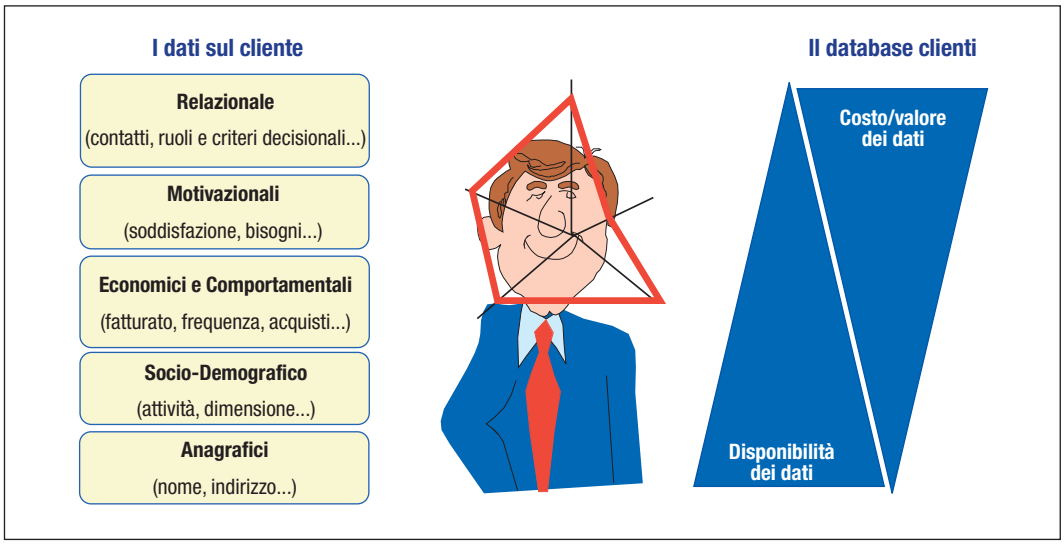


FIGURA 2
Il profilo del cliente
(Fonte: COSMO)

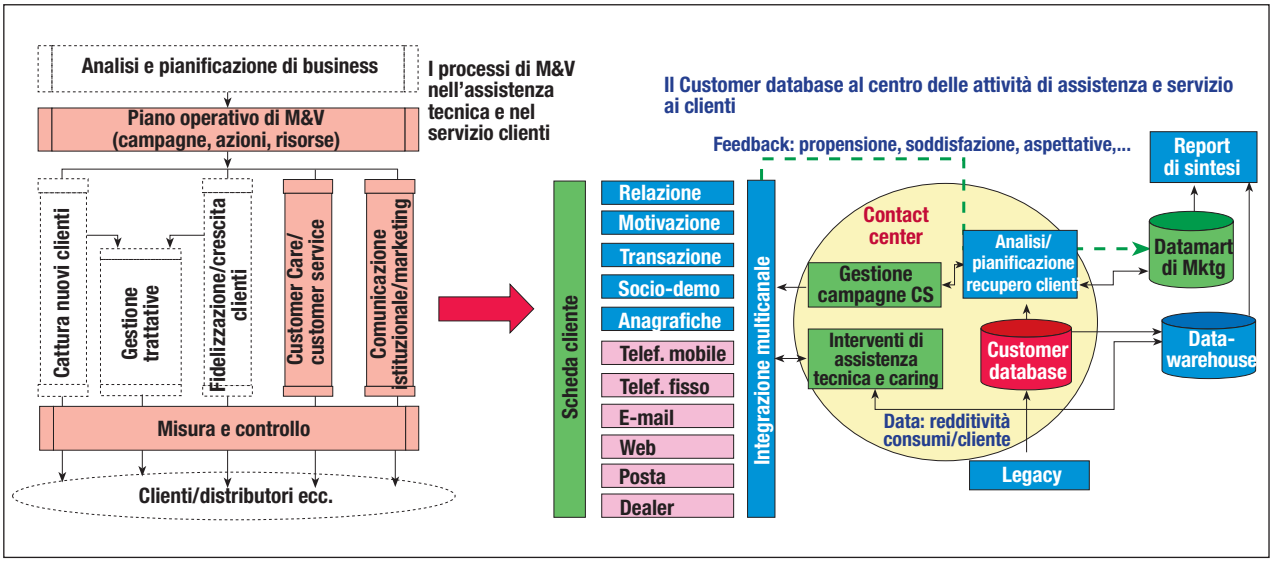


FIGURA 3

La configurazione di una piattaforma CRM adeguata a sostenere i processi di Customer Care e Customer Services (Fonte: Allaxia, Metodologia Cosmo™)

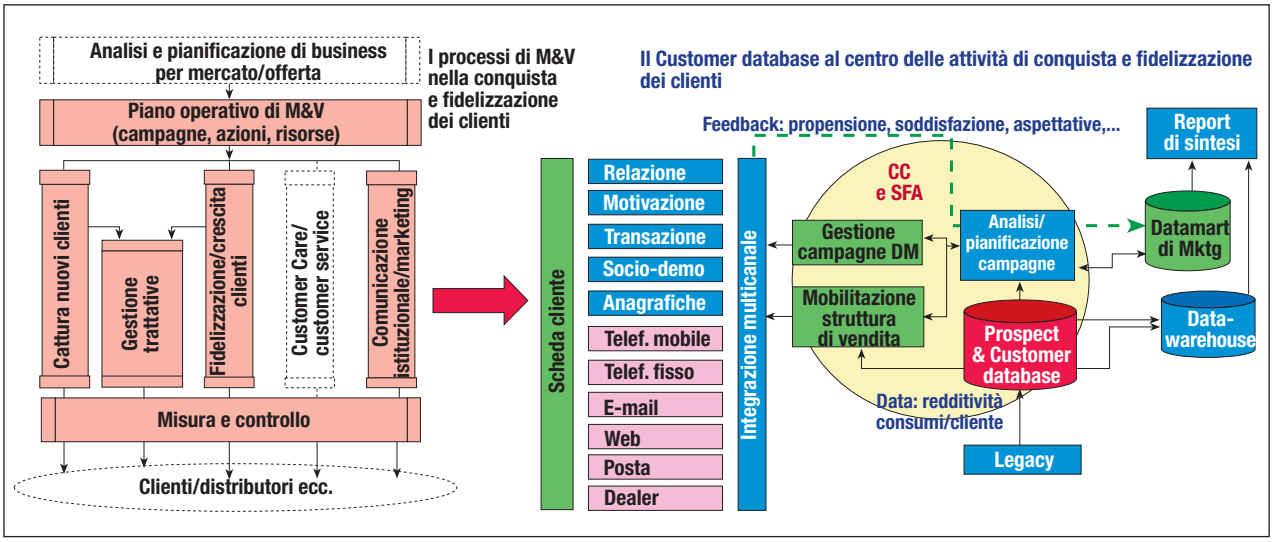


FIGURA 4

La configurazione di una piattaforma di CRM adeguata a sostenere i processi di conquista e fidelizzazione del cliente

del cliente" in termini di assistenza tecnica e servizi informativi e di supporto, che sono mirati a sostenere o migliorare la "soddisfazione del cliente", base essenziale perché al momento del nuovo acquisto il cliente si orienti di nuovo sulla marca già "provata"; tali relazioni sono sostenute essenzialmente dai tecnici dell'assistenza e dagli operatori del Contact Center (Figura 3).

2. Acquisizione e Fidelizzazione: l'insieme delle relazioni legate alla promozione e alla vendita di prodotti/servizi o nuovi o sostitutivi di quelli già acquistati; tali relazioni sono sostenute essenzialmente dai venditori dell'azienda e ancora dagli operatori del Contact Center (Figura 4).

Nel primo caso, i processi coinvolti sono quelli del Customer Care e Customer Service, con l'appoggio della comunicazione; la piattaforma di Contact Center permette a tutti i ruoli coinvolti di agire nei confronti del cliente conoscendo quali tipi di relazioni si sono svolte e aggiornare il suo profilo con quanto emerge per ogni contatto; il cliente si sente riconosciuto e, se il servizio di assistenza è puntuale e di qualità, anche "ben servito". I clienti la cui soddisfazione è andata al disotto dei livelli ritenuti critici per la fidelizzazio-

ne del cliente, devono essere assistiti con una cura particolare attraverso campagne apposite, prima di poter di nuovo essere sollecitati a un nuovo acquisto.
Nel secondo caso, i processi coinvolti sono quelli della conquista del cliente, della fidelizzazione e della trattativa, con l'appoggio della pubblicità e della comunicazione aziendale; il venditore dell'azienda o quello del *dealer* possono venire aiutati nel loro lavoro di identificazione di un nuovo cliente attraverso campagne di *leads qualification* cioè di selezione, nell'universo dei potenziali clienti (o prospect), di quelli che si dimostrano sensibili, o meglio ancora interessati a una proposta; la fornitura al venditore di una lista di prospect, magari anche accompagnata dall'indicazione di una data e luogo di appuntamento, costituisce un importante *support* al venditore in termini di risparmio di tempo e di aumento della probabilità di successo per una trattativa.
Se poi è disponibile nel Customer Data Base la data dell'ultimo acquisto, in funzione del ciclo di vita del prodotto, è possibile lanciare, per i clienti soddisfatti, una campagna di promozione per la sostituzione del prodotto/servizio con uno nuovo; anche in questo caso, è la combinazione dell'attività dell'operatore di Contact Center con quella del venditore supportato da strumenti di Sales Force Automation che può produrre un risultato più efficace, sia in termini di produttività dell'azione commerciale (minore tempo per trattativa andata a buon fine), che di redditività del cliente (allo stesso cliente viene venduto di più nello stesso arco di tempo).
In questo caso, i processi di vendita vengono supportati dalla piattaforma di CRM, comprendente sia funzionalità di Contact *Center in outbound* che di SFA-*Sales Force Automation*, come indicato nello schema, in cui vengono messe in evidenza le campagne di Direct Marketing, il Piano di Marketing e la strumentazione di automazione delle attività di vendita.
Completiamo ora l'analisi con le varie figure professionali necessarie al funzionamento del Sistema Strutturato e Operante di Relazione con il Cliente, di cui dobbiamo identificare l'unità organizzativa di appartenenza, i ruoli e le responsabilità.

Le varie figure professionali comprendono:
- Nella Direzione Marketing e Pubblicità:
  - Il *Marketing Manager*: ha l'obiettivo critico a livello aziendale di mantenere o migliorare la quota di mercato; spinge le soluzioni CRM come investimento strategico per perseguire il suo obiettivo critico; è responsabile della impostazione e poi realizzazione del Piano di Marketing e degli obiettivi relativi.
  - L'*Offer Manager*: ce ne possono essere più di uno all'interno di un'unica azienda; ognuno ha l'obiettivo critico di assicurare il massimo obiettivo di vendita dell'offerta di cui è responsabile, d'accordo con la capacità della Produzione di sostenere la disponibilità dei volumi previsti.
  - Il *Responsabile della Pubblicità e della Comunicazione*: ha l'obiettivo critico di promuovere l'immagine aziendale e, per quanto riguarda i prodotti, di aumentarne la visibilità e il grado di conoscenza da parte di clienti e di prospect in modo da favorire l'azione di vendita sui vari canali; per sviluppare azioni di Direct Marketing ha la necessità assoluta di avere disponibile un Contact Center, o interno o in *outsourcing*.
- Nella Direzione Vendite
  - Il *Responsabile Vendite*: ha l'obiettivo critico di aumentare le vendite e il fatturato, o almeno il primo margine di contribuzione, a livello aziendale; spinge la soluzione CRM in particolare per quanto riguarda la soluzione di SFA, che permette di migliorare la produttività dei commerciali e monitorare la redditività per singolo cliente.
  - Il *Venditore* (o *Sale Representative*): ha l'obiettivo critico di aumentare le vendite e il fatturato sul portafoglio clienti assegnato; utilizza la soluzione SFA che gli permette di migliorare la propria produttività dei commerciali e monitorare l'andamento del proprio fatturato e, quindi, i propri incentivi.
- Nella Direzione Customer Care:
  - Il *Responsabile del Contact Center*: ha l'obiettivo critico di migliorare il servizio ai clienti, di sviluppare le campagne di direct marketing e di presidiare la misura della customer satisfaction; gestisce il Contact Center come un'unità di produzione di servizi per l'azienda, cercando di introdurre sistemi di tipo industriale nei vari processi

di supporto al marketing e alle vendite che è chiamato a presidiare.

- L'*Operatore di Contact Center*: ha l'obiettivo critico di mantenere e sviluppare una relazione di tipo fiduciario con il cliente con cui entra in contatto.

❑ Nella Direzione Assistenza Tecnica:

- Il *Responsabile della Logistica e Assistenza Tecnica*: ha l'obiettivo critico di migliorare i tempi e i costi della logistica (consegne) e la qualità delle installazioni e dell'assistenza tecnica ai clienti; pianifica e gestisce, sulla base delle richieste pervenute dallo SFA, il programma degli interventi a esecuzione delle installazioni, mentre pianifica e gestisce, sulla base delle richieste pervenute dal Contact Center, gli interventi di soluzione di malfunzionamenti del prodotto/servizio installato/acquistato dal cliente.
- Il *Tecnico di Assistenza*: ha l'obiettivo critico di migliorare i tempi e la qualità degli interventi presso il portafoglio dei clienti assegnato; riceve dal Contact Center le indicazioni per eventuali interventi a soluzione di malfunzionamenti del prodotto/servizio installato/acquistato dal cliente, alimenta il CC e/o lo SFA con le indicazioni sugli interventi eseguiti.

❑ Nella Direzione Sistemi Informativi:

- Il *Responsabile Sistemi Informativi*: ha la responsabilità critica del funzionamento ottimale del Sistema Informativo aziendale, nel cui ambito è gestita la soluzione CRM.
- Il *Responsabile CRM*: ha l'obiettivo critico di ottimizzare la gestione del complesso sistema di supporto alla gestione delle relazioni con il cliente, dovendo interfacciare un numero assai rilevante di tipi di utenza, compresa quella del cliente finale nel caso in particolare di società di servizi.

❑ Il *Responsabile Pianificazione e Controllo*: ha l'obiettivo critico di impostare il Piano operativo aziendale e di presidiarne il monitoraggio con gli adeguati indicatori di prestazione sui vari livelli di responsabilità richiesti a livello aziendale; il monitoraggio del sistema di indicatori previsti nel Piano di marketing&vendite rientra fra le sue responsabilità.

❑ Il *Direttore Generale*: ha l'obiettivo critico di sviluppare e gestire il Piano Integrato di sviluppo aziendale, di cui il Piano di marketing&vendita ne è ovviamente un componente essenziale.

Rispetto ai processi di marketing&vendita identificati nel modello possiamo cercare di identificare le responsabilità di ciascun ruolo rispetto al risultato che si vuole ottenere dal processo stesso.

La figura 5 illustra in modo sintetico il contributo che ciascuna Funzione Organizzativa e, nel suo ambito, ciascun ruolo, fornisce rispetto ai processi in cui è coinvolto.

Per quanto riguarda il processo di pianificazione delle strategie di business e di marketing, l'istruttoria viene fatta dal responsabile marketing, con la collaborazione dei Responsabili della Offerta, della Pubblicità, della Vendita, della Logistica e del Sistema Informativo, con il supporto della Pianificazione e Controllo: il piano deve essere discusso e approvato dalla direzione generale.

Il Piano Operativo di Marketing viene sviluppato in termini di istruttoria dall'Offer Manager con la collaborazione dei responsabili della Pubblicità, del Contact Center e dei Sistemi Informativi, oltre che con il contributo dei venditori e dei tecnici dell'assistenza: il Piano deve essere approvato dal responsabile Marketing, in coerenza con le linee del Piano Strategico e di Business.

L'operatività delle azioni di vendita è sulle spalle dei venditori, con la collaborazione degli operatori del Contact Center dedicati all'*inbound* e del CRM Manager dei sistemi informativi, sotto la responsabilità del Direttore Vendite.

L'operatività del Customer Care e del Customer Service è sulle spalle degli operatori del Contact Center e dei Tecnici di Assistenza, sempre con il supporto del CRM Manager dei Sistemi Informativi.

L'operatività della Pubblicità e Comunicazione è sulle spalle delle agenzie di pubblicità selezionate dal responsabile e per la promozione in termini di Direct Marketing, su quelle degli operatori di Contact Center dedicati all'outbound.

L'operatività della Misura e Controllo è, infine, sulle spalle della Pianificazione e Controllo con la collaborazione di tutte le altre funzioni che devono fornire i relativi input per avere poi indietro i rispettivi indicatori di pre-

stazione aggiornati in modo da permettere la valutazione della distanza fra i risultati effettivi delle campagne e azioni varie e gli obiettivi definiti a livello di Piano.
Completiamo l'analisi delle 3 componenti di uno SMaRC con quella della piattaforma tecnologica che integra le 3 componenti del CRM (CRM Collaborativi, CRM Operazionale e CRM Analitico); la descrizione che segue identifica i vari componenti costitutivi in termini assolutamente generici.
I componenti del CRM Collaborativo, che supporta il contatto con i clienti, sono:

- PSTN, *Public Service Telecommunication Network*. È l'interfaccia con la rete di telecomunicazione pubblica.
- IP, *Internet Protocol*.
- PABX, *Private Branch eXchange*. È il componente centrale dell'architettura di un Contact Center. Un PABX controlla ogni *teleset* (telefono digitale) sulla postazione di ogni agente. Oltre a dirigere le chiamate sul corretto DN (*Directory Number*), realizza diverse funzioni tra cui:
- ACD, *Automatic Call Distribution*. È un meccanismo basato sul concetto di coda (coda ACD) dove la coda è costituita da un gruppo di DN appartenenti a operatori che realizzano gli stessi *task* o sequenze di task all'interno del Contact Center.
- IVR o VRU, *Interactive Voice Response*. È un componente che si interfaccia con il PBX. È dedicato a raccogliere informazioni preliminari o a eseguire processi completi su una chiamata in ingresso. In pratica, è il sistema che interagisce con il chiamante mediante una voce pre-registrata, con la quale è possibile effettuare delle scelte tra diverse opzioni.
- *Predictive Dialer*: è il componente che permette di riconoscere il numero del chiamante delle telefonate in inbound.
- *E-mail*: è il componente che permette la produzione e l'invio di posta elettronica e la sua ricezione.
- *Web*: è il componente che permette la navigazione sul portale aziendale e nello specifico permette l'accesso alle funzionalità di SFA.

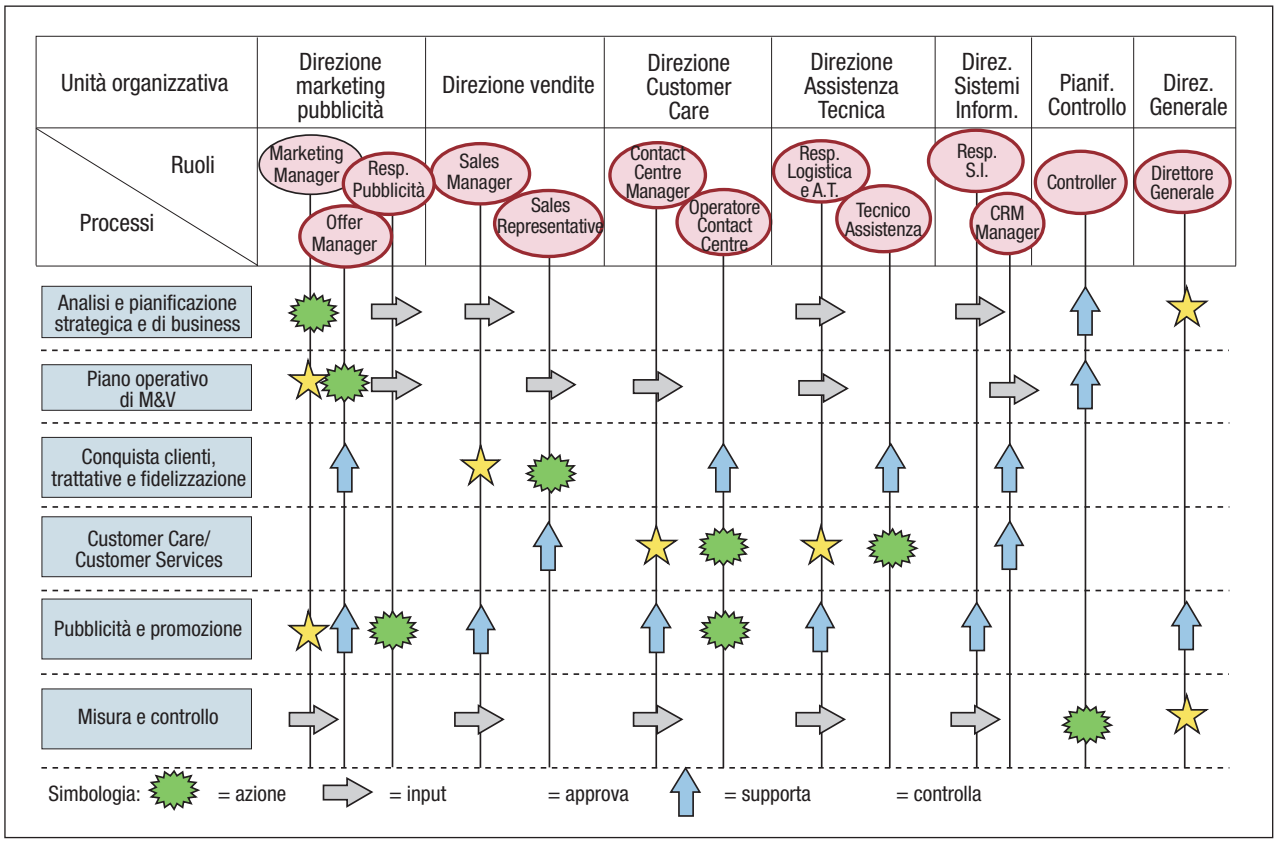


FIGURA 5

I principali ruoli e responsabilità nei processi di Marketing&Vendita supportati da una configurazione di piattaforma CRM (Fonte: COSMO)

Per la Gestione della Piattaforma sono presenti una *Console Real Time* per gli interventi sul sistema e un *Hystorical Reporting* per il tracciamento dello stato del sistema; entrambi permettono la Gestione dell'Interazione con i Clienti (*Customer Interaction Management*), basato su un sistema di Regole di Accodamento del Mix dei Media di contatto (*Business Roules Multimedia Queuing Blending*) che comprende:

- *Inbound Call Routing*: é il componente che permette l'instradamento della chiamate in ingresso al Contact Center verso il primo operatore libero.
- *Outbound Call Management*: è il componente che permette l'instradamento delle chiamate in uscita dal Contact Center verso il numero di telefono fisso o mobile, o di fax o l'indirizzo di e-mail, o quello postale, del nominativo indicato dall'operatore.
- *E-mail e Web Response Management*: sono i componenti che permettono la gestione delle risposte rispettivamente via e-mail e sul Web.

I componenti del CRM Operazionale, che supportano le attività degli Operatori del Contact Center, sono:

- Il *Browser Desktop*: permette la navigazione sulla Intranet aziendale.
- Gli *Scripted Workflow*: indicano il percorso da fare per ogni possibile richiesta dei clienti o tipo di campagna da lanciare.

- *Front Office*: è lo strumento che rende disponibile all'operatore l'identificazione e i dati di profilo del cliente chiamante e, quindi, di interagire con il cliente conoscendo tutti i suoi precedenti rapporti con l'azienda.
- *Applicazioni a supporto degli Operatori*: comprendono tutte le applicazioni aziendali di cui gli Operatori possono avere bisogno per gestire in modo adeguato la relazione con il cliente.

I Venditori, che invece hanno bisogno di essere supportati nelle attività relative alla conquista e fidelizzazione dei clienti, hanno a disposizione i seguenti componenti, propri di una piattaforma di Sales Force Automation, sempre nell'ambito del CRM Operazionale:

- L'*Agenda*, su cui possono pianificare i loro incontri e i contatti con la clientela; l'agenda di solito è condivisa con altri addetti aziendali in modo da permettere, ad esempio al tecnico dell'assistenza e/o all'operatore di Contact Center, di conoscere e utilizzare, in caso di contatto con il cliente, l'informazione relativa al prossimo appuntamento pianificato dal venditore.
- La *Scheda Cliente*, in cui sono disponibili tutti dati e le informazioni rilevanti del cliente sia in termini di profilatura che di eventuale segmentazione, completata dalle indicazioni sui rapporti in essere (prodotti acquistati, ammontare speso, eventuali insoluti, eventuali richieste di supporto al Contact Center ecc.); nel caso che si tratti ancora di un Prospect, i dati sulla scheda saranno sicuramente molto scarsi, ma in funzione del livello di qualificazione del Prospect (cioè dello stato di avanzamento dei contatti e quindi degli interessi/bisogni rilevati), potranno essere indicati sia il suo "potenziale" di acquisto che le principali opportunità da proporgli.
- La *Scheda Campagne*, su cui sono indicate le campagne promozionali e di comunicazione pianificate e su cui è richiesto al venditore di pianificare specifiche attività di supporto o di vendita.
- Le *Indicazioni di Budget*, per cliente e in generale, su cui viene valutata e riconosciuta la parte variabile del suo stipendio e l'andamento effettivo delle acquisizioni mese per mese.
- La *Scheda Offerta*, con le indicazioni delle offerte disponibili e i relativi listini prezzi.
- La *Scheda Concorrenza*, con l'indicazione dei principali concorrenti per ciascuna offerta, con l'indicazione dei punti di forza/punti di debolezza rispetto alla propria offerta.
- Il *Cruscotto Personale* con l'indicazione delle prestazioni obiettivo (ordini, fatturato, incassato ecc.) e di quelle reali progressive.

Infine, per quanto riguarda i componenti del CRM Analitico, che permettono al Market Research Manager di analizzare e monitorare i comportamenti del mercato e dei vari segmenti che lo costituiscono e a ciascun Venditore di analizzare e monitorare il comportamento dei clienti che ha nel proprio portafoglio, sono disponibili:

- il *Customer Data Base*, che comprende tutti i dati sui prospect e sui clienti, a partire dai dati anagrafici fino ai comportamenti di acquisto, alla identificazione dei loro bisogni e alle relative motivazioni;
- la funzione di *Business Intelligence*, basata sul Datamart di marketing&vendita, che se-

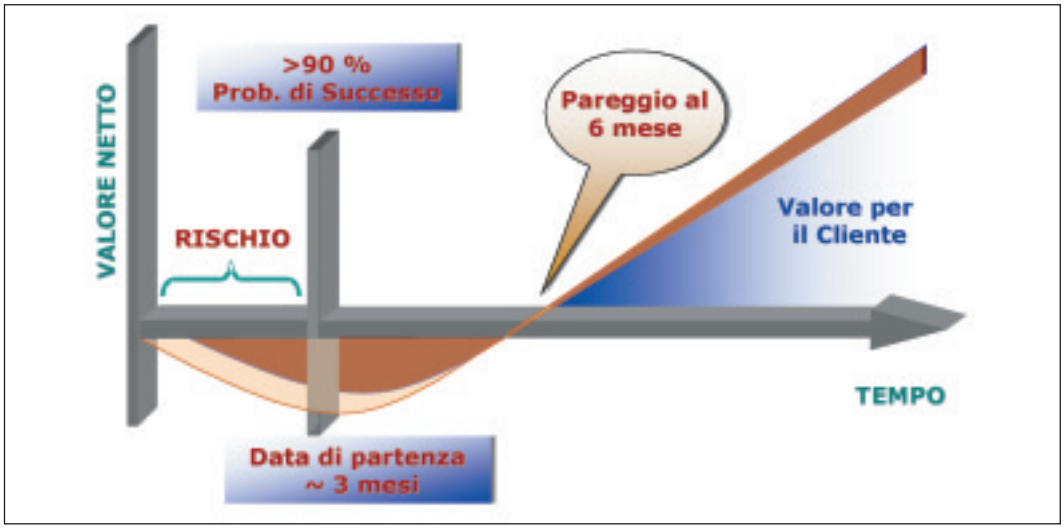


FIGURA 6
Stima del ROI di una soluzione CRM come "information utility"

leziona dalle varie Applicazioni di impresa e dai Sistemi Legacy i dati necessari per questo tipo di analisi e li elabora attraverso strumenti di Olap/Data Mining;

- la funzione di *Customer Portfolio Analysis*, sempre basata sul Datamart di M&V, che comprende *tool* per elaborare il grado di concentrazione dei clienti e analizzare le loro *performance* dal punto di vista dell'azienda;

4. IL CAMBIAMENTO DI SCENARIO DELLA TECNOLOGIA CRM

Come abbiamo visto, la complessità delle funzioni d'uso che una piattaforma di CRM integrata deve mettere a disposizione per una soluzione che supporti l'insieme dei processi di marketing&vendita, è molto alta. Tale complessità si traduce, inoltre, nella difficoltà di portare a lavorare in modo sincrono, rispetto alle esigenze del cliente, una serie di ruoli aziendali abituati a muoversi, invece, con modalità totalmente asincrone; un'indagine dell'Economist del 2002 riporta a questo proposito che il numero degli insuccessi, e quindi il rischio che l'implementazione di sistemi operanti di CRM fallisca, è significativamente alto. Le analisi condotte da vari specialisti sugli investimenti in piattaforme di CRM disponibili sul mercato, mostrano che è in corso una consistente riduzione dei costi di investimento, attraverso due strade parallele.

Intanto, come per tutte le tecnologie innovative, assistiamo a un forte decremento della componente tecnologica del sistema operante, limitato alle componenti di *hardware, software e middleware*: Nucleus Research stimò, nel 2002, che insieme a una soluzione Siebel, considerato il leader in questo settore per i sistemi complessi e completi che devono equipaggiare almeno alcune centinaia di utenze, quotabile intorno ai 18.000 dollari per costo annuo di una singola utenza, sarebbero diventate disponibili dal 2003 anche soluzioni più semplificate a un costo annuo per posto di utenza non superiore a qualche migliaio di dollari.

In secondo luogo, emerge anche nella fornitura di soluzioni di CRM, la possibilità di forniture di "piattaforme di servizio" di CRM vista come *Information Utility* (Figura 6). Gartner Group, in proposito, applica il concetto di *Return On Investement* (ROI) per valutare l'opportunità dell'introduzione di una piattaforma CRM a supporto del sistema operante di marketing&vendita: tenendo conto del costo totale per la sua implementazione (non solo della componente tecnologica, ma anche dell'attività di progettazione, avviamento e formazione del personale coinvolto), della probabilità che il sistema fallisca (o rischio di interruzione del progetto) e del tempo necessario per rientrare dell'investimento valutando il differenziale di redditività della clientela e di produttività commerciale senza e con un sistema operante di CRM, Gartner dimostra che un CRM tradizionale sviluppato *in house*, visto come una piattaforma tecnologica complessa, ha alti costi fissi iniziali e di manutenzione, una

probabilità di successo non superiore al 50%, e un tempo relativamente lungo per raggiungere il *break-even* (27 mesi), con la conseguenza di avere un effetto minimo sul valore totale per il cliente.
Se, invece, il CRM viene visto come una Information Utility, questa ha costi iniziali e di manutenzione più bassi, variabili e prevedibili con certezza poiché commisurati all'uso che ne viene fatto. Inoltre, ha una probabilità di successo di almeno il 90% e un *break-even point* decisamente più breve (intorno ai 6 mesi), con la conseguenza di avere un significativo effetto sul valore totale per il cliente.
Le implicazioni di questa analisi sono più chiare se si considerano le varie voci che contribuiscono alla formazione del costo di investimento o al costo d'uso nella due alternative indicate; dalla tabella 1 si vede come, alle voci di costo relative al personale, alla tecnologia e alla gestione del processo, della soluzione come Information Utility, comunque presenti, si aggiungano nella soluzione in house (tradizionale), i costi della progettazione, realizzazione e messa in opera del sistema operante, a cui si devono aggiungere poi i costi di manutenzione.
Naturalmente questo non significa che i costi di progettazione e realizzazione non ci siano: in una soluzione come Information Utility questi costi sono a carico del Fornitore della "utility di CRM", come d'altra parte avviene per ogni fornitura di questo genere, e il fornitore poi li ribalta in termini di "prezzo d'uso" sull'insieme dei clienti, dove ogni cliente paga per l'uso di ciascuno dei posti di lavoro equipaggiati dalla utility.
In tal modo, l'obiettivo del fornitore diventa quello di remunerare il proprio costo di gestione e l'ammortamento del proprio investimento su un numero di "posti di lavoro" tale da permettergli di arrivare a un punto di break even, al di sopra del quale possa cominciare ad avere un margine di contribuzione positivo.
Entrambe le soluzioni sono valide: quale che sia quella da scegliere diventa un esercizio di valutazione della convenienza fra *make or buy* che ciascuna azienda può fare come per qualunque altro tipo di acquisto come, ad esempio, le auto della flotta aziendale, le mura di uno stabilimento di produzione ecc., tutti beni o acquisibili come investimento proprio oppure affittabili con un contratto di gestione "tutto incluso" da un gestore specializzato.
In conclusione, dalla combinazione delle due linee di tendenza indicate, emerge che il prezzo a cui è possibile equipaggiare un singolo posto di utenza sia dell'ordine di 1.000 dollari a partire dal 2003-2004.

5. IL MERCATO ITALIANO DEL CRM

La dizione CRM, per i fornitori di tecnologia, si è ormai estesa a comprendere tutte le componenti che hanno un'influenza anche

TABELLA 1
Le componenti di costo di un progetto per la realizzazione di uno SMaRC nella soluzione tradizionale "in house" e nella soluzione "information utility"

Tipo di soluzione CRM	Costo delle persone	Costo della tecnologia	Costo di gestione
	Costi comunque presenti		
Come Information Utility	• Commerciali • Marketing	• Integrazione • Pulizia dei dati	• Strategia
"In house", di tipo tradizionale	Costi addizionali		• Legale
	• Personale interno, del Sistema Informativo e del Fornitore • Supporto interno ed esterno • Formazione • Competenze tecniche	• Software • Middleware • Hardware (Servers, Storage Devices) • Manutenzione del sistema e della Applicazione • Aggiornamenti del sistema e della Applicazione • Accessori • Telecomunicazioni	• Competenze • Vendite • Marketing • Servizio • Finanza Corporate • Contabilità • Risorse umane

Anni	2000	2005	Variazione %	Variazione % Anni
CRM e Call Center	151,1	858,9	468,4	93,7
SFA	19,3	168,9	775,5	155,1
DW & BI	83,6	212,3	154,0	30,8
Totale area ICM	**254,0**	**1.240,0**	**388,2**	**77,6**

TABELLA 2
Le previsioni di spesa per il software applicativo nell'area ICM in Italia (Fonte: Assintel/Sirmi)

indiretta nella gestione del cliente: se includiamo come supporto ai processi di marketing&vendita non solo le componenti tecnologiche che permettono un contatto diretto con il cliente, ma l'integrazione completa di tutte le tecnologie che concorrono a gestire le relazioni con i clienti, inclusi il *Data Warehouse* (DW) e il *Data Mining* (DM) e i processi di *back office* e di *front office*, questo insieme viene indicato con il termine di *Integrated Customer Management* (ICM).
Per quanto in Italia la domanda di soluzioni ICM sia ancora relativamente limitata, le aspettative di sviluppo si presentano molto interessanti già a breve periodo: la stima è che il mercato per le soluzioni ICM cresca dai 254 miliardi del 2000 agli oltre 1.240 stimati per il 2005, con un incremento annuo del 77% (Tabella 2). Dove si concentra questo sviluppo? Mentre le grandi imprese riducono gli investimenti in piattaforme ICM troppo complesse e cercano di risolvere le difficoltà di implementazione adottando un approccio più graduale, il numero dei potenziali utenti (o posti di lavoro) nella Piccola e Media Impresa è enorme. Microsoft ha "visto" con interesse l'opportunità e sta entrando sul mercato con una offerta da 79,00$ per posto di lavoro al mese. Limitandosi solo alla offerta di soluzioni CRM, è interessante valutare anche il mercato potenziale che possono avere tali soluzioni per l'Italia; se adottiamo la definizione di PMI proposta da Customer Relationship Management for Small and Medium-size Enterprises[1] in una recente ricerca di mercato identifichiamo come:

- *Piccole Imprese*, quelle che hanno da 1 a 10 addetti (fra personale di marketing, venditori, addetti ai servizi ecc.) che già usano o potrebbero usare utilmente un sistema CRM.
- *Medie Imprese*, quelle che hanno da 11 a 100 addetti (fra personale di marketing, venditori, addetti ai servizi ecc.) che già usano o potrebbero usare utilmente un sistema CRM.
- *Grandi Imprese*, quelle che hanno più di 100 addetti (fra personale di marketing, venditori, addetti ai servizi ecc.) che già usano o potrebbero usare utilmente un sistema CRM.

L'elaborazione dei dati con le definizioni indicate porta a identificare per il mercato USA:

- Un mercato di Piccole Imprese, configurabile in una dimensione di circa 9,1 mln, che esprime un potenziale di circa 20,5 mln di Posti di Lavoro equipaggiabili con CRM (PdL CRM, che quindi già usano o potrebbero usare utilmente).
- Un mercato di Medie Imprese, configurabile in una dimensione di circa 340 mila, che esprime un potenziale di circa 11,4 mln di PdL CRM.

Sviluppando lo stesso tipo di analisi a livello Italia (ancora su dati Istat 1996, ma non dovrebbero cambiare sostanzialmente quando saranno disponibili i nuovi dati Istat 2001), si arriva a una segmentazione del mercato dei PdL CRM come indicato in tabella 3.
In sintesi, emerge per l'Italia un mercato configurato in:

- 3,3 mln di Piccole Imprese che esprimono un potenziale di circa 5,9 mln di PdL CRM;
- 22,6 mila Medie Imprese, che esprimono un potenziale di circa 0,86 mln di PdL CRM;
- 1,8 mila Grandi Imprese, che esprimono un potenziale di circa 1,0 mln di PdL CRM;

In altre parole, il mercato delle Piccole e Medie Imprese in Italia ha un potenziale di circa 6,8 mln di PdL, che al valore unitario, tanto per fare un esempio, di 1.000 € all'anno, fanno circa 10,5 mila milioni di €!

[1] www.crm4sme.com

N° addetti	N° di imprese Italiane	PdL potenziali per impresa	PdL potenziali totali	
Piccole	**3.350.028**	**1,8**	**5.959.446**	
da 1 a 9	3.168.564	1,5	4.622.965	
da 10 a 49	181.464	7,4	1.336.481	
Medie	**22.684**	**38,1**	**865.388**	
da 50 a 99	13.368	32,9	439.464	
da 100 a 199	6.010	27,9	167.525	
da 200 a 499	3.306	78,2	258.399	
Totale piccole e medie	**3.372.712**	**2,0**	**6.824.835**	**54,9%** del totale addetti di imprese fino a 500
grandi, oltre 500	1.811	561,7	1.017.261	**20,0%** del totale addetti di imprese con più di 500
Totale generale	**3.374.523**	**2,3**	**7.842.096**	**44,8%** del totale addetti universo

TABELLA 3

Il mercato potenziale di Posti di Lavoro equipaggiabili con CRM in Italia (Fonte: ISTAT - Censimento intermedio Industria - 1996, elaborazione R. Bellini)

6. I BISOGNI DI SUPPORTO DI MARKETING DELLE RETI DISTRIBUTIVE GOVERNATE E L'OPPORTUNITÀ PER SOLUZIONI SMARC E SMART

Una volta identificato il potenziale di mercato delle piattaforme ICM nelle Piccole e Medie Imprese, possiamo cercare di capire come si possa arrivare a dotare tali imprese di quello che abbiamo chiamato Sistema strutturato e operante di Management della Relazione con il Cliente (SMaRC).

La soluzione, infatti, non è semplice: non si tratta solo di un problema di tecnologia disponibile a prezzi accettabili, ma i fallimenti dei progetti condotti da medie e grandi imprese hanno chiarito che, oltre all'aspetto relativo alla "tecnologia disponibile", per lo sviluppo di un Sistema SMaRC ci sono altri problemi e, in particolare, come mostrano i risultati di un test di una ricerca condotta in ambito MIP-Politecnico di Milano sul tema dei Fattori Critici di Successo di Progetti CRM in Italia (Project Work Corso MRC-2002):

- è difficile far riconoscere all'azienda che sia necessaria una nuova *vision* aziendale, quella che pone il cliente al centro dell'attenzione di tutte le strutture aziendali;
- è difficile riuscire a definire correttamente quale sia inizialmente il budget reale; spesso, a fronte di un budget iniziale, si scopre che sono necessarie più revisioni prima di poter considerare il progetto operativo e adeguato rispetto agli obiettivi posti e alle aspettative;
- la difficoltà di definire il budget iniziale porta in parallelo una difficoltà nella definizione dei tempi reali di cui il progetto ha bisogno; spesso il tempo effettivo supera di molto il tempo inizialmente previsto;
- ancora, è molto difficile misurare in anticipo il valore dei risultati ottenibili, per cui diventa molto difficile anche misurare i risultati realmente ottenuti a fronte dell'investimento sostenuto;
- resta, infine, difficile ottenere sia l'implementazione dei nuovi processi di marketing&vendita che il cambiamento culturale delle risorse coinvolte.

In altre parole, il problema che si trovano(o troveranno) a fronteggiare le imprese nella realizzazione del proprio progetto SMaRC è molto legato agli aspetti relativi al cambiamento introdotto da un più forte orienta-

mento al cliente e a quello derivante dall'adattamento a soluzioni operative diverse rispetto a quelle preesistenti alla introduzione dello SMaRC; si evidenzia la necessità di guidare il cambiamento con un doppio livello di presidio:

1. il presidio sul piano della implementazione della soluzione tecnologica;

2. il presidio sul piano della implementazione del nuovo modello operativo sostenuto dalla soluzione tecnologica.

In particolare, questo secondo aspetto è quello su cui la Piccola e Media Impresa ha maggiori difficoltà, per l'insufficienza, strutturale, di personale dedicato alle attività di supporto interno professionale.

Un modo adeguato per sviluppare soluzioni SMaRC per la Piccola Impresa è quello di guardare alle catene distributive costituite da reti di piccole imprese commerciali di prodotti/servizi di grandi imprese, cioè alle reti di imprese del *trade* o del cosiddetto "canale indiretto".

Troviamo catene distributive di questo genere in molti settori merceologici, come ad esempio:

1. nel settore dell'informatica; sono stimate in almeno 5.000 le imprese specializzate che operano nel settore rivendendo e aggiungendo valore ai prodotti hardware e ai pacchetti software messi sul mercato dai grandi fornitori;

2. nel settore della telefonia fissa e mobile; solo i 4 grandi operatori di Telecomunicazioni (Telecom/Tim, Vodafone, Wind e 3) gestiscono le loro 4 reti per un totale di almeno 2.500-3.000 imprese;

3. nel settore dell'auto si stimano di nuovo in circa 10.000 le concessionarie auto che presidiano il mercato dei circa 20 grandi fornitori;

4. ci sono ancora le reti delle varie Compagnie di Assicurazioni, le reti di distribuzione di materiali elettrici e di quelli per la casa ecc..

Diciamo che queste reti sono "governate" non tanto e non solo perché sono legate da un contratto di commercializzazione con la casa fornitrice, che oltretutto spesso ha dimensioni più grandi delle singole imprese che commercializzano i suoi prodotti/servizi, ma piuttosto perché sono finalizzate dallo stesso obiettivo che insiste sullo stesso cliente: catturare più clienti sia per sè che per il proprio fornitore, servirli bene e "sfruttarli" meglio con campagne di *cross-selling* e *up-selling*. In altre parole, queste reti di imprese sono "governate dal comune interesse a catturare e fidelizzare il comune cliente finale".

Ognuna delle imprese che fa parte di queste reti va considerata sia dal punto di vista della relazione fra il fornitore e l'impresa distributrice, sia da quello della relazione fra l'impresa distributrice e il cliente finale.

Per quanto riguarda la relazione fra il fornitore e l'impresa distributrice (*trader*) l'obiettivo del fornitore (G.Dellisanti, News Letter Omar 98) nei confronti di un operatore del trade è quello di motivarlo a promuovere e a vendere i propri prodotti, anziché quelli dei concorrenti, a target specifici, in volumi crescenti, con qualità trasferita crescente (servizi accessori, customer satisfaction ecc.) e con margini decrescenti.

L'obiettivo del trader è, invece, un po' diverso:

- i concorrenti del fornitore non sono i *suoi* concorrenti, ma sono altri rivenditori, o altri canali di vendita;
- il trader vende i prodotti del fornitore nella misura in cui sono richiesti dal mercato, consentendo di applicare ragionevoli margini e soprattutto non generando eccessivi problemi (e quindi costi) *pre* e *post* vendita;
- il trader si considera il vero specialista del marketing, della vendita e della relazioni con la clientela;
- il trader (evoluto) considera l'innovazione dei canali di vendita un'opportunità (se gestita da lui, per rafforzare il suo ruolo nei confronti del fornitore) mentre la considera una minaccia, se gestita dal fornitore contro di lui.

I due punti di vista richiedono un notevole impegno per farli diventare conciliabili, dove il *gap* di interessi e obiettivi non si colma complicando le formule di sconto o i termini e condizioni di fornitura, che anzi sta ampliandosi a seguito della crescente pressione competitiva "in essere" sia sui fornitori che sui canali distributivi.

La soluzione si trova tenendo conto che il trader è un piccolo imprenditore, con una sua autonomia gestionale che sviluppa nell'ambito di un rapporto con il suo principale forni-

tore di esclusività o quasi esclusività sul prodotto venduto; se da un lato, il fornitore deve rispettare tale autonomia gestionale, a rischio di entrare in rotta di collisione con l'imprenditore, dall'altro lato il rapporto privilegiato fra fornitore e trader impone a quest'ultimo una serie di vincoli su tempi e modi di commercializzazione, ma in cambio può chiedere di essere supportato da una serie di servizi, come per esempio la pubblicità di prodotto piuttosto che la disponibilità di servizi gestionali e/o prodotti nuovi che devono essere lanciati e promossi annualmente e che gli permettono di mantenere e sviluppare il rapporto che il trader stesso ha con la sua clientela locale.

In questo senso, il trader appartiene a una *rete* rispetto alla quale deve individuare tutti i vantaggi per lo sviluppo della sua attività.

Il vero fattore di differenziazione del trader rispetto al fornitore è costituito:

- dalla conoscenza unica che ciascun trader ha sul territorio in cui opera, non tanto in termini geografici ma piuttosto in termini economici sia del mercato del consumatore finale che di quello delle imprese;
- dall'insieme di relazioni che nel corso della sua vita il trader è riuscito a sviluppare e che applica allo specifico tipo di prodotto/servizio sulla cui commercializzazione fonda il proprio benessere economico e i propri obiettivi di successo ulteriore.

Fra queste relazioni sono importanti quelle con la clientela acquisita ma molto di più pesano le conoscenze del contesto in cui opera; se lo guardiamo dal punto di vista dei comportamenti di acquisto del cliente finale, il trader ha una buona conoscenza sia delle caratteristiche dei suoi clienti potenziali e attuali, che dei cosiddetti influenzatori del processo di acquisto.

Per quanto riguarda i clienti attuali, li conosce, anche individualmente, li segue nei loro cambiamenti, intuisce di cosa possono avere bisogno e ne anticipa le richieste con proposte adeguate; per i clienti potenziali, invece, si trova nella faticosa situazione di doverli andare a cercare e coltivare a lungo prima di conquistarli, utilizzando, comunque, il proprio sistema di relazioni come grimaldello per aprire nuove porte.

Il punto di forza del trader è costituito dalla conoscenza degli influenzatori di acquisto, rispetto ai quali vanta una vasta gamma di riferimenti, come possono essere i centri di assistenza tecnica, le scuole di formazione professionale, le società di consulenza locali, le locali associazioni di impresa e professionali, le formazioni politiche e i club culturali locali ecc., e dei suoi concorrenti trader che operano con altre marche.

Conosce, infine, e sa lavorare, sulle linee di cambiamento della geografia sociale ed economica locale, rispetto alla quale può anticipare alcuni suoi comportamenti nel modo di promuovere o integrare la propria offerta, pur fortemente condizionata dalle politiche del fornitore.

In altre parole, la vera conoscenza del trader riguarda il contesto in cui opera la sua azienda e la sua intelligenza viene esercitata nel capire e anticipare i cambiamenti di tale contesto in modo da facilitare la propria attività.

In qualche modo, si ribalta il peso che riconosciamo nei tradizionali processi di vendita delle grandi organizzazioni: in queste, infatti, tutto il peso è centrato sulla strutturazione del processo e sulla costruzione di aggregati che permettono di ragionare in termini medi, senza che queste medie corrispondano, per definizione, ad alcun soggetto reale, mentre nella gestione del trader prevale il peso dedicato alla manutenzione del suo sistema di relazioni unico, assai poco strutturate e strutturabili, ma con una grande attenzione ai contenuti e ai significati delle relazioni stesse accoppiate ai rispettivi portatori e al lavorio di associazione e integrazione di tali coppie contenuto-portatore tali da potergli portare nuovi affari.

Caratteristica di tale attività è molto più l'efficacia dell'attività che la sua efficienza; è molto più importante da questo punto di vista coltivare al tennis una relazione con un "influenzatore" che sviluppare sistematicamente una serie di contatti senza conoscere la natura e le inclinazioni dell'interlocutore che si vuole contattare.

Di cosa ha bisogno il trader per migliorare la sua competitività locale e che il fornitore può aiutarlo ad avere?

Il trader è molto sensibile:

- ai processi di promozione locale, compren-

dendo anche tentativi di inserimento pubblicitario sui media locali; considera stimolante il fatto di far conoscere il nome del suo ufficio o del suo negozio o della sua officina anche nell'ambiente in cui vive dal punto di vista sociale, soprattutto se associabile a quello di un grande marchio noto a livello nazionale o internazionale; su questo specifico aspetto della pubblicità si gioca una prima importante componente della cooperazione dell'impresa locale con l'impresa che opera a livello nazionale;

- ai problemi di efficienza operativa e, quindi, al margine lordo e al profitto e ancora al problema finanziario, e anche su questo può chiedere al fornitore un aiuto per introdurre innovazione di processo nel suo sistema operativo.

Viceversa, il trader è molto poco sensibile e oppone forte resistenza a:

- tutte le richieste di pianificazione e programmazione di marketing: se costretto dal suo principale fornitore a formulare un budget o una campagna, farà molta fatica anche solo a rispondere alle domande poste per formalizzare un piano di attività che lo coinvolga; l'unica misura del successo/insuccesso accettata è quella del fatturato ottenuto rispetto all'obiettivo, su cui misura la propria commissione commerciale;
- tutte le misurazioni che "fanno sprecare tempo e risorse" ma che d'altra parte costituiscono la base sia per il suo autocontrollo sia per il controllo, dall'esterno, da parte del fornitore; ad esempio, il fornitore può introdurre strumenti che permettano al trader sia di verificare i cambiamenti di importanza della clientela acquisita, sia di facilitare l'identificazione di potenziali nuovi clienti; il miglioramento della conoscenza del comportamento della clientela nel tempo permette, infatti, al trader di associare a tali comportamenti i cambiamenti del contesto che conosce in modo non strutturato ma molto bene, per cui riesce poi a formulare delle regole di carattere più generale su cui sviluppare la sua "intelligenza del cliente e del mercato"; acquisire consapevolezza che un cliente importante ha ridotto i suoi acquisti, piuttosto che un piccolo cliente sta aumentando in modo significativo il proprio giro di affari o il proprio reddito viene messo in relazione ad altri elementi di contesto che solo lui conosce e sa interpretare.

Da queste considerazioni nasce la proposta di attrezzare i trader di una rete distributiva con l'introduzione di un doppio sistema tecnologico a sostegno dello SMaRC locale e dello SMaRT-Sistema strutturato e operante per il Management della Relazione con il Trader: il primo permette di costruire e aggiornare il Customer Data Base locale, dove sono registrate tutte le informazioni utili alla gestione del cliente, in comune fra il Trader e il Fornitore; il secondo, invece, è mirato a supportare il processo di marketing&vendita del Trader e, in particolare, del suo personale di vendita; il servizio per il trader consiste nel miglioramento della capacità di intelligence del suo mercato locale, a favore quindi del miglioramento dei risultati di vendita del Trader stesso, mentre l'obiettivo del fornitore è quello di accumulare conoscenza in modo strutturato sul mercato di riferimento di ciascun trader.

Ci sono in particolare due processi di Trade Marketing che possono essere considerati interessanti sia dal fornitore che dal trader, per iniziare la collaborazione in termini di "marketing cooperativo":

- il piano operativo di marketing annuale;
- il piano di campagne di cattura di nuovi clienti e di fidelizzazione dei clienti acquisiti.

Il piano di Trade Marketing avrà maggiori possibilità di successo se chi lo propone (cioè il Trade Marketeer del fornitore) terrà conto dei seguenti criteri;

- sostenere la doppia identità, del singolo trader e del fornitore nazionale, ottenuta integrando le specificità del mercato locale gestito da ciascun trader con gli obiettivi di vendita richiesti dal fornitore, facendo bene attenzione a valorizzare i sistemi di relazione unici che ciascun trader ha sviluppato o sta sviluppando;
- generare consenso e cooperazione della rete di trader nei confronti del fornitore; questo sarà possibile solo se ci sarà stato un coinvolgimento nella formulazione sia degli obiettivi che nelle modalità operative di sviluppo della varie azioni;
- garantire un sufficiente controllo sui risultati critici del piano (e non sulle procedure di applicazione).

Il piano delle campagne di cattura e fidelizzazione avrà maggiore possibilità di successo se sarà finalizzato a mettere a disposizione del trader liste di nuovi clienti potenziali (generazione di lead qualificati) su cui sia poi lui stesso a sviluppare l'attivazione delle azioni di vendita nel suo contesto.

7. STRUMENTI DI CRM PER IL TRADE GOVERNATO

Abbiamo già visto l'opportunità di avere soluzioni tecnologiche a basso costo per attrezzare la Piccola e Media Impresa per introdurre il doppio sistema SMaRC locale e SMaRT. Il vantaggio di applicare queste soluzioni a una rete distributiva governata può essere articolato almeno nei seguenti 3 benefici:

1. la condivisione della conoscenza dei vari segmenti di clientela e dei loro comportamenti di acquisto e di utilizzo dei prodotti/servizi erogati costituisce un enorme *know how* che ciascun Trader può utilizzare per migliorare la gestione e le iniziative sul proprio portafoglio clienti;

2. essendo il progetto complesso e tale da incidere sugli aspetti "culturali" sia del fornitore che dei Trader facenti parte della rete distributiva governata, la soluzione di un qualunque problema che si presenta a uno degli operatori coinvolti diventa immediatamente riutilizzabile su tutti gli altri;

3. essendo evidente, infine, sia un vantaggio per il fornitore delle imprese della rete distributiva che un vantaggio per ciascuno dei Trader della rete distributiva, sia l'investimento per l'attivazione che il costo di gestione possono essere suddivisi fra il fornitore e ogni operatore della rete.

Vediamo una soluzione particolare a titolo assolutamente esemplificativo, sviluppata per le reti distributive governate e basata sui principi del marketing cooperativo, come quella proposta da Allaxia con il nome di Connate (Figura 7).

Si tratta di un portafoglio di "capacità" costituite da competenze tecnologiche, applicative, processuali e di servizio integrate, supportate da una *Web Applications Services Platform* (WASP) che integra in una architettura unitaria le componenti necessarie alla gestione di un Sistema SMaRC, coerentemente a quanto indicato dal modello COSMO.

Il principale obiettivo di Connate è quello di permettere la gestione del "ciclo di vita del cliente" attraverso la convergenza di 3 domini che sono:

- il flusso dei processi di business;
- il ciclo di vita del cliente;

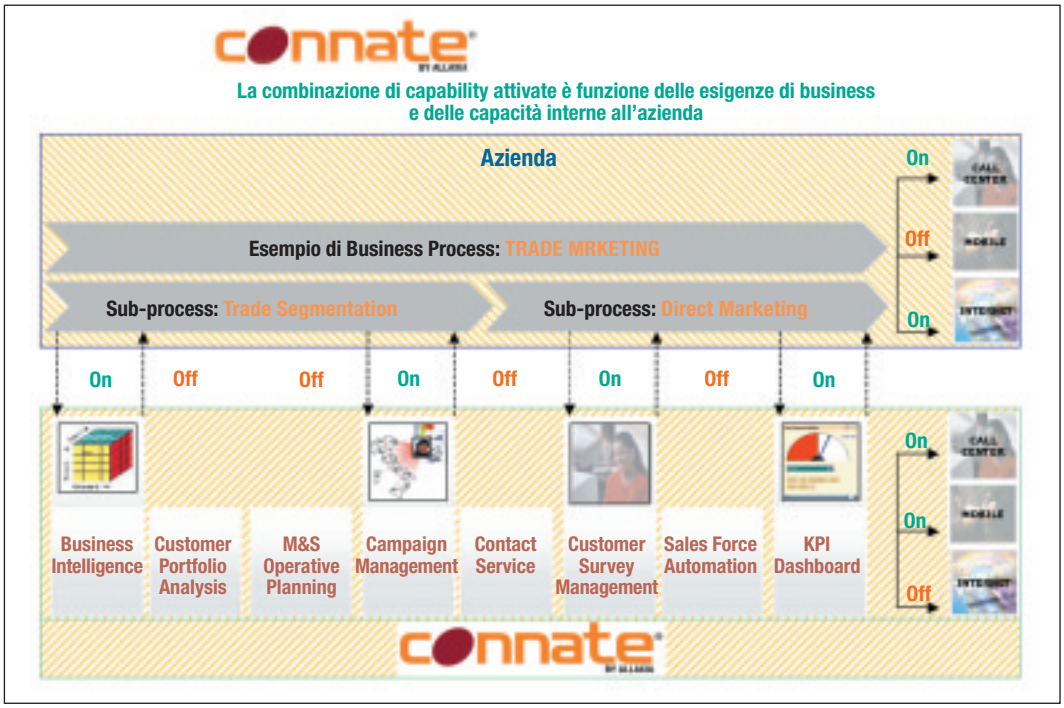


FIGURA 7
Un framework per un ambiente flessibile di servizio a sostegno di uno SMaRC locale
(Fonte: Allaxia)

il sistema tecnologico e di erogazione dei servizi.
La macro funzioni d'uso della *suite* sono:
❑ *Business Intelligence*: comprende il Datamart di marketing&vendita, una funzione Olap/Data Mining, il Reporting e il Narrow Casting; consente di navigare, visualizzare e analizzare i dati per segmenti multidimensionali, con funzionalità avanzate di statistica, grafica e generazione di report; consente, infine, azioni di *narrowcasting* (*data/event based*).
❑ *Customer Portfolio Analysis*: comprende un tool per elaborare il grado di concentrazione dei clienti e analizzare le matrici di migrazione dei clienti; consente di misurare le performance commerciali, analizzando la "scala" (grado di concentrazione) e la "migrazione" del portafoglio clienti nel tempo e assegnando un "colore" corrispondente alla dinamica di ogni cliente.
❑ *Marketing&Sales Operative Planning*: comprende lo *scoring* per cliente, la definizione degli Obiettivi per cliente, la Pianificazione con il calcolo del ROI, la definizione delle Campagne per Cliente; consente di valutare il potenziale, definire gli obiettivi di vendita, e pianificare campagne, metodi e media differenziati per cliente valutandone il ritorno dell'investimento.
❑ *Campaign Management*: comprende la Gestione delle liste, la Gestione delle azioni per cliente, la Supervisione delle campagne; consente la gestione, l'esecuzione e la supervisione delle campagne di Marketing, Vendita e Comunicazione.
❑ *Contact Services*: comprende la Gestione del Numero Verde e dell'accesso a Internet, per fornire informazioni sui prodotti, supportare un operatore di *help-desk* per l'assistenza ai clienti.
❑ *Customer Survey Management*: comprende la Progettazione delle indagini multicanale, la Elaborazione dei Dati e il Reporting automatico; consente di progettare, condurre e analizzare i risultati di indagini realizzabili attraverso un'ampia varietà di metodi e media di rilevazione.
❑ *Sales Force Automation*: comprende l'agenda per la programmazione dei contatti, la scheda contatto, la scheda opportunità, l'identificazione delle tattiche di vendita in funzione della concorrenza, la scheda di qualificazione dell'ordine ecc. a partire dalla costruzione del Profili del Cliente; consente di gestire, quindi, tutte le attività di vendita e le azioni di marketing con i relativi esiti, di formulare le previsioni di vendita e di valutare l'andamento dello sforzo di vendita *ex post*.
❑ *KPI Dasboard*: comprende il Cruscotto degli indicatori personalizzati di performance; consente di rilevare, rappresentare e gestire in modo facile ed efficace gli indicatori di performance e supervisione di processi aziendali creando un "cruscotto personalizzato" ai vari livelli organizzativi e decisionali.
Il Customer DataBase è fisicamente uno solo, gestito a livello centrale da parte del fornitore, ma strutturato in modo tale che ciascun trader abbia a propria disposizione il Customer Data Base dei propri clienti; un'opportuna scheda cliente, in formato elettronico e scaricabile via web, gli permette di aggiornare i dati del cliente per ogni contatto.
Gli strumenti di pianificazione, di market e *customer intelligence*, di gestione delle campagne di Direct Marketing, di impostazione ed esecuzione delle *survey* di mercato o di quelle di customer satisfaction, di supporto allo sforzo di vendita e, infine, gli indicatori di prestazione dei cruscotti per ogni ruolo aziendale identificato, sono tutti accessibili e a sua disposizione per essere supportati dalla suite Connate.
La logica di impostazione è quella indicata di "information utility", e cioè di servizi di "cooperative marketing" per migliorare le performance della rete di vendita e aumentare la profittabilità del portafoglio clienti. Le soluzioni modulari e configurabili sono attivabili in base al tipo di business e al livello di investimenti e agli obiettivi, senza la necessità di modificare strutture e processi aziendali, e con costi sotto controllo e legati all'effettivo utilizzo dei singoli moduli che vengono attivati e resi disponibili (Figura 8).
La soluzione consente di passare da una logica di campagne di marketing gestite centralmente (*top-down*) a una modalità "collaborativa" che coinvolge la rete distributiva con le sue esperienze e conoscenze. I principali benefici sono:

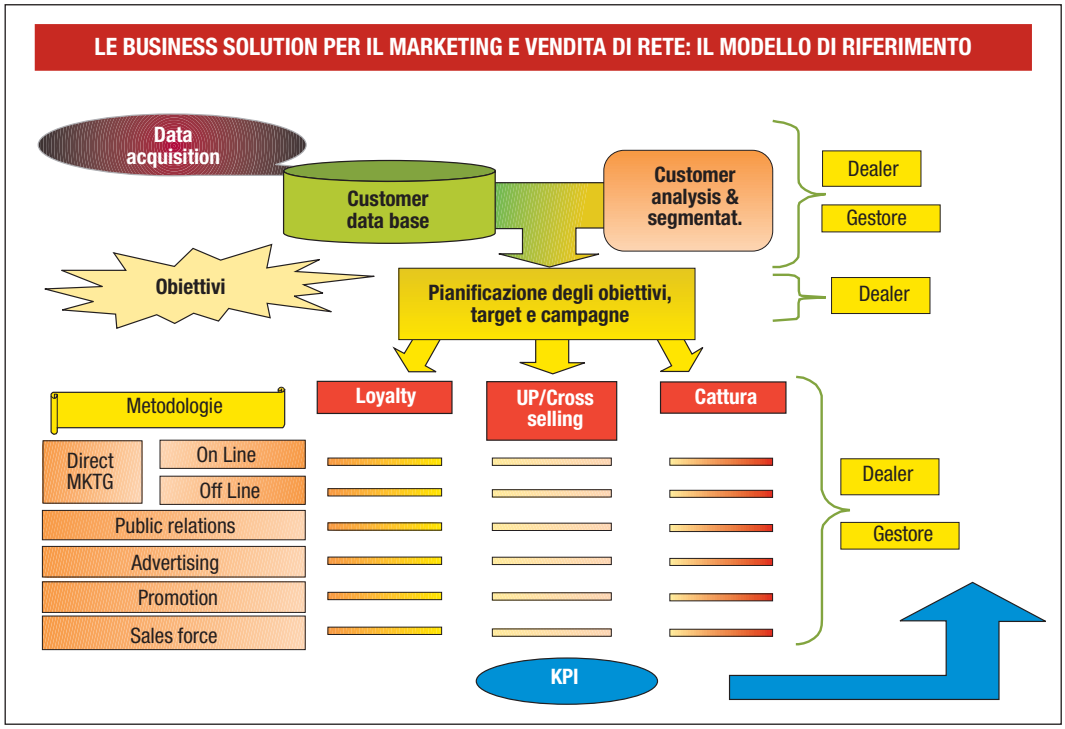


FIGURA 8
Il modello di riferimento per una soluzione di Marketing&Vendita per la Rete

- Misurare la soddisfazione dei clienti;
- Gestire la fedeltà dei clienti (programmi fedeltà);
- Aumentare la pedonalizzazione nel punto vendita;
- Acquisire nuovi clienti;
- Aumentare i volumi di vendita;
- Valorizzare l'azione commerciale migliorandone qualità e profittabilità;
- Monitorare le performance della rete commerciale.

I servizi sono fatturati in modalità *pay-per-use*, in proporzione all'utilizzo e ai risultati effettivamente conseguiti, e sono accessibili secondo due possibili architetture:

- In OUT-SOURCING: servizi totalmente gestiti dal fornitore su specifiche del singolo cliente;
- In CO-SOURCING: servizi integrabili con customer database, contact centre, altri sistemi o processi nell'ambito delle attività di marketing del Cliente.

Ma la cosa molto importate, associata alla soluzione tecnologica, è la disponibilità di una Rete di Business Partner, cioè di consulenti, in grado di "accompagnare" e sostenere il cambiamento presso ciascuna azienda della rete cliente secondo modalità di intervento strutturate con apposite metodologie e strumentate con l'utilizzo dei dati e delle informazioni disponibili attraverso il cruscotto degli indicatori di prestazione per ogni ruolo aziendale, del fornitore e del Trader: si tratta della introduzione del doppio sistema SMaRC e SMaRT.

I costi per un pacchetto di intervento base che comprende sia la soluzione tecnologica dedicata che alcune giornate di consulenza di accompagnamento partono dai 7.000-8.000 € in su per il primo anno; in questo pacchetto l'incidenza della componente tecnologica non supera il 20-25% del totale dell'investimento del Trader, come è giusto che sia per arrivare a qualunque Sistema Operante basato su nuove tecnologie.

Questo tipo di soluzione è stata utilizzata con successo su alcune grandi reti distributive governate; di seguito, viene sinteticamente riportata l'esperienza fatta per una rete di concessionari di un costruttore di auto.

8. IL CASO DI UNA RETE DI CONCESSIONARI AUTO

La rete distributiva comprende alcune centinaia di Concessionari che vendono auto e servizi di assistenza.

L'esigenza da soddisfare con l'introduzione di una soluzione SMaRC locale era quel-

la di migliorare la quota e la penetrazione in un mercato ipercompetitivo, aumentando la capacità di vendita di ciascun concessionario e migliorando l'efficacia e l'efficienza di tutta la rete distributiva, a partire dalla messa in comune, fra concessionaria e costruttore, dei nominativi dei prospect e delle informazioni sul loro rispettivo profilo. La messa a fattore comune dei profili dei prospect e dei clienti determina la possibilità di analizzare ciascun profilo, ottenuto rilevando i dati di una articolata Scheda Cliente in cui vengono registrati i dati delle seguenti 3 sezioni:

- L'anagrafica e i mezzi di contatto.
- Gli esiti previsti della vendita con la sollecitazione al venditore a formulare le proprie valutazioni sulle caratteristiche del prospect in termini di valori associati al modello di auto, al suo uso e alle principali funzioni d'uso del prospect, ai principali fattori di scelta.
- Le argomentazioni di vendita utilizzate dal venditore e i fattori che hanno determinato successo (acquisto da parte del prospect) o insuccesso (prospect che non compra e sparisce).

L'analisi dei successi e degli insuccessi della rete di vendita (di tutta la rete di vendita, indipendentemente dal Trader) mette a disposizione dei venditori un insieme di indicazioni che continuano ad arricchirsi alimentando la costruzione di un vero e proprio Knowledge Data Base finalizzato alla vendita, che nel tempo dovrebbe permette di guidare ogni singola azione commerciale verso un tasso di successo crescente.

I servizi messi poi a disposizione della rete dei concessionari con una soluzione SMaRT erano complessivamente:

- la gestione del processo di vendita;
- l'impostazione del piano di marketing;
- la formazione del personale.

Più specificamente, gli obiettivi della gestione del processo di vendita erano:

- identificare e qualificare liste di prospect;
- attirare persone (prospect, clienti acquisiti e clienti persi) presso il punto vendita;
- convincerle ad acquistare nuovi prodotti e a utilizzare i servizi della concessionaria;
- ampliare il Prospect Data Base, raccogliendo dati per successive azioni di marketing mirate;
- migliorare la conoscenza del proprio mercato potenziale.

La soluzione realizzata si è articolata nei seguenti componenti:

- una "doppia" rete di strumenti *on-line* e di competenze locali (*Business Partner*) per il supporto della rete dei Concessionari
- il supporto alla rete di vendita per aumentare il *sell out* tramite:
  - supporto a campagne esistenti;
  - sviluppo e implementazione di nuove campagne locali one-to-one a breve termine (qualificazione contatto);
  - il miglioramento della produttività di vendita: da contatti qualificati a contratti conclusi;
- l'analisi e il miglioramento dei processi di Marketing e Vendita della Rete di supporto e di ciascun Concessionario, migliorando l'orientamento al cliente.

9. IL VALORE CHE L'INTRODUZIONE DI SISTEMI SMARC E SMART PORTANO NEL BUSINESS DI RETI DISTRIBUTIVE GOVERNATE

Riassumiamo, infine, qual è il contributo, per ciascun livello di interesse coinvolto, che l'innovazione del CRM può dare in termini di valore per il business delle reti distributive governate.

Contributi al valore per il singolo cliente finale

- Poter esprimere le proprie motivazioni ed aspettative.
- Ricevere comunicazioni ed offerte mirate e personalizzate.
- Diversificare e ampliare i metodi e i canali di comunicazione con il dealer e/o con il fornitore.
- Fruire di un continuo miglioramento della qualità dei servizi.

Contributi al valore per il singolo Trader parte della rete distributiva

- Supporto alla individuazione dei punti di forza e delle aree critiche per il recupero dei livelli di efficienza.
- Supporto alla pianificazione economica dell'attività (impatto sulla struttura finanziaria, organizzazione e dimensionamento delle risorse)

In particolare, dal punto di vista marketing:

- Customer Intelligence e azioni di vendita mirate.

- Fidelizzazione dei clienti.
- Aumento del traffico sul punto vendita.
- Recupero dei clienti insoddisfatti e inattivi.
- Raccolta di suggerimenti per il miglioramento del proprio punto vendita.
- Diversificazione ed ampliamento dei metodi e canali di relazione con i clienti.

Contributi al valore per il gestore della rete (per esempio, il trade marketing manager)

- Supporto alla progettazione e sviluppo del Trader Data Base.
- Trasferimento di metodologie e strumenti per il *Trader Profiling* (gestionale e di mercato) e il *Trader Rating*.
- Misurazione dei risultati degli interventi sulla rete.
- Migliorare e qualificare la conoscenza dei Trader.
- Supporto alla individuazione dei punti di forza e delle aree critiche per il recupero dei livelli di efficienza.

In particolare, dal punto di vista marketing:
Aumento dell'efficacia della rete attraverso:

- azioni proattive sui clienti;
- miglioramento della qualità dei processi/strumenti;
- ottimizzazione, grazie alla disponibilità di metodologie e indicatori di performance, ripetibili e comparabili.

- Condivisione di obiettivi commerciali e di marketing, secondo una logica cooperativa.
- Fidelizzazione della rete stessa, tramite la fornitura di servizi personalizzati che "aiutano a vendere".
- Trarre vantaggio dalle nuove modalità di comunicazione, attraverso la gestione di comunità di interessi.

Contributi al valore per l'azienda fornitrice della rete (per esempio il marketing manager)

- Miglioramento del posizionamento competitivo della rete.
- Misura delle performance per dealer e per periodo.
- Ottimizzazione del canale di rivendita.
- Pianificazione economica ed organizzativa dello sviluppo della rete.
- Rilevazione percezione del cliente verso nuovi canali distributivi.
- Analisi e valutazione della opportunità di apertura di nuovi canali.

10. CONCLUSIONI

Rispetto al tema specifico affrontato in questo articolo, si possono fare 3 considerazioni di sintesi sui risultati dell'analisi dello stato dell'arte.

1. Le tecnologie CRM hanno superato la fase di *start up* e sono entrate in una fase di maggiore maturazione che permette di progettarle, realizzarle e metterle in opera con elevata modularità per diversi livelli di complessità e con la gradualità necessaria per soluzioni SMaRC che hanno un significativo impatto sul piano dei processi di vendita, dei ruoli professionali coinvolti e degli obiettivi da raggiungere; tali soluzioni sono disponibili sia per progetti sviluppati completamente "in house", sia per progetti in cui il committente ritiene di poter utilizzare per la propria configurazione SMaRC piattaforme di servizio accessibili via Web come "information utility".

2. I costi di investimento e di gestione delle soluzioni SMaRC sono, di conseguenza, anch'essi modulabili per le diverse situazioni aziendali; in particolare, per quanto, riguarda le piccole e medie imprese, la soluzione come "information utility" basata su un'offerta *pay-per-use* per singolo posto di lavoro è di per sè alla portata di qualunque impresa; per quanto riguarda le piccole imprese di reti distributive governate (nel senso dato in questo articolo, "governate dalle esigenze del cliente finale" che hanno in comune il Trader e il suo fornitore) il sistema SMaRC locale si combina con un sistema SMaRT per supportare i processi di tutti gli attori in Rete e le rispettive piattaforme tecnologiche possono essere co-finanziate dagli attori stessi con contributi proporzionali al peso di ciascuno nella rete;

3. il successo della progettazione di una configurazione CRM è però determinato dall'essere disegnata, sviluppata e messa in opera coerentemente alle funzioni d'uso richieste dai processi di Marketing&Vendite aziendali e alle competenze di ruoli professionali adeguati definiti nell'ambito del Sistema Obiettivo SMaRC-Sistema strutturato e operante di Management della Relazione con il Cliente.

Un'ultima riflessione riguarda, invece, la natura di un progetto di sviluppo di un sistema SMaRC: le piattaforme CRM possono essere considerate come tecnologie per l'automazione dei processi di Marketing&Vendita che introducono modelli operativi e paradigmi di valutazione delle prestazioni di tipo industriale; si può cioè cominciare a parlare di ingegnerizzazione di questi processi, fino a ieri considerati costituiti da attività assolutamente non strutturabili e, inoltre, realizzate da personale con una forte componente di indipendenza.

Bibliorafia

[1] Boldizzoni D., Serio L. (a cura di): Innovazione e crescita nella piccola impresa. *Il Sole 24 Ore*-2003.

[2] www.crm4sme.com

[3] Franklin C.: *Why Innovation Fails*. Spiro Press.

[4] Molineux P.: *Exploiting CRM-connecting with customers*. MCA, 2002.

[5] Peppers D., Rogers M.: I*mpresa one to one - il marketing relazionale nell'era della Rete*. Apogeo, 2001.

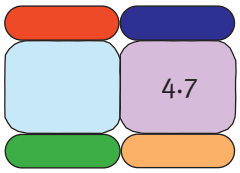
Roberto Bellini, ingegnere elettronico, è docente di Gestione e Marketing dell'Innovazione presso il Politecnico di Milano e, nell'ambito del MIP, ha contribuito ad impostare e realizzare il Corso Avanzato di Management della Relazione con il Cliente.
Presiede la Sezione AICA di Milano e collabora al progetto Osservatorio delle Competenze e delle Certificazioni Informatiche, sviluppato da Fondazione Politecnico.

LINGUISTICA COMPUTAZIONALE

STRUMENTI E RISORSE PER IL TRATTAMENTO AUTOMATICO DELLA LINGUA

Nicoletta Calzolari
Alessandro Lenci

Le ricerche sul TAL hanno aperto nuove prospettive per la creazione di applicazioni per l'accesso intelligente al contenuto documentale. Sviluppi significativi riguardano i sistemi per l'analisi "robusta" del testo, i metodi per l'acquisizione automatica di conoscenza dai documenti e le infrastrutture per lo sviluppo e gestione di risorse linguistiche di grandi dimensioni, grazie ai quali è oggi possibile realizzare modelli e strumenti per il trattamento della lingua utilizzabili in contesti operativi reali.



1. IL TRATTAMENTO AUTOMATICO DELLA LINGUA

Nella società dell'informazione differenti categorie di utenti (professionisti, amministratori pubblici e comuni cittadini) devono confrontarsi con la necessità quotidiana di accedere a grandi quantità di contenuti digitali *semi-strutturati o non strutturati*, all'interno di basi documentali in linguaggio naturale disponibili sul Web o su Intranet locali. Un'alta percentuale delle conoscenze e processi che regolano le attività di gruppi di lavoro, istituzioni e imprese risiede, infatti, all'interno di documenti dalle forme e tipologie più varie (testi normativi, manuali, agenzie stampa, rapporti tecnici, *e-mail* ecc.), talvolta in lingue diverse e, sempre più di frequente, accompagnati da materiale multimediale. La natura non strutturata di tale informazione richiede due passi fondamentali per una sua gestione efficace: ovvero, la selezione dei documenti rilevanti rispetto alle necessità specifiche dell'utente e l'estrazione dell'informazione dai testi, per garantire il suo impiego in altre applicazioni o per compiti specifici. La facilità di tale accesso, la capacità di recuperare l'informazione adeguata in tempi rapidi, la sua gestione e usabilità sono, dunque, parametri chiave per garantire il successo di imprese economiche, lo sviluppo imprenditoriale, la competitività professionale, così come anche l'integrazione sociale e occupazionale e la formazione permanente.

Gli sviluppi più recenti della *linguistica computazionale* e del *natural language engineering* hanno creato soluzioni tecnologiche dalle enormi potenzialità per migliorare la ricerca e gestione intelligente dell'informazione contenuta nei documenti testuali. Le nuove tecnologie della lingua, infatti, permettono ai sistemi informatici di accedere ai contenuto digitale attraverso il *Trattamento Automatico della Lingua* (TAL) o *Natural Language Processing* (NLP). Il problema di come acquisire e gestire la conoscenza depositata nei documenti testuali dipende dal suo essere codificata all'interno della rete di strutture e relazioni grammaticali e lessicali che costituiscono la natura stessa della comunicazione linguisti-

ca. Sono il lessico e le regole per la combinazione delle parole in strutture sintatticamente complesse che nel linguaggio si fanno veicoli degli aspetti multiformi e creativi dei contenuti semantici. Attraverso l'analisi linguistica automatica del testo, gli strumenti del TAL sciolgono la tela del linguaggio per estrarre e rendere espliciti quei nuclei di conoscenza che possono soddisfare i bisogni informativi degli utenti. Dotando il computer di capacità avanzate di elaborare il linguaggio e decodificarne i messaggi, diventa così possibile costruire automaticamente rappresentazioni del contenuto dei documenti che permettono di potenziare la ricerca di documenti anche in lingue diverse (*Crosslingual Information Retrieval*), l'estrazione di informazione rilevante da testi (*Information Extraction*), l'acquisizione dinamica di nuovi elementi di conoscenza su un certo dominio (*Text Mining*), la gestione e organizzazione del materiale documentale, migliorando così i processi di elaborazione e condivisione delle conoscenze.

2. UN PO' DI STORIA: IL TAL IERI E OGGI

Nata come disciplina di frontiera, di fatto ai margini sia del mondo umanistico che delle applicazioni informatiche più tradizionali, la linguistica computazionale in poco più di 50 anni è riuscita a conquistare una posizione di indiscussa centralità nel panorama scientifico internazionale. In Italia, alla storica culla pisana rappresentata dall'Istituto di Linguistica Computazionale del CNR – fondato e diretto per lunghi anni da Antonio Zampolli – si sono affiancati molti centri e gruppi di ricerca attivi su tutto il territorio nazionale. Sul versante applicativo, le numerose iniziative imprenditoriali nel settore delle tecnologie della lingua testimoniano l'impatto crescente della disciplina (sebbene con ritmi molto più lenti che nel resto dell'Europa, come risulta dal rapporto finale del progetto comunitario Euromap [12]) al di fuori dello specifico ambito accademico, prova del fatto che i tempi sono diventati maturi perché molti dei suoi risultati affrontino la prova del mercato e della competizione commerciale.
Quali i motivi di questa crescita esponenziale? Sebbene facilitato dai progressi nel settore informatico e telematico, unitamente all'effetto catalizzante di Internet, sarebbe improprio spiegare lo sviluppo della disciplina solo in termini di fattori meramente tecnologici. In realtà, la linguistica computazionale possiede, oggi, una sua maturità metodologica nata dalla conquista di un preciso spazio di autonomia disciplinare anche rispetto alle sue anime originarie, l'indagine umanistica e la ricerca informatica. Questa autonomia si contraddistingue per un nuovo e delicato equilibrio tra lingua e computer. Le elaborazioni computazionali sono, infatti, chiamate a rispettare la complessità, articolazione, e multidimensionalità della lingua e delle sue manifestazioni testuali. Al tempo stesso, i documenti testuali emergono come una risorsa di conoscenza che può essere gestita ed elaborata con le stesse tecniche, metodologie e strumenti che rappresentano lo stato dell'arte nella tecnologia dell'informazione.
A tale proposito è utile ricordare come la linguistica computazionale affondi le sue radici in due distinti paradigmi di ricerca. Da un lato, è possibile trovare i temi caratteristici dell'applicazione di metodi statistico-matematici e informatici allo studio del testo nelle scienze umane, di cui Padre Roberto Busa e Antonio Zampolli rappresentano i pionieri nazionali. Il secondo paradigma fondante è rappresentato dall'*Intelligenza Artificiale* (IA) e, in particolare, dall'ideale delle "macchine parlanti", che hanno promosso temi di ricerca rimasti "classici" per il settore, come la traduzione automatica, i sistemi di dialogo uomo-macchina ecc..
Il TAL si è sviluppato alla confluenza di queste due tradizioni promuovendo il faticoso superamento di alcune forti dicotomie che hanno caratterizzato le anime della linguistica computazionale ai suoi esordi, dicotomie riassumibili proprio in diverse, e a tratti ortogonali, concezioni della lingua e dei metodi per le sue elaborazioni computazionali. Da un lato, la lingua, come prodotto complesso e dinamico realizzato nella variabilità delle sue tipologie testuali, si è a lungo opposta alla lingua in *vitro* di esperimenti da laboratorio troppo spesso decontestualizzati e riduttivi rispetto alle sue reali forme e usi. A questo bisogna unire anche la prevalenza dei metodi statistici per lo studio delle regolarità distribuzionali delle pa-

role tipico di molta linguistica matematica applicata al testo, in forte contrasto col prevalere di tecniche simboliche che hanno costituito, per lungo tempo, il modello dominante per la progettazione dei primi algoritmi per il TAL. Il superamento di tale dicotomia è stato reso possibile grazie al radicale mutamento di paradigma avvenuto nel TAL, a partire dalla seconda metà degli anni '80, caratterizzato dal diffondersi, e poi dal netto prevalere, di un'epistemologia neo-empirista. Questo cambiamento si è concretizzato nella diffusione dei metodi statistico-quantitativi per l'analisi computazionale del linguaggio [19], e nella rinnovata centralità dei dati linguistici.

La disponibilità crescente di *risorse linguistiche*, in particolari *corpora* testuali e lessici computazionali, ha costituito un fattore determinante in questa svolta metodologica e tecnologica nel TAL. La disponibilità di *corpora* di grandi dimensioni è diventata una variabile fondamentale in ogni fase di sviluppo e valutazione degli strumenti per l'elaborazione dell'informazione linguistica. Gli strumenti per il TAL sono, infatti, ora chiamati a confrontarsi non con pseudolinguaggi di laboratorio, ma con testi di grande complessità e variabilità linguistica e strutturale. A sua volta, questo ha portato al diffondersi di tecniche di elaborazione linguistica più "robuste" di quelle simboliche tradizionali, in grado di affrontare la variabilità lessicale e strutturale del linguaggio, e anche quel suo continuo resistere ai vincoli grammaticali che è così evidente in molte sue manifestazioni, prima fra tutte, la lingua parlata. La possibilità di accedere a quantità sempre crescenti di dati linguistici digitali ha indubbiamente facilitato tale innovazione metodologica, fornendo i dati linguistici necessari per un uso intensivo dei metodi statistici, che hanno incominciato a ibridare le architetture e gli algoritmi più tradizionali. Un ulteriore fattore di accelerazione è stato fornito dalla necessità della tecnologia della lingua di passare da prototipi di laboratorio a sistemi funzionanti in grado di offrirsi agli utenti come affidabili strumenti per la gestione dell'informazione linguistica. Il banco di prova del *World Wide Web*, per sua natura risorsa di informazione documentale multiforme e magmatica, ha imposto ai sistemi per il TAL di acquisire una capacità di adeguarsi alle complessità della lingua reale, prima impensabile.

3. DAL TESTO ALLA CONOSCENZA

All'interno dell'ampio spettro di attività del TAL, che coinvolgono quasi tutti i domini dell'*Information Technology*[1], di particolare interesse e impatto sono le possibilità offerte dalle più recenti tecnologie della lingua per *trasformare i documenti testuali in risorse di informazione e conoscenza*. Alla base di questo processo di accesso e analisi del contenuto digitale risiedono tre tipi di tecnologie, fondamentali per ogni sistema basato sul TAL:

1. *strumenti per l'analisi linguistica di testi e l'acquisizione dinamica di conoscenza* – analizzatori morfologici, parser sintattici[2], acquisitori automatici di terminologia e informazione semantica dai testi ecc.;

2. *risorse linguistiche* – lessici computazionali, reti semantico-concettuali multilingui, *corpora* testuali anche annotati sintatticamente e semanticamente per lo sviluppo e la valutazione di tecnologia del linguaggio;

3. *modelli e standard per la rappresentazione dell'informazione linguistica* – ontologie per il *knowledge sharing* e la codifica lessicale, modelli per la rappresentazione e interscambio di dati linguistici.

Grazie anche alle nuove opportunità offerte dalla tecnologia XML (*eXtensible Markup Language*) è possibile realizzare una maggiore integrazione tra i diversi moduli per l'elaborazione della lingua, e la standardizzazione della rappresentazione dei dati, necessaria per assicurare la loro interscambiabilità

[1] Questi vanno dal riconoscimento automatico del parlato alla traduzione automatica, dallo sviluppo di interfacce uomo-macchina multimodali ai sistemi di *question-answering* che permettono di interrogare una base documentale formulando la richiesta come una domanda in linguaggio naturale. Un'ampia rassegna delle varie applicazioni del TAL è disponibile in [11, 13].

[2] *Il parsing* è il processo di analisi linguistica attraverso cui viene ricostruita la struttura sintattica di una frase, rappresentata dall'articolazione dei costituenti sintagmatici e dalle relazioni di dipendenza grammaticale (esempio soggetto, complemento oggetto ecc.).

e la coerenza del trattamento dell'informazione. Strumenti di analisi, risorse linguistiche e standard di rappresentazione vengono, dunque, a costituire un'infrastruttura per il TAL che attraverso l'analisi linguistica automatica dei documenti testuali permette di estrarre la conoscenza implicitamente contenuta in essi, trasformandola in conoscenza esplicita, strutturata e accessibile sia da parte dell'utente umano che da parte di altri agenti computazionali (Figura 1).
È importante sottolineare l'aspetto di stretta interdipendenza tra i vari componenti per il TAL, illustrata in maggior dettaglio in figura 2. Gli strumenti di analisi linguistica costruiscono una rappresentazione avanzata del contenuto informativo dei documenti attraverso elaborazioni del testo a vari livelli di complessità: analisi morfologica e lemmatizzazione, analisi sintattica, interpretazione e disambiguazione semantica ecc.. I moduli di elaborazione sono solitamente interfacciati con *database* linguistici, che rappresentano e codificano grandi quantità di informazione terminologica e lessicale, morfologica, sintattica e semantica, che ne permettono sofisticate modalità di analisi. Le analisi linguistiche forniscono l'*input* per i moduli di estrazione, acquisizione e strutturazione di conoscenza. La conoscenza estratta costituisce una risorsa per l'utente finale, e permette allo stesso di popolare ed estendere i repertori linguistico-lessicali e terminologici che sono usati in fase di analisi dei documenti. Si realizza, così, un ciclo virtuoso tra strumenti per il TAL e risorse linguistiche. Le risorse linguistiche lessicali e testuali permettono di costruire, ampliare, rendere operativi, valutare modelli, algoritmi, componenti e sistemi per il TAL, sistemi che sono, a loro volta, strumenti necessari per alimentare dinamicamente ed estendere tali risorse.
Un esempio di architettura per il trattamento automatico dell'Italiano è *Italian NLP*, sviluppato dall'Istituto di Linguistica Computazionale – CNR in collaborazione con il Dipartimento di Linguistica – Sezione di Linguistica Computazionale dell'Università di Pisa. *Italian NLP* è un ambiente integrato di strumenti e risorse che consentono di effettuare analisi linguistiche incrementali dei

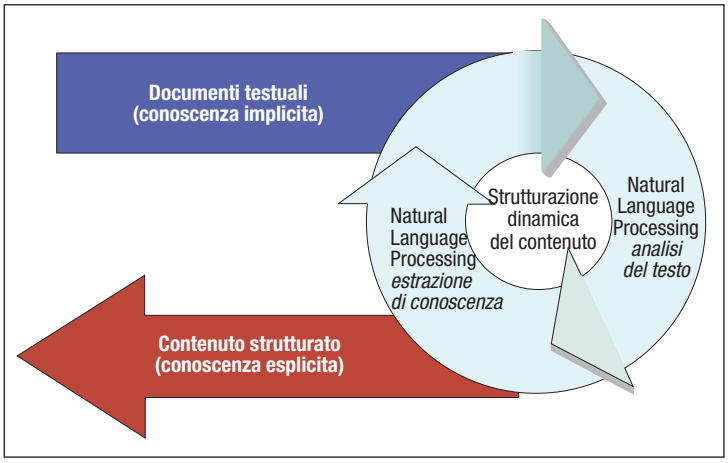


FIGURA 1
Dalla conoscenza implicita alla conoscenza esplicita

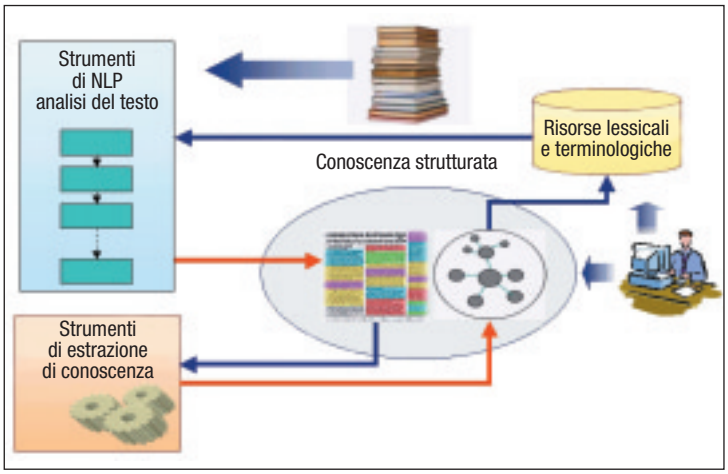


FIGURA 2
Un'architettura per l'estrazione di conoscenza dai testi basata sul TAL

testi. Ciascun modulo di *Italian NLP* procede all'identificazione di vari tipi di unità linguistiche di complessità strutturale crescente, ma anche utilizzabili singolarmente come fonte di informazione sull'organizzazione linguistica dei testi.
Come si vede in figura 3, un aspetto significativo di *Italian NLP* è il carattere ibrido della sua architettura. Moduli simbolici di *parsing* (basati su metodologie consolidate nella linguistica computazionale, come le tecnologie a stati finiti) sono affiancati a strumenti statistici che sono usati per operare disambiguazioni sintattiche e semantiche, filtrare "rumore" dalle analisi e anche arricchire le risorse lessicali con informazioni direttamente estratte dai testi oggetto di analisi, permettendo l'aggiornamento e specializzazione continua delle risorse linguistiche, e garantendo una maggiore robustezza e portabilità

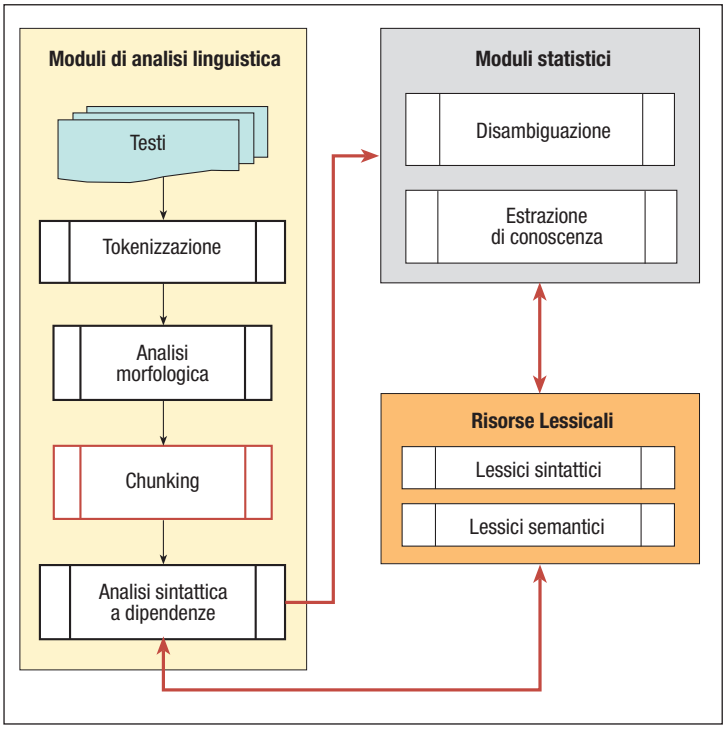


FIGURA 3
Strumenti di analisi e risorse linguistiche in Italian NLP

degli strumenti di analisi del linguaggio su domini e registri linguistici diversi.

Uno dei livelli di analisi linguistica più impegnativi è l'analisi sintattica automatica. In *Italian NLP*, questa è realizzata in due fasi successive. Dopo un processo di tokenizzazione[3] e analisi morfologica, viene effettuato un *parsing* "leggero" del testo (*shallow parsing*), in cui un *chunker* realizza contemporaneamente la disambiguazione morfosintattica delle parole, cioè l'identificazione della categoria sintattica con cui una forma occorre in un dato contesto linguistico, e la segmentazione del testo in sequenze di gruppi sintattici non ricorsivi (*chunk*) di cui vengono individuati il tipo (nominale, verbale ecc.) e la testa lessicale [15]. Per esempio, la frase "Il Presidente della Repubblica ha visitato la capitale della Francia" viene segmentata dal chunker nel modo seguente[4]:

$[_{N\_C}$ *Il Presidente*$][_{P\_C}$ *della Repubblica*$][_{FV\_C}$ *ha visitato*$][_{N\_C}$ *la capitale*$][_{P\_C}$ *della Francia*$]$

Come risultato del *chunking*, si ottiene dunque una strutturazione del testo in unità linguisticamente rilevanti sia per processi di estrazione dell'informazione e text mining, sia come input per la seconda fase di parsing in cui il testo segmentato è analizzato a livello sintattico-funzionale, per identificare relazioni grammaticali tra gli elementi nella frase come soggetto, oggetto, complemento, modificatore ecc.. In *Italian NLP* questo tipo di analisi è realizzato da IDEAL, *Italian DEpendency AnaLyzer* [1, 2], un compilatore di grammatiche a stati finiti definite su sequenze di chunk. Le regole della grammatica fanno uso di test sulle informazioni associate ai chunk (per esempio, informazioni morfosintattiche, tratti di accordo) e su informazioni lessicali esterne (il lessico che viene usato a questo fine comprende circa venticinquemila *frame sintattici di sottocategorizzazione*)[5]. L'output di IDEAL è costituito da relazioni grammaticali binarie tra una testa lessicale e un suo dipendente che forniscono una rappresentazione della struttura sintattica come la seguente[6]:

sogg	(visitare, presidente)
comp	(presidente, repubblica.<intro=di>)
ogg	(visitare, capitale)
comp	(capitale, Francia.<intro=di>)

Simili rappresentazioni della struttura linguistica del testo forniscono l'input fondamentale per processi di estrazione della conoscenza. Un esempio di applicazione di questo tipo è l'acquisizione semi-automatica di on-

[3] La *tokenizzazione* consiste nella segmentazione del testo in unità minime di analisi (parole). In questa fase l'input è sottoposto a un processo di normalizzazione ortografica (esempio separazione di virgolette e parentesi della parole, riconoscimento dei punti di fine frase ecc.), nell'ambito del quale vengono anche identificate le sigle, gli acronimi e le date.

[4] N_C, P_C e FV_C stanno rispettivamente per chunk di tipo nominale, preposizionale e verbale

[5] Un f*rame di sottocategorizzazione* specifica il numero e tipo di complementi che sono selezionati da un termine lessicale. Per esempio, il verbo *mangiare* seleziona per un complemento oggetto opzionale (cfr. *Gianni ha mangiato un panino; Gianni ha mangiato*), mentre il verbo *dormire*, in quanto intransitivo, non può occorrere con un complemento oggetto.

[6] sogg = soggetto; comp = complemento; ogg = oggetto diretto

tologie (*ontology learning*) da testi come supporto avanzato alla gestione documentale [8, 18]. Un'*ontologia* [9, 22] è un sistema strutturato di concetti e relazioni tra concetti che viene a costituire una "mappa" della conoscenza di un certo dominio od organizzazione. Gli strumenti e le risorse del TAL permettono di trasformare le conoscenze implicitamente codificate all'interno dei documenti testuali in conoscenza esplicitamente strutturata come un'ontologia di concetti. Attraverso il TAL è possibile, dunque, dotare i sistemi informatici di una chiave di accesso semantica alle basi documentali, consentendo agli utenti di organizzare e ricercare i documenti su base concettuale, e non solo attraverso l'uso di parole chiave. Le ontologie estratte dinamicamente dai testi vengono a costituire un ponte tra il bisogno di informazione degli utenti - rappresentato da idee, concetti o temi di interesse - e i documenti in cui l'informazione ricercata rimane nascosta all'interno dell'organizzazione linguistica del testo, che spesso ne ostacola il recupero. Anche in un linguaggio tecnico e apparentemente controllato, infatti, lo stesso concetto può essere espresso con una grande variazione di termini, e la scelta di uno di questi da parte dell'utente in fase di ricerca o indicizzazione, può impedire il recupero di documenti ugualmente rilevanti, ma in cui lo stesso concetto appare sotto forme linguistiche diverse. Le tecnologie della lingua rendono possibile lo sviluppo di *un ambiente per la creazione dinamica di ontologie a partire dall'analisi linguistica dei documenti*. Diventa così possibile velocizzare il processo di gestione dell'indicizzazione e della classificazione della base documentale, e ridurre il grado di arbitrarietà dei criteri di classificazione. La questione è, infatti, come fare a determinare i concetti rilevanti e più caratterizzanti per i documenti di un certo dominio di interesse. Per affrontare questo problema le tecniche linguistico-computazionali si basano su un'ipotesi molto semplice: i documenti sono estremamente ricchi di termini che con buona approssimazione veicolano i concetti e i temi rilevanti nel testo. Termini sono nomi propri, nomi semplici come *museo* o *pinacoteca,* oppure gruppi nominali strutturalmente complessi come *museo archeologico*, *ministero dei beni culturali, soprintendenza archeologica* ecc.. I termini possono essere a loro volta raggruppati, in quanto esprimono concetti molto simili. Per esempio, *scultura, affresco* e *quadro* condividono tutti un concetto più generico di "opera artistica" a cui possono essere ricondotti a un certo grado di astrazione. Attraverso l'uso combinato di tecniche statistiche e di strumenti avanzati per l'analisi linguistica come quelli di *Italian NLP* è possibile analizzare il contenuto linguistico dei documenti appartenenti a un dato dominio di conoscenza, individuare i termini potenzialmente più significativi e ricostruire una "mappa" dei concetti espressi da questi termini, ovvero costruire un'ontologia per il dominio di interesse. Come si vede nella figura 4, alla base dell'ontologia risiede un glossario di termini (semplici e complessi) estratti dai testi dopo una fase di analisi linguistica, effettuata con moduli di parsing. I termini estratti vengono successivamente filtrati con criteri statistici per selezionarne i più utili per caratterizzare una certa collezione di documenti. I termini sono organizzati e strutturati come in un Thesaurus di tipo classico, sulla base di alcune relazioni semantiche di base. L'ontologia viene, dunque, a essere composta di unità concettuali definite come insiemi di termini semanticamente affini. I concetti possono, inoltre, essere organizzati secondo la loro maggiore o minore specificità articolando l'ontologia come una tassonomia. Dal momento che un sistema di conoscenza non è fatto solo di concetti che si riferiscono a entità del dominio, ma anche di processi, azioni ed eventi che vedono coinvolte queste entità secondo ruoli e funzioni diverse, uno stadio più avanzato di estrazione può puntare anche all'identificazione di relazioni non tassonomiche tra concetti (per esempio, la funzione tipica di una certa entità, la sua locazione ecc.). È importante sottolineare che il processo di *ontology learning* attraverso l'analisi linguistica dei documenti avviene generalmente in stretta cooperazione con gli utenti, che sono chiamati a intervenire nelle varie fasi di estrazione della conoscenza per validarne i risultati. Come in altri settori di applicazione del TAL, anche in questo caso le tecnologie della lingua utilmente contribuiscono alla gestione dei contenuti di

informazione a *supporto* dell'esperto umano, senza pretendere di sostituirsi ad esso.
Gli strumenti di *Italian NLP* sono usati in molteplici contesti applicativi, in cui hanno dimostrato l'ampiezza e rilevanza delle opportunità pratiche offerte dal TAL. Tra gli esempi più significativi a livello nazionale è possibile citare i moduli linguistico-computazionali SALEM (*Semantic Annotation for LEgal Management*) [3] - sviluppato nell'ambito del progetto *Norme in Rete* (NIR) del Centro Nazionale per l'Informatica nella Pubblica Amministrazione (CNIPA) - e T2K (Text-2-Knowledge) - realizzato nell'ambito del progetto *TRAGUARDI* del Dipartimento della Funzione Pubblica - FORMEZ[7]. SALEM è un modulo per l'annotazione automatica della struttura logica dei documenti legislativi, integrato nell'editore normativo *NIREditor*, sviluppato dall'Istituto di Teoria e Tecnica dell'Informazione Giuridica - CNR. Attraverso l'analisi computazionale del testo, SALEM rende espliciti gli aspetti più rilevanti del contenuto normativo, individuando elementi quali il destinatario della norma, la sanzione prevista ecc.. Questi elementi di contenuto sono annotati esplicitamente sul testo con metadati XML, garantendo una migliore gestione e ricerca della documentazione legislativa. Il modulo T2K è, invece, finalizzato alla costruzione semi-automatica di thesauri di termini e di ontologie di metadati semantici per la gestione documentale nella pubblica amministrazione. A livello internazionale, gli strumenti per il TAL, illustrati sopra, sono stati applicati in numerosi progetti finanziati dall'Unione Europea, tra i quali si vogliono qui citare POESIA (*Public Open-source Environment for a Safer Internet Access*) [10], dedicato alla creazione di sistemi avanzati di *filtering* di siti web, e VIKEF (*Virtual Information and Knowledge Environment Framework*)[8], in cui gli strumenti di *Italian NLP* sono utilizzati per l'annotazione semantica di testi e la costruzione di ontologie, nell'ambito delle iniziative relative al Semantic Web. Questi sono solo alcuni dei numerosi esempi di progetti e iniziative in cui i prodotti del TAL la-

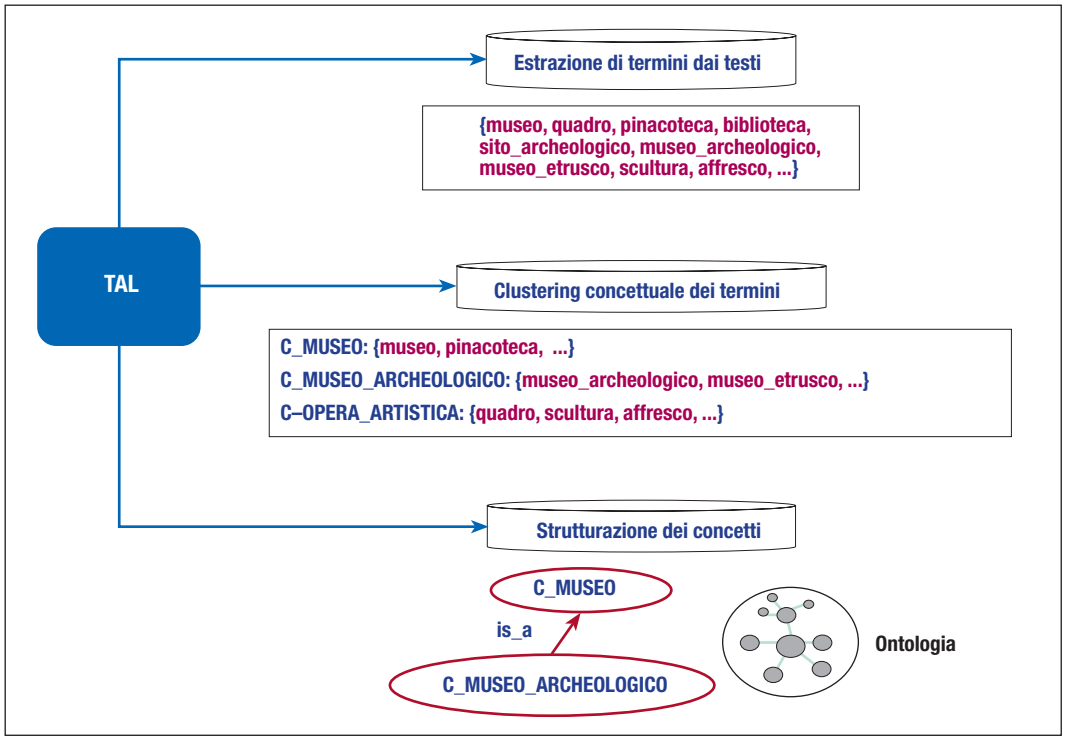


FIGURA 4
TAL e ontology learning

[7] Il progetto TRAGUARDI, di cui è responsabile la dott.ssa Anna Gammaldi di FORMEZ, è un'azione di sostegno alle pubbliche amministrazioni per la gestione dei fondi strutturali.

[8] http://www.vikef.net

sciano i centri di ricerca per entrare a diretto contatto con l'utenza e il mercato. Inoltre è importante notare come i contesti applicativi riguardino tipologie di testi completamente diverse, che vanno dai documenti legislativi alla documentazione della pubblica amministrazione, fino al linguaggio dei siti web. Questo testimonia la versatilità della ricerca attuale sul TAL nella sua capacità di affrontare il linguaggio naturale nella complessità delle sue più diverse e varie manifestazioni.

4. RISORSE LESSICALI PER IL TAL

Gli strumenti e le applicazioni del TAL hanno bisogno di poter interpretare il significato delle parole, porta di accesso al contenuto di conoscenza codificato nei documenti. I *lessici computazionali* hanno lo scopo di fornire una rappresentazione esplicita del significato delle parole in modo tale da poter essere direttamente utilizzato da parte di agenti computazionali, come, per esempio, parser, moduli per Information Extraction ecc.. I lessici computazionali multilingui aggiungono alla rappresentazione del significato di una parola le informazioni necessarie per stabilire delle connessioni tra parole di lingue diverse.
Nell'ultimo decennio numerose attività hanno contribuito alla creazione di lessici computazionali di grandi dimensioni. All'esempio più noto, la rete semantico-concettuale WordNet [7] sviluppata all'università di Princeton, si sono affiancati anche altri repertori di informazione lessicale, come PAROLE [21], SIMPLE [14] e EuroWordNet [23] in Europa, Comlex e FrameNet negli Stati Uniti, ecc.. Per quanto riguarda l'italiano, è importante citare i lessici computazionali ItalWordNet e CLIPS, entrambi sviluppati nell'ambito di due progetti nazionali finanziati dal MIUR e coordinati da Antonio Zampolli.
ItalWordNet è una rete semantico-lessicale per l'italiano, strutturata secondo il modello di WordNet e consiste in circa 50.000 entrate. Queste sono costituite da uno o più sensi raggruppati in *synset* (gruppi di sensi sinonimi tra loro). I synset sono collegati tra loro principalmente da relazioni di *iperonimia*[9], che permettono di strutturare il lessico in gerarchie tassonomiche. I nodi più alti delle tassonomie sono a loro volta collegati agli elementi di una ontologia (*Top Ontology*), indipendente da lingue specifiche, che ha la funzione di organizzare il lessico in classi semantiche molto generali. Infine, ogni *synset* della rete è collegato, tramite una relazione di equivalenza, a synset del WordNet americano. Questo collegamento costituisce l'indice interlingue (*Interlingual Index* – ILI) e attraverso di esso ItalWordNet viene a essere integrata nella famiglia di reti semantiche sviluppata nel progetto europeo EuroWordNet, diventando così una vera e propria risorse lessicale multilingue (Figura 5). L'ILI è anche collegato alla *Domain Ontology*, che contiene un'ontologia di domini semantici. Oltre all'iperonimia, il modello ItalWordNet comprende anche una grande varietà di altre *relazioni semantiche* che, collegando sensi di lemmi anche appartenenti a categorie morfosintattiche differenti, permettono di evidenziare diverse relazioni di significato, operanti sia a livello paradigmatico sia a livello sintagmatico.
Il progetto SIMPLE (*Semantic Information for Multipurpose Plurilingual LExica*) ha portato alla definizione di un'architettura per lo sviluppo di lessici computazionali semantici e alla costruzione di lessici computazionali per 12 lingue europee (Catalano, Danese, Finlandese, Francese, Greco, Inglese, Italiano, Olandese, Portoghese, Spagnolo, Svedese, Tedesco). I lessici di SIMPLE rappresentano un contributo estremamente innovativo nel settore delle risorse lessicali per il TAL, offrendo una rappresentazione articolata e multidimensionale del contenuto semantico dei termini lessicali. Il modello di rappresentazione semantica di SIMPLE è stato usato anche per la costruzione di CLIPS, che include 55.000 entrate lessicali con informazione fonologica, morfologica, sintattica e semantica.
Il modello SIMPLE costituisce un'architettura

[9] Un termine lessicale *x* è un *iperonimo* di un termine lessicale *y* se, e solo se, *y* denota un sottoinsieme delle entità denotate da *x*. Per esempio, *animale* è un iperonimo di *cane*. La relazione simmetrica è quella di *iponimia*, per cui *cane* è un iponimo di *animale*.

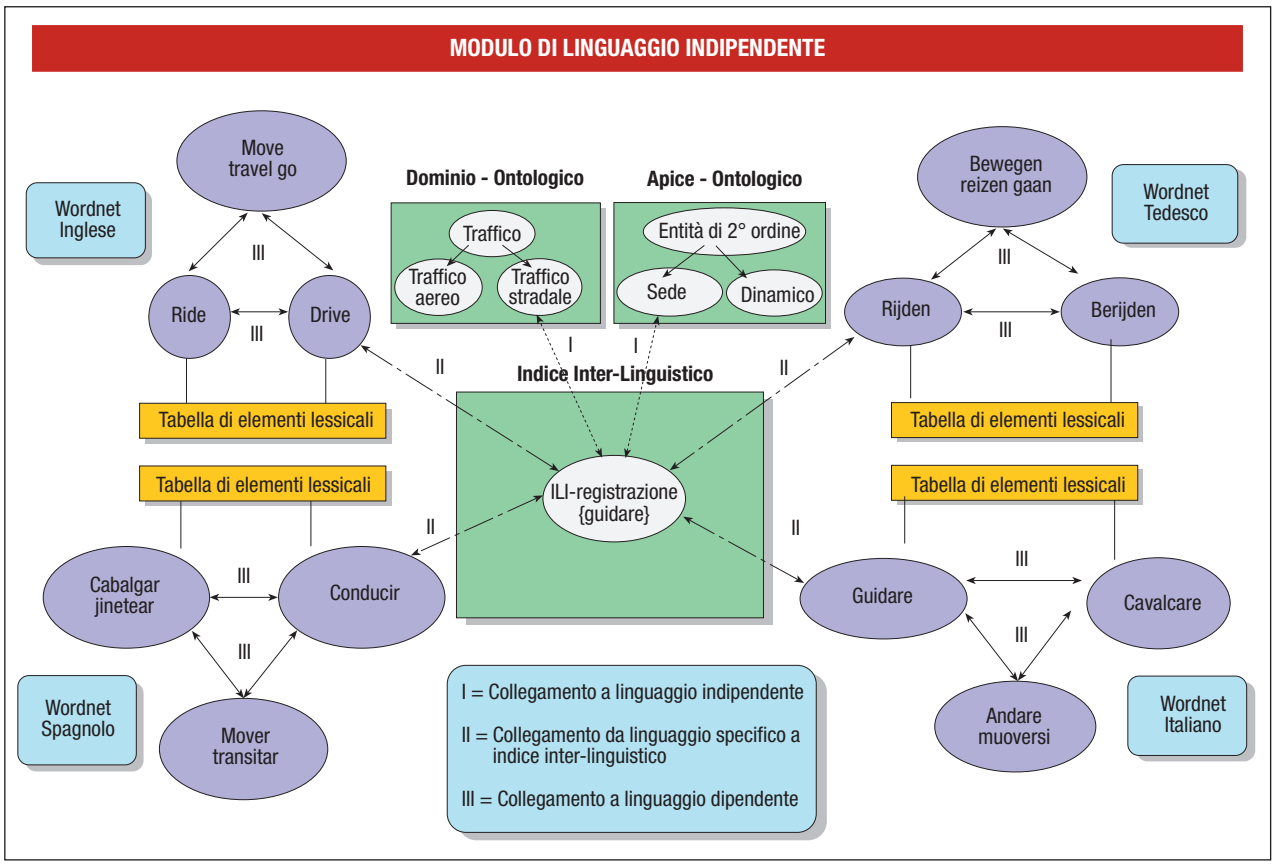


FIGURA 5

L'architettura di EuroWordNet

TABELLA 1

Un frammento della Core Ontology di SIMPLE

1. TELIC
2. AGENTIVE
2.1. *Cause*
3. CONSTITUTIVE
3.1. *Part*
3.1.1. Body_part
3.2. *Group*
3.2.1. Human_group
3.3. *Amount*
4. ENTITY
4.1. *Concrete_entity*
4.1.1. Location
...

per lo sviluppo di lessici computazionali nel quale il contenuto semantico è rappresentato da una combinazione di diversi tipi di entità formali [14] con i quali si cerca di catturare la multidimensionalità del significato di una parola. In tal modo, SIMPLE tenta di fornire una risposta a importanti questioni che coinvolgono la costruzione di ontologie di tipi lessicali, facendo emergere allo stesso tempo problemi cruciali relativi alla rappresentazione della conoscenza lessicale. Al cuore del modello SIMPLE è possibile trovare un repertorio di *tipi semantici* di base e un insieme di informazioni semantiche che devono essere codificate per ciascun senso. Tali informazioni sono organizzate in *template*, ovvero strutture schematiche che rappresentano formalmente l'articolazione interna di ogni tipo semantico, specificando così vincoli semantico-strutturali per gli oggetti lessicali appartenenti a quel tipo. I tipi semantici formano la *Core Ontology* di SIMPLE (Tabella 1), uno dei cui modelli ispiratori è la *Struttura Qualia* definita nella teoria del Lessico Generativo [5, 20]. I tipi semantici sono, infatti, organizzati secondo principi ortogonali, quali la funzione tipica delle entità, la loro origine o costituzione mereologica ecc., nel tentativo di superare i limiti quelle ontologie che troppo spesso ap-

Lemma:	Violino
SEMU_ID:	#V1
POS:	N
GLOSS:	Tipo di strumento musicale
DOMAIN:	**MUSIC**
SEMANTIC_TYPE:	**Instrument**
FORMAL_ROLE:	**Isa** *strumento_musicale*
CONSTITUTIVE_ROLE:	**Has_as_part** *corda*
	Made_of *legno*
TELIC_ROLE:	**Used_by** *violinista*
	Used_for *suonare*
Lmma:	**Guardare**
SEMU_ID:	#G1
POS:	V
GLOSS:	Rivolgere lo sguardo verso qualcosa per osservarlo
SEMANTIC_TYPE:	**Perception**
EVENT_TYPE	**Process**
FORMAL_ROLE:	**Isa** *percepire*
CONSTITUTIVE_ROLE:	**Instrument** *occhio*
	Intentionality = yes
PRED_REPRESENTATION:	Guardare (**Arg0: aniùate**) (**Arg1: entity**)
SYN_SEM_LINKING:	**Arg0 = subj_NP**
	Arg1 = obj NP

TABELLA 2
Entrate lessicali di SIMPLE per violino e guardare

piattiscono la ricchezza concettuale sulla sola dimensione tassonomica.
Il modello di SIMPLE fornisce le specifiche per la rappresentazione e la codifica di un'ampia tipologia di informazioni lessicali, tra le quali il tipo semantico, l'informazione sul dominio, la struttura argomentale per i termini predicativi, le preferenze di selezione sugli argomenti, informazione sul comportamento azionale e aspettuale dei termini verbali, il collegamento delle strutture predicative semantiche ai *frame* di sottocategorizzazione codificati nel lessico sintattico di PAROLE, informazioni sulle relazioni di derivazione tra parole appartenenti a parti del discorso diverse (per esempio, *intelligente – intelligenza; scrittore – scrivere* ecc.). In SIMPLE, i sensi delle parole sono codificati come *Unità Semantiche o SemU*. Ad ogni *SemU* viene assegnato un tipo semantico dall'ontologia, più altri tipi di informazioni specificate nel *template* associato a ciascun tipo semantico. La tabella 2 fornisce una rappresentazione schematica di due entrate lessicali (per il nome *violino* e il verbo *guardare*) codificate secondo le specifiche del modello SIMPLE. Il potere espressivo di SIMPLE è costituito da un ampio insieme di relazioni organizzate lungo le quattro dimensioni della Struttura Qualia proposta nel Lessico Generativo come assi principali della descrizione lessicale, cioè *Formal Role, Constitutive Role, Agentive Role e Telic Role*. Le dimensioni Qualia vengono usate per cogliere aspetti diversi e multiformi del significato di una parola. Per esempio il Telic Role riguarda la funzione tipica di un'entità o l'attività caratteristica di una categoria di individui (esempio, la funzione prototipica di un professore è insegnare). L'Agentive Role riguarda, invece, il modo in cui un'entità è creata (esempio, naturalmente o artificialmente), mentre il Constitutive Role rappresenta la composizione o struttura interna di un'entità (per esempio, le sue parti o il materiale di cui è composta). In SIMPLE, è possibile discriminare fra i vari sensi delle parole calibrando l'uso dei diversi tipi di informazione resi disponibili dal modello. Per esempio, la figura 6 mostra una possibile caratte-

rizzazione di una porzione di spazio semantico associato alla parola "ala" il cui contenuto può essere articolato in quattro SemU che hanno in comune lo stesso tipo semantico (ovvero *PART*), ma che possono comunque essere distinte attraverso le relazioni che esse hanno con altre unità semantiche. Per esempio, se da una parte la SemU_1 e la SemU_3 sono simili per quanto concerne la dimensione della funzionalità (entrambe si riferiscono a entità usate per volare), sono distinte per quanto riguarda gli aspetti costitutivi, poiché la SemU_1 si riferisce a una parte di un aereo e la SemU_3 alla parte di un uccello ecc.. Nonostante si sia ancora lontani dal poter fornire rappresentazioni veramente soddisfacenti del contenuto di una parola, l'architettura di SIMPLE tenta di avvicinarsi alla complessità del linguaggio naturale fornendo un modello altamente espressivo e versatile per descrivere il contenuto linguistico.

I lessici computazionali devono essere concepiti come sistemi dinamici il cui sviluppo si integra strettamente con processi di acquisizione automatica di informazione dai testi. Dal momento che i significati delle parole vivono, crescono e mutano nei contesti linguistici in cui occorrono, la loro rappresentazione nei repertori lessicali deve tenere necessariamente in considerazione le modalità con le quali l'informazione lessicale emerge dal materiale testuale e come quest'ultimo contribuisce alla creazione e alla variazione del significato. Conseguentemente, i lessici computazionali – anche di grandi dimensioni – non possono essere mai concepiti come repertori statici e chiusi. Al contrario, i lessici computazionali sono in grado al più di fornire nuclei di descrizione semantica che comunque devono essere costantemente personalizzati, estesi e adattati a diversi domini, applicazioni, tipologie di testo ecc.. In questo senso, il processo di creazione di risorse semantico-lessicali si deve accompagnare allo sviluppo di strumenti e metodologie per il *lexical tuning*, ovvero per l'adattamento dell'informazione semantica ai concreti contesti d'uso [4]. Questa sembra essere una condizione essenziale affinché le risorse linguistiche possano diventare strumenti versatili e adattativi per l'elaborazione del contenuto semantico dei documenti.

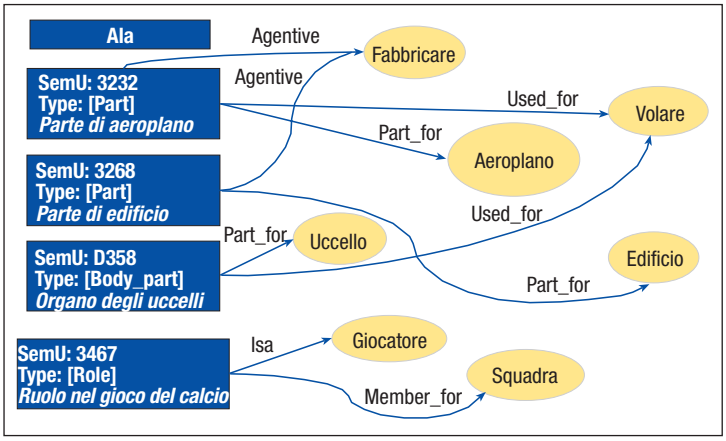


FIGURA 6

Rappresentazione dei significati di ala in SIMPLE

Gli strumenti per affrontare questo problema vengono dalla ricerca sull'*acquisizione automatica della conoscenza* e, più in generale, dall'uso di tecniche di *apprendimento automatico*, sia supervisionato che non supervisionato. Molti di questi metodi sono basati su un modello distribuzionale del significato, secondo il quale il contenuto semantico di una parola o termine è derivabile dal modo in cui esso si distribuisce linguisticamente, ovvero dall'insieme dei contesti in cui è usato [16]. Secondo questo approccio, a ciascuna parola di un testo viene associata una rappresentazione in forma di vettore distribuzionale. Le dimensioni dei vettore sono date dalle dipendenze grammaticali del termine con altri termini lessicali (verbi, nomi, aggettivi ecc.) nei documenti, oppure più semplicemente dalle parole che occorrono con il termine all'interno di una certa finestra di contesto. I vettori distribuzionali vengono generalmente estratti dai testi in maniera automatica con gli strumenti del TAL. Attraverso l'applicazione di algoritmi di *clustering* alle rappresentazioni vettoriali è possibile ricostruire spazi di similarità semantica tra i termini, ovvero classi di termini o parole semanticamente simili [17]. Infatti, il grado di similarità semantica tra due termini è proporzionale al grado di similarità della loro distribuzione grammaticale nei testi. In questo modo, è possibile arricchire ed estendere le risorse lessicali con nuove informazioni sul comportamento semantico delle parole e che direttamente rispecchiano il loro l'uso nei testi. Nuovi sensi o usi specifici di un certo domi-

nio o registro linguistico sono, quindi, derivabili automaticamente attraverso l'uso combinato del TAL e di algoritmi di apprendimento. Una maggiore comprensione dei problemi riguardanti le profonde interrelazioni tra rappresentazione e acquisizione del significato dei termini lessicali potrebbe avere importanti ripercussioni su come le risorse linguistiche verranno in futuro costruite, sviluppate e usate per le applicazioni.

5. STANDARD PER LE RISORSE LINGUISTICHE

Un altro aspetto di fondamentale importanza per il ruolo delle risorse lessicali (e più in generale linguistiche) nel TAL è come ottimizzare la produzione, mantenimento e interscambio tra le risorse linguistiche, così come il processo che porta alla loro integrazione nelle applicazioni. La precondizione essenziale per raggiungere questi risultati è stabilire una struttura comune e standardizzata per la costruzione dei lessici computazionali che possa garantire la codifica dell'informazione linguistica in maniera tale da assicurare la sua riutilizzazione da parte di applicazioni diverse e per compiti diversi. In questo modo, si può rafforzare la condivisione e la riusabilità delle risorse lessicali multilingui promuovendo la definizione di un linguaggio comune per la comunità degli sviluppatori e utilizzatori di lessici computazionali. Un'importante iniziativa internazionale in questa direzione è stata rappresentata dal progetto ISLE (*International Standards for Language Engineering*) [6], continuazione di EAGLES (*Expert Advisory Group for Language Engineering Standards*), ambedue ideati e coordinati da Antonio Zampolli. ISLE è stato congiuntamente finanziato dall'Unione Europea e dal *National Science Foundation* (NSF) negli USA e ha avuto come obiettivo la definizione di una serie di standard e raccomandazioni in tre aree cruciali per le tecnologie della lingua:

1. lessici computazionali multilingui,
2. interattività naturale e multimedialità,
3. valutazione.

Per quanto riguarda il primo tema, Il *Computational Lexicon Working Group* (CLWG) di ISLE si è occupato di definire consensualmente un'infrastruttura standardizzata per lo sviluppo di risorse lessicali multilingui per le applicazioni del TAL, con particolare riferimento alle specifiche necessità dei sistemi di traduzione automatica e di *Crosslingual Information Retrieval*. Nel corso della sua attività, ISLE ha fatto suo il principio metodologico secondo il quale il processo di standardizzazione, nonostante per sua natura non sia intrinsecamente innovativo, deve comunque procedere a stretto contatto con la ricerca più avanzata. Il processo di standardizzazione portato avanti da ISLE ha, infatti, perseguito un duplice obiettivo:

1. la definizione di standard sia a livello di contenuto che di rappresentazione per quegli aspetti dei lessici computazionali che sono già ampiamente usati dalle applicazioni;
2. la formulazione di raccomandazioni per le aree più di "frontiera" della semantica computazionale, ma che possono comunque fornire un elevato contributo di innovazione tecnologica nel settore del TAL.

Come strumento operativo per raggiungere questi obiettivi, il CLWG di ISLE ha elaborato MILE (*Multilingual ISLE Lexical Entry*), un modello generale per la codifica di informazione lessicale multilingue.

MILE è uno schema di entrata lessicale caratterizzata da un'architettura altamente *modulare e stratificata* [6]. La modularità riguarda l'organizzazione "orizzontale" di MILE, nella quale moduli indipendenti ma comunque correlati coprono diverse dimensioni del contenuto lessicale (monolingue, multilingue, semantico, sintattico ecc.). Dall'altro lato, al livello "verticale" MILE ha adottato un'organizzazione stratificata per permettere vari gradi di granularità nelle descrizioni lessicali. Uno degli scopi realizzativi di MILE è stato quello di costruire un ambiente di rappresentazione comune per la costruzioni di risorse lessicali multilingui, allo scopo di massimizzare il riutilizzo, l'integrazione e l'estensione dei lessici computazionali monolingui esistenti, fornendo al tempo stesso agli utilizzatori e sviluppatori di risorse linguistiche una struttura formale per la codifica e l'interscambio dei dati. ISLE ha, dunque, cercato di promuovere la creazione di un'infrastruttura per i lessici computazionali intesi come risorse di dati linguistici aperte e distribuite. In questa prospettiva, MILE agisce come un meta-modello lessicale per facilitare l'intero-

perabilità a livello di contenuto in due direzioni fondamentali: interoperabilità tra risorse linguistiche, per garantire la riusabilità e integrazione dei dati e interoperabilità tra risorse linguistiche e sistemi del TAL che devono accedere ad esse.
Il *ruolo infrastrutturale* delle risorse linguistiche nell'ambito del TAL richiede che esse vengano armonizzate con le risorse di altre lingue, valutate con metodologie riconosciute a livello internazionale, messe a disposizione della intera comunità nazionale, mantenute e aggiornate tenendo conto delle sempre nuove esigenze applicative. All'interno di questo contesto si inserisce, oggi, il disegno, promosso da chi scrive di un "cambiamento di paradigma" nella produzione e uso di una nuova generazione di risorse e strumenti linguistici, concepiti come *Open Linguistic Infrastructure*, attraverso l'utilizzo di metadati e di standard che permettono la condivisione di tecnologie linguistiche sviluppate anche in ambiti diversi, e il loro uso distribuito in rete. Questa nuova concezione è anche determinante per realizzare appieno la visione del *Semantic Web*, ovvero l'evoluzione del web in uno spazio di contenuti effettivamente "comprensibili" dal calcolatore e non solo da utenti umani e con accesso multilingue e multiculturale.

6. CONCLUSIONI E PROSPETTIVE

Una delle priorità a livello nazionale ed europeo è costruire una società basata sulla informazione e sulla conoscenza. La *lingua è veicolo e chiave di accesso alla conoscenza*, e oggi più che mai è urgente la realizzazione di una infrastruttura consolidata di tecnologie linguistiche. Gli sviluppi recenti nel TAL e la crescente diffusione di contenuti digitali mostrano che i tempi sono maturi per una svolta nella capacità di elaborare grandi quantità di documenti testuali al fine di renderli facilmente accessibili e usabili per un'utenza sempre più vasta e composita.
Alcuni temi su cui articolare il TAL per una società della conoscenza sono:
1. *accesso "intelligente" all'informazione multilingue e trattamento del "contenuto" digitale* - è urgente aumentare la disponibilità di strumenti e risorse capaci di automatizzare le operazioni linguistiche necessarie per produrre, organizzare, rappresentare, archiviare, recuperare, elaborare, navigare, acquisire, accedere, visualizzare, filtrare, tradurre, trasmettere, interpretare, utilizzare, in una parola *condividere* la conoscenza;
2. *interattività naturale e interfacce intelligenti* - si devono sviluppare sistemi che agevolino la naturalezza dell'interazione uomo-macchina e aiutare la comunicazione interpersonale mediando l'interazione tra lingue diverse;
3. *il patrimonio culturale e il contenuto digitale* - le tecnologie del TAL favoriscono la crescita dell'industria dei "contenuti", con ampie opportunità per un Paese, come l'Italia, tradizionale produttore di industria culturale:
4. *promozione della ricerca umanistica nella società dell'informazione* - le tecnologie del TAL forniscono nuovi strumenti anche per le scienze umanistiche, facilitando la produzione e fruizione dei contenuti culturali, e evidenziano il contributo potenziale anche delle ricerche umanistiche sul piano delle opportunità economiche e dello sviluppo sociale.
Per realizzare l'obiettivo di un accesso avanzato al contenuto semantico dei documenti è necessario affrontare la complessità del linguaggio naturale. L'attuale esperienza nel TAL dimostra che una tale sfida si può vincere solo adottando un approccio interdisciplinare e creando un ambiente altamente avanzato per l'analisi computazionale della lingua, l'acquisizione di conoscenze attraverso l'elaborazione automatica dei testi e lo sviluppo di una nuova generazione di risorse linguistiche basate sul rappresentazioni avanzate e standardizzate del contenuto lessicale.

Bibliografia

[1] Bartolini R., Lenci A., Montemagni S., Pirrelli V.: *Grammar and Lexicon in the Robust Parsing of Italian: Towards a Non-Naïve Interplay*. Proceedings of the Workshop on Grammar Engineering and Evaluation, COLING 2002 Post-Conference Workshop, Taipei, Taiwan, 2002.

[2] Bartolini R., Lenci A., Montemagni S., Pirrelli V.: *Hybrid Constraints for Robust Parsing: First Experiments and Evaluation*. Proceedings of LREC 2004, Lisbona, Portugal, 2004.

[3] Bartolini R., Lenci A., Montemagni S., Pirrelli, V., Soria C.: Semantic *Mark-up of Italian Legal Texts through NLP-based Techniques*. Proceedings of LREC 2004, Lisbona, Portugal, 2004.

[4] Basili R., Catizone R., Pazienza M-T., Stevenson M., Velardi P., Vindigni M., Wilks Y.: *An Empirical Approach to Lexical Tuning*. Proceedings of the LREC1998 Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, Granada, Spain, 1998.

[5] Busa F., Calzolari N., Lenci A., Pustejovsky J.: *Building a Semantic Lexicon: Structuring and Generating Concepts*. In Bunt H., Muskens R., Thijsse E. (eds.): Computing Meaning Vol. II. Kluwer, Dordrecht, 2001.

[6] Calzolari N., Bertagna F., Lenci A., Monachini M.: *Standards and best Practice for Multilingual Computational Lexicons and MILE (Multilingual ISLE Lexical Entry)*. ISLE deliverables D2.2 – D3.2 http://lingue.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm, 2003.

[7] Fellbaum C., (ed.): *WordNet. An Electronic Lexical Database*. MIT Press, Cambridge (MA), 1998.

[8] Gómez-Pérez A., Manzano-Macho D.: *A Survey of Ontology Learning Methods and Techniques*. Ontoweb Deliverable 1.5 http://ontoweb.aifb.uni-karlsruhe.de/About/Deliverables, 2003.

[9] Gruber T.R.: A Translation Approach to Portable Ontologies. *Knowledge Acquisition*, Vol. 5, 1993.

[10] Hepple M., Ireson N., Allegrini P., Marchi S., Montemagni S., Gomez Hidalgo J.M.: *NLP-enhanced Content Filtering within the POESIA Project*. Proceedings of LREC 2004, Lisbona, Portugal, 2004.

[11] Jackson P, Moulinier I.: *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins, Amsterdam, 2002.

[12] Joscelyne A., Lockwood R.: *Benchmarking HLT Progress in Europe*. HOPE, Copenhagen, 2003.

[13] Jurafsky D., Martin J.H.: *Speech and Language Processing*. Prentice Hall, Upper Saddle River (NJ), 2000.

[14] Lenci A., Bel N., Busa F., Calzolari N., Gola E., Monachini M., Ogonowsky A., Peters I., Peters W., Ruimy N., Villegas M., Zampolli A.: SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, Vol. 13, 2000.

[15] Lenci A., Montemagni S., Pirrelli V.: CHUNK-IT. An Italian Shallow Parser for Robust Syntactic Annotation. *Linguistica Computazionale*, Vol. 16-17, 2003.

[16] Lenci A., Montemagni S., Pirrelli V., (eds.): *Semantic Knowledge Acquisition and Representation, Giardini Editori*. Pisa, in stampa.

[17] Lin D., Pantel P.: *Concept Discovery from Text*. Proceedings of the Conference on Computational Linguistics 2002, Taipei, Taiwan, 2002.

[18] Maedche A., Staab S.: Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*, Vol. 16, 2001.

[19] Manning C.D., Sch_tze H.: *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge (MA), 1999.

[20] Pustejovsky J.: *The Generative Lexicon*. MIT Press, Cambridge (MA), 1995.

[21] Ruimy N., Corazzari O., Gola E., Spanu A., Calzolari N., Zampolli A.: *The European LE-PAROLE Project: The Italian Syntactic Lexicon*. Proceedings of the LREC1998, Granada, Spain, 1998.

[22] Staab S., Studer R. (eds.): *Handbook of Ontologies*. Springer Verlag, Berlin, 2003.

[23] Vossen P.: Introduction to EuroWordNet. *Computers and the Humanities*. Vol. 32, 1998.

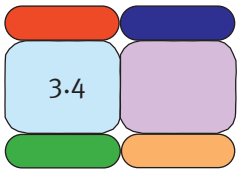
Nicoletta Calzolari è direttore dell'Istituto di Linguistica Computazionale del CNR di Pisa. Lavora nel settore della Linguistica Computazionale dal 1972. Ha coordinato moltissimi progetti nazionali, europei e internazionali, è membro di numerosi Board Internazionali (ELRA, ICCL, ISO, ELSNET ecc.), Conference Chair di LREC 2004, *invited speaker* e membro di Program Committee dei maggiori convegni del settore.
glottolo@ilc.cnr.it

Alessandro Lenci è ricercatore presso il Dipartimento di Linguistica dell'Università di Pisa e docente di Linguistica Computazionale. Ha conseguito il perfezionamento alla Scuola Normale Superiore di Pisa e collabora con l'Istituto di Linguistica Computazionale del CNR. Autore di numerose pubblicazioni, i suoi interessi di ricerca riguardano la semantica computazionale, i metodi per l'acquisizione lessicale, e le scienze cognitive.
alessandro.lenci@ilc.cnr.it

GRID COMPUTING: DA DOVE VIENE E CHE COSA MANCA PERCHÉ DIVENTI UNA REALTÀ?

Mauro Migliardi

Il Grid Computing rappresenta la frontiera della ricerca nel campo delle architetture di calcolo parallelo. Questo termine rappresenta la formulazione di un'idea, quella della condivisione delle risorse di calcolo, sviluppatasi negli ultimi tre decenni a seguito della rapida evoluzione delle tecnologie informatiche. Verrà descritta l'evoluzione di questo modello di calcolatore "non convenzionale" nelle soluzioni presenti e passate identificando quali sono i problemi ancora irrisolti.



1. INTRODUZIONE

Nel campo delle architetture di calcolo parallelo, il Grid Computing rappresenta oggi, senza alcun dubbio, la frontiera delle attività di ricerca. Tale fatto è testimoniato sia dal numero di progetti di ricerca dedicati a tale parola chiave [1], sia dal numero di grandi ditte di sistemi di calcolo, *software* e informatica in generale, che si sono dotate di una loro "soluzione Grid" [2, 3, 4, 5], sia dalle iniziative nazionali [6] o sovranazionali [7] connesse alla parola chiave "Grid". Tuttavia, questo termine, oggi così attuale, è, se si considerano i progetti di ricerca e l'evoluzione delle tecnologie informatiche nel loro complesso, solo la punta di un iceberg. Il termine "Grid", infatti, rappresenta la formulazione particolarmente fortunata di un'idea, quella della condivisione delle risorse di calcolo, sviluppatasi faticosamente negli ultimi tre decenni a seguito della rapida evoluzione delle tecnologie informatiche. Negli ultimi dieci anni, poi, a partire dai primi esperimenti di macchine parallele virtuali realizzate su reti locali sino a giungere a concezioni avveniristiche quali la potenza di calcolo su richiesta o l'integrazione dinamica di componenti simulativi sviluppati indipendentemente, un filo conduttore può essere individuato nella necessità di superare il problema dell'eterogeneità, sia a livello *hardware* che a livello software, al fine di fornire un'immagine integrata del sistema. Questo problema dell'eterogeneità, tuttavia, si presenta in nuove forme ogni qualvolta, trovata una soluzione accettabile per dominare il livello di eterogeneità precedentemente affrontato, si voglia superare la concezione precedente per aprire a nuove idee e possibilità.

Scopo del presente articolo è fornire una visione prospettica dell'evoluzione dei sistemi informatici dedicati al calcolo parallelo/distribuito. In quest'ottica, si forniranno al lettore una tassonomia del fenomeno Grid, una panoramica dell'evoluzione dei sistemi di condivisione delle risorse di calcolo collegata alle problematiche e agli sviluppi tecnologici che l'hanno guidata e un'analisi delle problematiche ancora aperte e che necessitano di trovare una soluzione prima di poter

osservare un effettivo avvento del Grid Computing nel tessuto economico e sociale del mondo tecnologicamente evoluto.

2. UNA TASSONOMIA DEL FENOMENO GRID

Secondo la definizione del progetto Globus, uno dei primi, più importanti e, dal punto di vista dei finanziamenti ottenuti, più fortunati progetti relativi al Grid Computing, "Le Grid sono ambienti persistenti che rendono possibile realizzare applicazioni software che integrino risorse di strumentazione, di visualizzazione, di calcolo e di informazione che provengono da domini amministrativi diversi e sono geograficamente distribuite[1]". La definizione di Grid si basa, quindi, sul concetto di condivisione di risorse di diverse tipologie e si presta, quindi, a categorizzare diverse tipologie di Grid: le *Grid computazionali* [21, 36], le *Grid di dati* [7, 8, 36], le *Grid di applicazioni e servizi* [36], le *Grid di strumentazione* [9, 18] e, infine, le *Grid di sensori* [10, 29, 37].
Una Grid computazionale è l'aggregazione di risorse di calcolo provenienti da domini di sicurezza e gestione differenti. Tale aggregazione è finalizzata a fornire a un insieme di utenti potenza di calcolo *on-demand*, in modo disaccoppiato dalla provenienza, cioè dai nodi che la stanno fisicamente fornendo. Si consideri, ad esempio, il caso di una società multinazionale con sedi sparse in tutto il mondo. Ognuna di queste sedi possiede, oggi, un parco di architetture di calcolo dimensionato o secondo un *worst-case*, dimensionato cioè per soddisfare non le necessità medie ma i picchi di richiesta, oppure dimensionato per soddisfare le necessità medie e, quindi, non in grado di gestire le situazioni di picco.
Una struttura Grid in grado di sopperire alle richieste di picco con potenza computazionale proveniente dalle altre sedi della società permetterebbe di dimensionare in modo più consono alle reali necessità quotidiane il parco macchine realizzando così un notevole risparmio. Allo stesso tempo, lo stesso meccanismo di raccolta della potenza di calcolo sull'intero insieme delle risorse di calcolo della società permette di realizzare un supercalcolatore disponibile dinamicamente nel momento del bisogno. Le Grid computazionali, rappresentando una linea evolutiva delle architetture di calcolo, hanno una storia particolarmente ricca e variegata, sono quindi quelle di cui maggiormente si parlerà nel presente articolo.
Le Grid di dati possono essere considerate una delle forme evolutive del Web. Infatti, come il Web, nascono per contenere grandi moli di dati distribuite in domini differenti per gestione e posizione geografica. A differenza del Web, dove, pur data la ricchezza delle informazioni presenti, mancano sia meccanismi espliciti di standardizzazione dei significati associati a queste informazioni, sia strumenti di elaborazione concorrente delle informazioni ottenute, le Grid di dati coniugano la ricchezza delle sorgenti con gli strumenti adatti a far sì che le informazioni presenti possano essere facilmente correlate e vengano, quindi, a dotarsi di un alto valore aggiunto rispetto al mero contenuto. Si consideri, a titolo di esempio, il caso delle basi di dati contenenti i casi clinici dei singoli ospedali nel mondo. Ognuno di queste basi di dati possiede, già presa singolarmente, un fortissimo valore in quanto permette di analizzare quali siano state nel passato le decisioni dei clinici. Tale valore però, aumenta in modo non lineare se, aggregando le basi di dati in una Grid di dati, si rende possibile analizzare, confrontare e correlare le decisioni prese da clinici di scuole diverse in presenza di patologie uguali o, quantomeno, comparabili.
Le Grid di applicazioni e/o servizi, rappresentano una delle visioni più futuribili nel campo del Grid Computing. Esse, infatti, non si limitano a essere una versione globalizzata all'intera Internet del concetto di *Application Service Provider* (ASP), cioè della possibilità di affittare un certo tempo di esecuzione di una specifica applicazione su di un *server* remoto, ma contemplano anche la aggregazione di componenti applicativi sviluppati indipendentemente al fine di realizzare nuove e originali applicazioni. Si consideri, a titolo di esempio, il seguente caso. La costruzione di una nave è un processo che ormai non coinvolge più il solo settore cantieristico in senso stretto, ma, al contrario, richiede l'intervento di un largo insieme di fornitori dedicati ai singoli sottosiste-

[1] Fonte: http://www.globus.org

mi. Lo sviluppo di tali sottosistemi, per esempio la parte motoristica, la parte di stabilizzazione o la parte dedicata alla sensoristica di navigazione, viene affidato esternamente all'ente cantieristico nominalmente responsabile della costruzione navale. In uno scenario tradizionale, ognuno di questi fornitori ha una visione limitata al sottosistema di cui è incaricato e si viene così a perdere la visione olistica del prodotto finale sino al momento in cui, ottenuti tutti i singoli sottosistemi dai diversi fornitori, l'ente cantieristico sarà in grado di assemblare il prodotto finito. Questo non permette di valutare all'interno del *loop* di progettazione gli effetti dovuti alle interazioni tra i diversi sottosistemi, e rischia, quindi, di portare a scoprire indesiderati effetti collaterali solo in fase di collaudo dell'intero sistema. Una soluzione a questa *empasse* è ottenibile tramite l'integrazione in una Grid di applicazioni di elementi simulativi relativi ai diversi sottosistemi. Tuttavia, elementi simulativi sviluppati secondo lo standard internazionale attuale, cioè (HLA) *High Level Architecture* [11], non supportano l'assemblaggio dinamico a *run-time* di un sistema completo, hanno un supporto limitato per l'esecuzione distribuita e non permettono di schermare completamente agli altri utenti i dettagli interni al componente. Al contrario, una Grid di applicazioni deve permettere a ciascun fornitore:

- di mantenere il pieno controllo del componente simulativo relativo al suo prodotto;
- di esporre solo un'interfaccia opaca;
- di assemblare l'intero sistema nave in una simulazione completa.

In una situazione di questo genere, ovviamente, una Grid di applicazioni sussume al suo interno sia una Grid computazionale che una Grid di dati. Infatti, le risorse di calcolo utilizzate per eseguire i diversi componenti applicativi sono aggregate dinamicamente a partire da domini di gestione diversi. Contemporaneamente, i dati su cui si trova a operare l'applicazione complessivamente costruita provengono da sorgenti differenti e sono correlate tramite l'esecuzione coordinata delle componenti applicative.
Una Grid di strumentazione consiste nella generalizzazione del concetto di accesso remoto ad apparati di costo elevato o per rendere fruibile in forma remota un *setup* sperimentale di elevato valore didattico. In una Grid di strumentazione, apparati gestiti da enti diversi possono essere integrati per aggregare esperimenti complessi non possibili in nessuno dei singoli siti coinvolti nella Grid o per condividere strumentazione didattica in modo da rendere disponibile a corsi di studi afferenti ad atenei diversi una struttura condivisa per la realizzazione di esperimenti di comprovato valore didattico.
Il concetto di Grid di sensori proviene, principalmente, dal mondo militare. In ambito militare, tale concetto sta alla base della moderna teoria del combattimento come attività incentrata sull'interconnessione delle componenti dell'esercito belligerante. In tale visione, infatti, si parla della Grid di sensori come del livello atto a fornire la visione dettagliata dello stato corrente agli agenti operativi e decisionali. Recentemente, tuttavia, l'idea di una Grid di sensori destinata a fornire una visione dettagliata e a prova di guasto della situazione del campo è penetrata anche in ambiti non militari quali quelli dell'agronomia e della salvaguardia del territorio. Si consideri, a titolo di esempio, il caso di un insieme di sensori all'infrarosso sparsi per una foresta con lo scopo di segnalare tempestivamente lo svilupparsi di focolai di incendio. L'integrazione delle informazioni provenienti da un grande numero di sensori permette di effettuare una migliore separazione del calore ambientale dal calore proveniente da un focolaio d'incendio e, quindi, garantisce una consistente diminuzione della sensibilità del sistema ai falsi positivi. Le Grid di sensori, sono il ramo del Grid Computing che maggiormente si avvicina al concetto di *Ubiquitous Computing*, infatti, tutte le attività di ricerca dedicate a questa tipologia di Grid pongono una fortissima attenzione alle problematiche legate alle reti *wireless*, all'instradamento autonomo e alle reti senza infrastruttura statica, cioè le cosiddette reti *ad hoc*.

3. UN PO' DI STORIA

Per valutare il livello di tumultuosità evolutiva delle tecnologie dell'informazione, si considerino, per esempio, questi due parametri:

- il numero di transistori e la frequenza di *clock* del singolo processore (Figura 1 **A** e **B**), a rappresentare la sua capacità elaborativa;

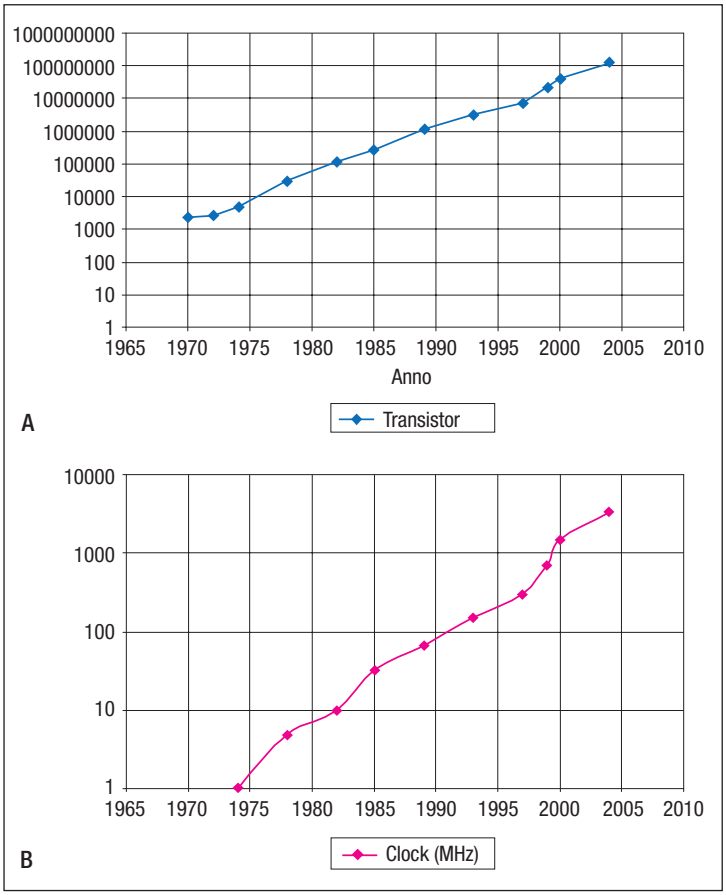


FIGURA 1

A Numero di transistor presenti su singola CPU. B Frequenza tipica di clock

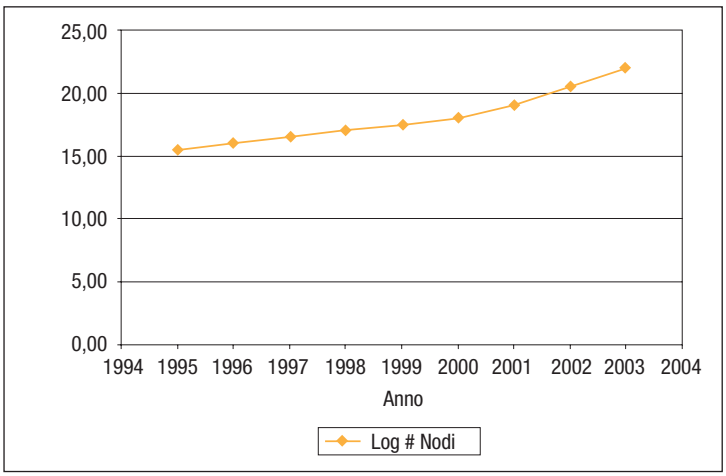


FIGURA 2

Numero di calcolatori interconnessi tramite Internet

il numero di calcolatori connessi a Internet a rappresentare la capacità di cooperare (Figura 2).

Questo aumento ha guidato la maccanica dell'evoluzione dei sistemi informatici in generale e dei meccanismi di condivisione delle risorse di calcolo in particolare.

Se si considera il fenomeno della condivisione delle risorse di calcolo, si possono identificare quattro "ere" (Tabella 1) caratterizzate da distinte metodologie e architetture dedicate al supercalcolo.

La prima era è caratterizzata dal concetto "un computer per molti utenti". In questa era il costo della singola risorsa di calcolo è tale da essere accessibile solo a entità grandi, in grado di ammortizzare l'investimento attraverso l'uso contemporaneo della risorsa da parte di numerosi utenti. In questa era, esisteva un ampio insieme di architetture dedicate al supercalcolo basate su soluzioni completamente *ad hoc* sviluppate sia da ditte specializzate che dai grandi venditori di hardware generico. Negli anni '80, la discesa dei costi dell'hardware porta all'inizio del decennio ad avere sistemi di lavoro dedicati a singoli utenti e alla fine del decennio ad avere computer personali anche per usi non strettamente tecnici (per esempio, *desktop publishing* e *office automation*). Contemporaneamente, il campo delle architetture di supercalcolo conosceva la stagione delle macchine SIMD (*Single Instruction Multiple Data*) [20] e una fioritura di ditte specializzate nella produzione di architetture di supercalcolo basate su tale paradigma. Tuttavia, già alla fine degli anni '80, l'attenzione di alcuni gruppi di ricercatori inizia a focalizzarsi sulla possibilità di sfruttare l'accresciuta potenza di calcolo disponibile su una singola *workstation* e la disponibilità di tecnologie a basso costo atte a permettere l'interconnessione in rete locale di tali workstation per la costruzione di macchine parallele virtuali [39].

Gli anni '90 sanciscono l'avvento del concetto di "Molti calcolatori per un singolo utente". La definitiva affermazione della legge di Moore [26, 34] nel campo dei *personal computer*, porta a rendere economicamente impraticabile ogni tentativo di sviluppare architetture di calcolo *ad hoc*. Infatti, come testimoniato dal fallimento di esperienze quali quella della LISP (*List Processing*) Machine [27, 30], il tempo richiesto dal ciclo di progettazione, test e produzione di un'architettura dedicata risulta essere nettamente superiore a quello richiesto dalle CPU *mainstream* per raggiungere prestazioni tali da permettere a sistemi software eseguiti su di esse di superare nettamente le

Era	Relazione tra calcolatori e utenti	Architettura di condivisione	Architetture di super calcolo
Anni '70	Uno a molti	Sistemi Time sharing	Supercalcolatore
Anni '80	Uno a uno	PC e Workstation	Supercalcolatore
Anni '90	Molti a uno	Clusterizzazione	COW
Terzo millennio	Molti a molti	Grid	Grid

TABELLA 1
Ere computazionali

prestazioni dei sistemi hardware dedicati. Allo stesso modo, l'enorme incremento prestazionale dimostrato con scadenza annuale dalle CPU mainstream esclude dal mercato tutte le architetture di supercalcolo basate su processori speciali, come le architetture SIMD. Contemporaneamente, gli anni '90 testimoniano il definitivo affermarsi come tecnologie mainstream di quelle che, a fianco dell'incredibile evoluzione delle CPU, rappresentano la seconda colonna portante dell'idea stessa di Grid: le reti di calcolatori e Internet.

3.1. Dal Supercalcolatore al Cluster di Workstation

Il passo successivo viene preannunciato dal colpo di coda della Thinking Machine Corporation. Nel 1993, dopo un decennio di ardito pionierismo nel campo delle architetture SIMD, per tentare di salvare un futuro che, sfumati i grandi finanziamenti militari per la guerra fredda, appare assai oscuro, tenta la produzione della Connection Machine 5 [35], un supercalcolatore basato sull'interconnessione tramite una struttura proprietaria di pochi, potentissimi[2] processori mainstream: gli SPARC prodotti da SUN. I costi di produzione del sistema restano comunque troppo alti a causa delle molte componenti altamente customizzate, quali per esempio la rete di interconnessione dei processori, ancora presenti nell'architettura, e l'esperimento risulta, quindi, economicamente fallimentare[3].

Nel 1994, presso l'università di Stanford, il progetto Beowulf inizia una nuova fase sottolineando con forza il concetto di architettura basata su *Components Off The Shelf* (COTS) o *Commodity Computing* (CC). Questo concetto di Commodity Computing, cioè di calcolo come servizio di base, è figlio dell'alleanza Intel/Microsoft che negli anni '90 porta alla realizzazione della visione "Un PC su ogni scrivania". Infatti, tra le caratteristiche che rendono particolarmente significativo il primo cluster Beowulf non vi sono *performance* di picco o peculiarità architetturali, ma bensì:

- l'uso come nodi computazionali di macchine comparabili con quelle destinate all'*office automation* (PC con processore Intel DX4);
- l'uso come tecnologia di interconnessione tra i nodi dello standard di rete progettato per l'*office automation* (*ethernet*);
- l'uso di un sistema operativo *open source* di uso generale, non specificamente progettato per il calcolo ad alte prestazioni (Linux o FreeBSD).

Nella seconda metà degli anni '90, nel mondo del calcolo ad alte prestazioni la ricerca verte in larga misura su architetture cluster, cluster di workstation (COW) e cluster di PC[4], cioè architetture caratterizzate dall'aggregazione statica di nodi di computazione dedicati completamente all'architettura cluster e interconnessi tramite un'interfaccia di rete di tipo standard.

È fondamentale notare che, tra la fine degli anni '80 e l'inizio degli anni '90, il processo che

[2] Si parla, ovviamente, in termini relativi ai primi anni '90.

[3] Nel vero senso della parola: nonostante la CM-5 abbia l'onore di apparire in un film di successo popolare quale *Jurassic Park*, la Thinking Machine Corporation dichiara fallimento nello stesso 1993.

[4] L'Università del Tennessee in Knoxville mantiene una lista dei 500 supercomputer più performanti al mondo che viene aggiornata due volte all'anno. Il primo cluster appare nella lista nell'anno 1995. Nel 2003 sette dei Top 10 sono architetture cluster.

Si parla di ***Single System Image*** (SSI) quando, in un ambiente costituito da risorse eterogenee e/o distribuite, si cerca di costruire a beneficio dell'utente un'interfaccia che unifichi la gestione e l'utilizzo di queste risorse mascherandone le differenti caratteristiche. È importante sottolineare come la definizione del concetto di SSI non dipenda da disomogeneità in termini di prestazioni, ma solo in termini di possibilità di utilizzo.
Un semplice esempio di SSI è quello fornito dalla condivisione delle risorse nei sistemi Microsoft. In tale versione di SSI, lo strato middleware incorporato nei sistemi Windows si occupa di mascherare la distribuzione e le disomogeneità delle risorse (per esempio, quelle derivanti da versioni diverse del sistema operativo) e le visualizza all'utente finale con un'interfaccia unificata. Tuttavia, la SSI fornita dal meccanismo di condivisione delle risorse non è assolutamente inficiata dalla presenza di sistemi più lenti in quanto l'unificazione cui si mira è meramente di interfaccia e non si applica alle prestazioni.

culmina con l'affermazione dei concetti COTS e CC segna un cambiamento sostanziale nella ricerca dedicata ai sistemi di calcolo. Se nel precedente ventennio tale ricerca era stata orientata principalmente all'hardware, essa diventa ora sostanzialmente orientata al software. Infatti, anche lo sviluppo di sistemi per il calcolo ad alte prestazioni risulta essere principalmente mirato alla definizione di strati software (middleware[5]) atti a mascherare l'eterogeneità delle risorse di calcolo costruendo in modo più o meno marcato una ***Single System Image***. Per i sistemi cluster di tipo Beowulf, il livello di SSI è dato fondamentalmente da:

- la possibilità di compilare un programma eseguibile sull'intero sistema cluster;
- la possibilità di schedulare l'esecuzione di un'applicazione su tutti o un sottoinsieme dei nodi del cluster tramite un sistema *batch*;
- la possibilità di monitorare ed eventualmente abortire le applicazioni in esecuzione.

3.2. Dal Cluster al Grid

L'inizio del passo successivo è segnato, nel 1996, da una fioritura di progetti mirati a definire sistemi per il *metacomputing*[6]. I sistemi di metacomputing possiedono molte caratteristiche in comune con i sistemi cluster, infatti, come questi ultimi, essi sono gruppi di nodo computazionali di tipo standard interconnessi tramite interfacce di rete di tipo standard. Tuttavia, esse possiedono in sè le caratteristiche seminali del concetto di Grid Computing, ovvero definiscono "sistemi formati dall'aggregazione dinamica di nodi computazionali non dedicati al metacomputer, interconnessi anche attraverso Internet e appartenenti a domini di sicurezza e gestione non omogenei". In tabella 2 sono schematizzate le principali differenze tra un sistema cluster e un sistema per il metacomputing.
Per due anni, il termine metacomputer o sistema di metacomputing viene utilizzato per rappresentare tutte le strutture di calcolo formate da nodi eterogenei distribuiti su Internet. Persino il progetto Globus, negli articoli pubblicati sino all'inizio dell'anno 1998, vie-

TABELLA 2
Principali differenze tra un sistema cluster e un sistema di Metacomputing

Sistemi Cluster	Sistemi di Metacomputing
Nodi computazionali dedicati completamente al cluster	Nodi computazionali aggregati dinamicamente su richiesta o base volontaristica
Nodi computazionali il più possibile omogenei fra loro	Nodi computazionali di qualunque genere
Rete di interconnessione dedicata al solo traffico dati delle applicazioni del cluster	Rete di interconnessione condivisa con traffico estraneo al metacomputer
Rete di interconnessione locale	Rete di interconnessione distribuita su Internet
Un singolo dominio di gestione e sicurezza	Più domini di gestione e sicurezza forniscono i nodi del metacomputer

[5] Secondo la base dati IEEE Explorer, il primo articolo scientifico relativo a middleware appare nel 1993. Nel 1994 gli articoli su middleware sono 7. Negli anni successivi circa 170 all'anno.

[6] Il termine metacomputing viene effettivamente coniato da L. Smarr, direttore NCSA, alla fine degli anni '80, tuttavia è solo nel 1997 che tale termine prende piede con il suo significato attuale. Nell'anno 1996 vengono pubblicati circa 150 articoli scientifici dedicati al metacomputing. Nel 1997, quasi 400 (Fonte: *Citeseer*).

Modello a scambio di messaggi
Il modello computazionale a scambio di messaggi viene introdotto negli anni '80 da C. A. R. Hoare. Esso si basa fondamentalmente su due primitive, *send* e *receive*, che permettono a processi strettamente sequenziali di comunicare e sincronizzarsi fra loro al fine di costruire un'applicazione concorrente o parallela di qualsivoglia complessità. Il modello a scambio di messaggi è quello più comunemente usato nel calcolo distribuito in quanto non richiede strutture complesse come la memoria condivisa, ma pone, invece, come unico vincolo architetturale la presenza di un canale di comunicazione tra tutte le componenti del sistema di calcolo.

ne descritto dai suoi autori "Un toolkit per un'infrastruttura di metacomputing" [22]. Tuttavia, tra l'inizio del 1998 [23] e la prima metà del 1999 [24], Ian Foster e Carl Kesselmann, i padri del progetto Globus, pubblicano la loro visionaria metafora di una "Griglia di distribuzione della potenza computazionale in cui, come i watt nella griglia di distribuzione dell'elettricità, la potenza computazionale può essere distribuita a chi ne fa richiesta senza badare alla sua provenienza". Questa futuristica visione ha un immediato successo e viene adottata dalla comunità scientifica sostituendo il termine metacomputing[7].

4. PROGRAMMARE IN PARALLELO

Sino a questo punto la nostra attenzione si è focalizzata principalmente sull'evoluzione delle architetture hardware che ha portato alla visione delle Grid. Si consideri ora il punto di vista complementare: cosa vuol dire scrivere programmi per una Grid computazionale?

4.1. Grid del Supercalcolo

Secondo la visione sponsorizzata dalla comunità del supercalcolo, una Grid computazionale consiste nell'interconnessione dei supercalcolatori disponibili nei diversi centri al fine di realizzare una struttura globale in grado di lanciare l'esecuzione dell'applicazione candidata su uno qualunque dei supercalcolatori parte della Grid che sia in grado di fornire le risorse (CPU, memoria e spazio disco) richieste dall'applicazione stessa. In tale visione, quindi, le applicazioni che devono essere scritte per una Grid computazionale sono sostanzialmente le stesse che venivano scritte per le architetture cluster in quanto la dinamicità del sistema Grid è, di fatto, limitata alla fase di scelta del supercalcolatore su cui lanciare l'esecuzione. In questa situazione, quindi, le tecniche di programmazione utilizzate sono quelle tradizionali dello scambio di messaggi e si basano, fondamentalmente, sull'uso delle librerie derivate dal modello CSP [19] quali PVM [39] e MPI [31]. Questa soluzione implica un modello di applicazione con un accoppiamento stretto tra le sue varie parti e, pur permettendo di ottenere livelli di prestazioni molto alti, pone vincoli tali da prevenire il completo utilizzo della natura dinamica di una Grid computazionale. Inoltre, a meno di accettare degradi prestazionali fortissimi, questo modello di applicazione richiede di ritagliare l'applicazione stessa sulle caratteristiche del sistema di calcolo che, quindi, non può essere soggetto alle variazioni dinamiche tipiche di una Grid computazionale. Infine, questo tipo di programmazione è estremamente complessa e richiede una conoscenza approfondita delle problematiche proprie del calcolo parallelo ad alte prestazioni. Quindi, pur essendo applicabile a una vasta gamma di tipologie di applicazione, la generazione di una nuova applicazione secondo il paradigma di scambio di messaggi o il trasporto di un'applicazione tradizionale allo stesso paradigma computazionale richiede uno sforzo di ingegnerizzazione o reingegnerizzazione estremamente elevato ed è affrontabile solo da un *pool* di programmatori esperti nelle problematiche proprie del calcolo parallelo ad alte prestazioni.
Per queste ragioni, questo tipo di soluzione, pur essendo già disponibile sul mercato, è di fatto appetibile solo per la grande industria, quella che possiede al suo interno uno staff IT di dimensioni tali da potersi permettere di addestrarne una parte in modo specifico e di averne una parte considerevole dedicata allo sviluppo e alla manutenzione di specifiche applicazioni.

[7] Nel 1999, il numero di articoli che parlano di metacomputing scende a 150, nel 2000 praticamente a 0. Contemporaneamente sale il numero di articoli relativi al Grid Computing. (Fonte: *Citeseer*).

4.2. Grid "Montecarlo"

Una seconda visione delle Grid computazionali è quella pionierizzata da progetti quali SETI@home [38] e realizzata al presente nei sistemi di ditte quali Entropia [12] o AVAKI [13] o dei grandi vendor di sistemi informatici citati all'inizio del presente articolo. In questa visione, un centro di controllo della Grid computazionale gestisce l'allocazione di parti dell'esecuzione globale dell'applicazione a nodi ottenuti dinamicamente in modo volontaristico, come nel caso di SETI@home, o identificando nodi disponibili all'interno di una intranet aziendale, come nel caso delle soluzioni commerciali. Le applicazioni generate in questo modo sono tutte di tipo Montecarlo, realizzate con un modello di computazione farmer-worker ad accoppiamento ridotto.

Questo tipo di soluzione permette la creazione di nuove applicazioni o il trasporto di applicazioni pre-esistenti verso piattaforme di tipo Grid computazionale con un costo di re-ingegnerizzazione piuttosto basso. Essa, infatti, non richiede una conoscenza estremamente approfondita delle problematiche proprie del calcolo parallelo ad alte prestazioni. Tuttavia, questa soluzione è adottabile solo per una specifica categoria di applicazioni, quelle appunto basate su simulazioni di tipo Montecarlo o, più in generale, quelle in cui è necessario eseguire una stessa sequenza di operazioni su di un enorme insieme di dati o casi di test che possono essere ripartiti in sottoinsiemi processabili in modo indipendente. L'indipendenza tra le diverse parti garantisce l'assenza di comunicazione tra i nodi che eseguono le operazioni. Questo permette di trascurare i problemi che potrebbero essere generati dall'eterogeneità delle interconnessioni tra i nodi della Grid computazionale e semplifica, quindi, sensibilmente l'applicazione.

Purtroppo, però, dati i vincoli sulla tipologia di applicazione utilizzabile con questa soluzione, essa non è generalizzabile a un ampio bacino di utenza ed è principalmente adottata per test di nuovi farmaci o ricerca di pattern all'interno di enormi moli di dati come nei casi del progetto SETI o dell'analisi del genoma umano.

4.3. Grid di Servizi

Una terza, nuova visione delle Grid computazionali, è quella di una Grid, orientata ai servizi, che utilizzi la tecnologia dei *Web Services*. Questa visione viene formulata da alcuni progetti di ricerca [8] ed è stata formalizzata nella prima metà dell'anno 2002 dal progetto *Open Grid Services Architecture* (OGSA), una *joint venture* tra IBM e il team del progetto Globus [14]. Tale visione si propone lo sviluppo di applicazioni capaci di utilizzare la tecnologia Grid-computing adottando un paradigma di computazione *multi-tier*, basato su servizi con interfacce standardizzate. Nel modello definito da OGSA, una Grid è sostanzialmente un'organizzazione virtuale in grado di permettere e coordinare l'uso di risorse condivise. Al fine di realizzare questa visione, OGSA definisce tre livelli di protocollo:

1. livello di connettività;
2. livello di controllo delle risorse;
3. livello di collettivizzazione delle risorse.

Il primo livello si occupa fondamentalmente di gestire la comunicazione tra le componenti della Grid e il controllo degli accessi, sia da parte di utenti finali che da parte di altre componenti della Grid, a ciascuna delle componenti stesse.

Il secondo livello, si occupa di definire le necessità di una richiesta ed effettuare un'associazione di tali necessità con le risorse atte a soddisfarle[8].

Il terzo livello, infine, si occupa di coordina-

Modello Farmer-Worker

Il modello computazionale Farmer-Worker viene comunemente utilizzato nel caso di applicazioni decomponibili in *task* di grana medio-grande (*Bag of Work*) tra di loro indipendenti, cioè che per il loro svolgimento non necessitano di comunicare con altri. Un Farmer genera un insieme di Bag of Work che vengono messe a disposizione dei Worker. Ogni Worker che si trova in stato ideale estrae una Bag of Work dall'insieme ed effettua l'esecuzione del task richiesto. Alla conclusione del task, il Worker restituisce i risultati ottenuti al Farmer e, trovandosi, quindi, ora nello stato ideale, torna a controllare se esiste altro lavoro da fare.

Questo meccanismo iterativo implementato dai Worker costituisce un naturale gestore del bilanciamento dei carichi in quanto un Worker più lento si limiterà a processare meno Bag of Work, mentre un Worker più efficiente se ne accaparrerà un numero maggiore.

Il modello Farmer-Worker può essere applicato sia ad applicazioni in cui esiste un solo passo di generazione delle Bag of Work, sia ad applicazioni in cui il Farmer, una volta raccolti i risultati ottenuti dal processamento di un insieme di Bag of Work, ne produca uno successivo (per esempio, applicazioni di simulazione a tempo discreto).

[8] Questo livello viene spesso definito "matchmaker", cioè paraninfo.

re l'utilizzo delle risorse condivise in modo da fornire un servizio di gestione collettiva di tali risorse.
Tale paradigma, facendo leva su tecnologie software mainstream quali Web Services, Java2 Enterprise Edition e Microsoft.NET, risulta essere più vicino alla sensibilità dei programmatori di applicazioni di tipo commerciale e presenta, quindi, una minor difficoltà di adozione da parte delle piccole medie imprese.
Una soluzione di questo genere, pur dovendo sacrificare una parte dell'ottimalità prestazionale raggiungibile solo con paradigmi di computazione dedicata alla computazione parallela quale quello di scambio di messaggi, permetterebbe di rendere fruibile la tecnologia Grid-computing anche a ditte medio-piccole non dotate di un gruppo di programmatori esperti nelle problematiche della computazione parallela. Una soluzione di questo genere permetterebbe, quindi, di:
- allargare il bacino di utenza della tecnologia Grid-computing rendendola disponibile non solo alla grande industria ma anche alle PMI;
- rendere disponibile alle PMI una potenza di calcolo decisamente superiore a quella correntemente a loro disposizione senza per altro richiedere grossi investimenti in hardware dedicato.

5. PROBLEMI APERTI

Purtroppo, il lavoro iniziato dal progetto OGSA è ben lungi dall'essere completato. Si consideri, a titolo di esempio, il caso del protocollo SOAP [9]. Questo protocollo, è il protocollo di riferimento per la trasmissione dati utilizzato dalle tecnologie Web Services e NET, tuttavia, le sue caratteristiche lo rendono del tutto inadatto a situazioni che richiedano alti livelli di prestazioni. Per questa ragione, il progetto OGSA ne suggerisce l'utilizzo soltanto per il livello di controllo delle risorse. Tuttavia, sebbene siano in corso diversi studi per identificare quali protocolli possano sostituire SOAP in ambienti ad alte prestazioni, tali studi sono ancora in fase sperimentale, quindi non esiste, ad oggi, una soluzione unificante che possa essere suggerita all'interno di OGSA. I modelli di programmazione per Grid computazionali disponibili sono, quindi, ad oggi, solo quello a scambio di messaggi e quello *farmer-worker* per applicazioni di tipo Montecarlo.
Un secondo problema aperto, è quello della gestione di un livello globale di eterogeneità all'interno delle risorse presenti in una Grid computazionale. Come spiegato in precedenza, il meccanismo di superamento del problema dell'eterogeneità e la costruzione di una SSI da super-imporre come colla unificante al di sopra delle diversità. Il progetto Globus, lo standard *de facto* per la costruzione di Grid di supercalcolo, ha definito un linguaggio di descrizione delle risorse che permette di descrivere in modo univoco le pur differenti caratteristiche dei nodi presenti in una Grid. Tuttavia, questa descrizione, seppure rappresenta una specie di SSI, si limita, di fatto, a demandare alle singole applicazioni la gestione dell'eterogeneità. Sono in corso esperimenti relativi alla costruzione estemporanea di Grid computazionali a partire da un insieme estremamente eterogeneo di risorse di calcolo, si consideri, per esempio, il caso di Flashmob1 [15] all'Università di San Francisco. Tuttavia, ad oggi, tutti i casi di successo per la costruzione di Grid computazionali provengono esclusivamente da due categorie:
- i casi in cui l'eterogeneità era resa innocua dalla natura stessa dell'applicazione (Grid montecarlo);
- i casi in cui l'eterogeneità viene limitata alla fase di identificazione delle risorse da usare al tempo dell'esecuzione (Grid di supercalcolo).

Allo stesso modo, risulta irrisolto il problema di garantire un livello di qualità del servizio in presenza di risorse che non solo sono eterogenee, ma cambiano la loro disponibilità nel tempo.
Infatti, se si escludono i due casi delineati sopra, ad oggi risulta praticamente impossibile garantire un livello di qualità del servizio alle applicazioni in mancanza di una quantificazione minima e massima delle risorse disponibili. In un ambiente dinamico come una Grid, un ulteriore livello di complessità è dato dalla necessità di scoprire le risorse che si rendono via via disponibili. La soluzione classica a questo tipo di problema è la realizzazione di un servizio di *directory*: tuttavia, la maggior parte dei servizi di directory oggi disponibili si basano su architetture con server centralizzato che,

ovviamente, mal si prestano ad una struttura potenzialmente globale come una Grid. Un esempio di directory globale e, in quanto completamente distribuita, estremamente scalabile, viene dal *Domain Name System* (DNS). Tuttavia, l'architettura del DNS è progettata per una frequenza di cambiamento dei dati presenti estremamente bassa, quindi non è in grado di supportare un sistema Grid in cui le risorse possono comparire e scomparire molto in fretta. Si consideri, a titolo di esempio, il caso di un nodo di computazione del sistema SETI@home. Esso è, di fatto, un PC che, essendo rimasto inattivo per alcuni minuti, ha fatto partire lo *screen saver*. Questo fatto, però, non fornisce alcuna garanzia sul prolungarsi della disponibilità della risorsa, al contrario è altrettanto probabile che il proprietario, disturbato dalla partenza dello *screen saver*, smetta di sognare ad occhi aperti e ricominci a utilizzare il suo PC. Per queste ragioni, le tecniche di scoperta delle risorse che oggi vengono utilizzate per aggregare le Grid di supercalcolo non si prestano a essere applicate a tipologie diverse di Grid e limitano, quindi, l'utilizzo di tale tecnologia a poche situazioni particolari.
Come dimostrato dal recente attacco [16] orchestrato ai danni del maggior sistema Grid degli Stati Uniti d'America, TeraGRID [17], il problema della sicurezza rappresenta uno dei maggiori punti di vulnerabilità di un sistema massicciamente distribuito e sottoposto a domini e regolamentazioni eterogenee come un sistema Grid. Ancora una volta, una soluzione proviene dal progetto Globus. Tale progetto ha, infatti, definito una struttura di controllo degli accessi alla Grid basata su *Public Key Infrastructure* (PKI) [32] e un sistema di certificati standard unico per l'intera Grid. Tuttavia, tale sistema si basa sulla reciproca fiducia tra le componenti costitutive della Grid. Questo fatto ha in sè un fattore di rischio e una limitazione di uso. Il fattore di rischio proviene dal fatto che, come dimostrato dall'attacco sopra citato, un sistema che assuma la affidabilità di ogni sua componente è sicuro solo come la più debole delle sue componenti. Quindi, nel caso di un sistema massicciamente distribuito tra diversi domini di gestione è praticamente impossibile garantire un effettivo livello di sicurezza. La limitazione proviene dal fatto che l'assunzione di fiducia completa di ogni componente una Grid in ogni altra componente taglia fuori qualsiasi meccansimo volontaristico di acquisizione di risorse e, quindi, limita grandemente i possibili usi della tecnologia. In particolare, ne limita la convergenza con il fenomeno del *peer-to-peer* con il quale, al contrario, condivide il concetto fondamentale di condivisione dinamica di risorse geograficamente distribuite.

7. CONCLUSIONI

In questo articolo, si è voluto rendere disponibile al lettore una visione prospettica del fenomeno Grid fornendogli una tassonomia di questo fenomeno, una descrizione di come la sua evoluzione sia profondamente legata all'evoluzione tumultuosa delle tecnologie informatiche e infine, una descrizione dei problemi irrisolti.
Per quanto il termine Grid sia oggi estremamente di moda, sono proprio problemi quali la gestione dell'elevatissimo livello di eterogeneità e la sicurezza che, restando irrisolti, rendono ancora improponibile un effettivo avvento del Grid in qualità di tecnologia pervasiva. Ad oggi, l'utilizzo della tecnologia Grid è limitato a grandi attori in grado di investire pesantemente in mano d'opera altamente specializzata, l'unica in grado di gestire e utilizzare i sistemi Grid di presente generazione.

Bibliografia

[1] AA.VV.: Enter the Grid website, http://enterthegrid.com/vmp/articles/contentsEnterTheGridResearch%20RepositoryResearch%20Projects.html

[2] AA.VV.: IBM Grid Computing, http://www-1.ibm.com/grid/

[3] AA.VV.: Sun Solutions: Grid Computing, http://www.sun.com/solutions/infrastructure/grid/

[4] AA.VV.: HP's grid solution - the grid stack, http://www.hp.com/techservers/grid/hp_grid_solution.html

[5] AA.VV.: Get on the Grid with Oracle 10g, http://www.oracle.com/events/grid/index.html?content.html

[6] AA.VV.: D-Grid, Iniziativa del governo tedesco per il Grid Computing, http://www.d-grid.de/

[7] AA.VV.: DataGRID Project Website, http://eu-datagrid.web.cern.ch/eu-datagrid/

[8] AA.VV.: The Globus Data Grid Initiative, disponibile on-line http://www.globus.org/datagrid/

[9] AA.VV.: sito progetto ISILab, http://isilab-esng.dibe.unige.it/

[10] AA.VV.: Has the military realized the power of grid computing?, disponibile *on-line* http://www.grid-today.com/02/1021/100580.html

[11] AA.VV.: Documentazione Ufficiale disponibile presso il sito del ministero della difesa Americano, https://www.dmso.mil/public/transition/hla/

[12] AA.VV.: Entropia website, www.entropia.com

[13] AA.VV.: AVAKI website, www.avaki.com

[14] AA.VV.: sito del progetto OGSA, http://www.globus.org/ogsa/

[15] AA.VV.: sito del progetto flashmobcomputing, http://www.flashmobcomputing.org/

[16] AA.VV.: Hackers hit supercomputing giants, disponibile on-line su http://www.cnn.com/2004/TECH/internet/04/15/hackers.supercomputers.ap/

[17] AA.VV.: Sito ufficiale del progetto TeraGrid, http://www.teragrid.org/

[18] Bagnasco A., Scapolla A.M.: *A grid of remote laboratory for teaching electronics*. Proc. of the 2nd International LeGE-WG Workshop, 3-4 Marzo, Parigi, Francia.

[19] C.A.R. Hoare: *Communicating Sequential Processes*. Prentice-Hall International, 1985.

[20] Flynn M.: Some Computer Organizations and Their Effectiveness. *IEEE Trans. Comput*., Vol. C-21, 1972, p. 94.

[21] Foster I., Kesselmann C., Tuecke S.: *The Anatomy of the Grid, Enabling Scalable Virtual Organizations*. International J. Supercomputer Applications, 2001.

[22] Foster I., Kesselman C., Globus: A metacomputing intrastructure toolkit. *International Journal of Supercomputer Applications*, Vol. 11, n. 2, 1997, p. 115-128.

[23] Foster I., Kesselman C.: The Globus Project: A Status Report, Proceedings of the Seventh Heterogeneous Computing Workshop. *IEEE Computer Society Press*, March 1998, p. 4-19.

[24] Foster I., Kesselman C., (eds.): The Grid: *Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Francisco, 1999.

[25] Gannon D., et al.: Programming the Grid: Distributed Software Components, P2P and Grid Web Services for Scientific Applications. *In Special Issue on Grid Computing, Journal of Cluster Computing*, Vol. 5, n. 3, 2002, p. 325-336. Kluwer Academic Publishers, 2002.

[26] Gelsinger P., Gargini P., Parker G., Yu A.: Microprocessors circa 2000. *IEEE Spectrum*, Ottobre, 1989.

[27] Graham P.: *Anatomy of a Lisp Machine*. AI Expert, December 1988.

[28] Gudgin M., Hadley M., Mendelsohn N., Moreau J., Frystyk Nielsen H.: *Simple Object Access Protocol* (1.2). Documentation available on line at http://www.w3.org/TR/SOAP/

[29] Kahn J.M., Katz R.H., Pister K.S.J.: *Mobile networking for smart dust*. Proc. Of MobiCom 99, Seattle, WA, USA, August 1999.

[30] Kogge P.M.: *The Architecture of Symbolic Computers*. McGraw-Hill 1991. ISBN 0-07-035596-7.

[31] Message Passing Interface Forum, MPI 2.0, disponibile on-line a http://www.mpi-forum.org/docs/docs.html

[32] Micali S.: *Enhanced Certificate Revocation System*. MIT/LCS/TM-542, 1995.

[33] Migliardi M., Kurzyniec D., Sunderam V.: *Standard Based Heterogeneous Metacompouting*. The Design of HARNESS II, Proc. of the Heterogeneous Computing Workshop of the International Parallel Distributed Processing Symposium 2002, Fort Lauderdale (FL) April 15-19, 2002.

[34] Moore G.: *Cramming more components onto integrated circuits*. Disponibile on-line, Electronics, Aprile, 1965.

[35] Ramachandramurthi S.: *A beginner's guide to the connection machine CM5, libro elettronico*. Disponibile on-line a http://csep1.phy.ornl.gov/cm5_guide/cm5_guide.html

[36] Saracco R.: Ubiquitous Computing. *Mondo Digitale*, Anno II, n. 7, Settembre 2003, p. 19-30.

[37] Shakkottai S., Srikant R., Shroff N.: *Unreliable Sensor Grids*. Coverage, Connectivity and Diameter, Proc. of INFOCOM 2003, S. Francisco, CA, USA, April 2003.

[38] Sullivan III T., Werthimer D., Bowyer S., Cobb J., Gedye D., Anderson D.: *A new major SETI project based on Project Serendip, data and 100,000 personal computers, Proc. of the Fifth Intl. Conf. on Bioastronomy*. IAU Colloq. n. 161, eds. C.B. Cosmovici, S. Bowyer, and D. Werthimer.

[39] Sunderam V.S., PVM: A Framework for Parallel Distributed Computing, Concurrency, Practice and Experience, 1990.

MAURO MIGLIARDI è nato a Genova nel 1966. Dopo esser stato uno dei principali ricercatori del progetto HARNESS per il meta-computing presso la Emory University di Atlanta, è ora ricercatore universitario presso l'Università di Genova. I suoi principali interessi di ricerca sono il meta-computing e il Grid-computing.
om@dist.unige.it

DENTRO LA SCATOLA

Rubrica a cura di

Fabio A. Schreiber

Il Consiglio Scientifico della rivista ha pensato di attuare un'iniziativa culturalmente utile presentando in ogni numero di Mondo Digitale un argomento fondante per l'Informatica e le sue applicazioni; in tal modo, anche il lettore curioso, ma frettoloso, potrà rendersi conto di che cosa sta "dentro la scatola". È infatti diffusa la sensazione che lo sviluppo formidabile assunto dal settore e di conseguenza il grande numero di persone di diverse estrazioni culturali che - a vario titolo - si occupano dei calcolatori elettronici e del loro mondo, abbiano nascosto dietro una cortina di nebbia i concetti basilari che lo hanno reso possibile. La realizzazione degli articoli è affidata ad autori che uniscono una grande autorevolezza scientifica e professionale a una notevole capacità divulgativa.

Le operazioni aritmetiche

Luigi Ciminiera

1. Introduzione

Questo è il secondo di una serie di articoli, di cui il primo è apparso nello scorso numero [1], che sono dedicati ad alcuni aspetti fondamentali di un sistema di elaborazione, visto come calcolatore, ovvero come macchina per effettuare calcoli. L'articolo precedente si è soffermato sulle rappresentazioni binarie, concentrandosi soprattutto sul modo in cui si rappresentano i numeri. In questo secondo articolo della serie ci si occuperà di descrivere gli algoritmi più comuni per l'effettuazione delle operazioni di somma/sottrazione, moltiplicazione e divisione.

L'algoritmo relativo a una qualsiasi delle operazioni aritmetiche non può prescindere dalla rappresentazione nell'articolo precedente, ovvero:

- **Numeri binari senza segno**, espressi utilizzando *n* bit come:

$$X = \sum_{i=0}^{n-1} x_i 2^i$$

- **Numeri in modulo e segno** che, ricordando che il bit x_{n-1} è quello che esprime il segno negativo (1) o positivo (0), sono espressi come:

$$X = (-1)^{x_{n-1}} \sum_{i=0}^{n-2} x_i 2^i$$

- **Numeri in complemento a 2**, espressi come:

$$X = - x_{n-1} 2^{n-1} + \sum_{i=0}^{n-2} x_i 2^i$$

Questo articolo si limita, inoltre, alla trattazione delle operazioni per i numeri a virgola fissa, rimandando alla letteratura citata in bibliografia, per le operazioni con i numeri a virgola mobile, molto diverse da quelle qui discusse, soprattutto a causa del numero molto maggiore di gradi di libertà.

2. Somma e sottrazione

L'operazione di somma con i numeri interi positivi deve essere effettuata in binario nello stesso modo in cui viene effettuata in decimale: si parte dalla colonna a destra e si calcola, per ogni colonna, la somma delle cifre corrispondenti degli addendi, più un eventuale riporto generato dalla precedente colonna. Per realizzare questo semplice algoritmo, è necessario però conoscere il valore della cifra di somma e quello del riporto che vengono generate da tutte le possibili combinazioni delle cifre nella base prescelta, che per la base 2, sono i seguenti:

0 + 0 + 0 = 0 *con riporto di* 0
1 + 0 + 0 = 0 + 1 + 0 = 0 + 0 + 1 = 1 *con riporto di* 0
1 + 1 + 0 = 1 + 0 + 1 = 0 + 1 + 1 = *o con riporto di* 1
1 + 1 + 1 = 1 *con riporto di* 1

Gli esempi seguenti mostrano come viene effettuata la somma dei numeri decimali 3 + 7 e 10 + 8, utilizzando una rappresentazione su 4 bit.

riporti	1110
3_{10}	0011
7_{10}	0111
10_{10}	1010

riporti	10000
10_{10}	1010
8_{10}	1000
2_{10}	0010

Nell'esempio di destra, si può notare come la somma produca un riporto generato nella colonna dei bit più significativi, che deborda a sinistra dalla nostra rappresentazione di 4 bit. Questo riporto in uscita a sinistra è l'indicazione che si è prodotto un *overflow*, ovvero il risultato è talmente grande che non può più essere rappresentato con il numero di bit prescelto; d'altra parte, è facile verificare che 4 bit consentono, in questa rappresentazione, di operare con numeri interi da 0 a 15, mentre il risultato corretto è 18, che richiederebbe almeno 5 bit.
Per poter essere sicuri di non incorrere in overflow, bisogna garantire che la rappresentazione della somma abbia almeno 1 bit in più rispetto a quella dell'operando più lungo. Nel caso di somma di k numeri da n bit, la lunghezza minima della rappresentazione del risultato, che garantisca contro gli overflow è $n + \lceil \log_2 k \rceil$.
Per quanto riguarda la sottrazione, le regole sono leggermente differenti: si procede sempre da destra verso sinistra, a ogni colonna bisogna sottrarre al bit del primo operando quello del secondo operando e anche un eventuale prestito utilizzato nella colonna precedente.

0 – 0 – 0 = 1 – 1 – 0 = 1 – 0 – 1 = 0 *con prestito di* 0
0 – 1 – 0 = 0 – 0 – 1 = 1 – 1 – 1 = 1 *con prestito di* 1
1 – 0 – 0 = 1 *con prestito di* 0
0 – 1 – 1 = 0 *con prestito di* 1

Gli esempi che seguono mostrano come vengono effettuate le sottrazioni 9 – 7 e 5 – 7, utilizzando le regole precedenti.

prestiti	01000
9_{10}	1001
5_{10}	0101
4_{10}	0100

prestiti	11100
5_{10}	0101
7_{10}	0111
14_{10}	1110

Anche in questo caso, l'esempio di destra mostra un'anomalia segnalata dal fatto che vi è un prestito in uscita dalla colonna più a sinistra. In questo caso, il problema consiste nel fatto che il risultato è negativo (–2), mentre si sta utilizzando una rappresentazione che considera solo, per definizione, numeri interi *positivi*. Per poter trattare anche i numeri interi negativi, bisogna ricorrere a una delle altre due rappresentazioni considerate. In particolare, la rappresentazione in complemento a 2 permette di effettuare le somme utilizzando esattamente lo stesso algoritmo precedentemente illustrato per i numeri senza segno, con l'ulteriore semplificazione che anche i bit di segno vengono trattati nello stesso modo rispetto a tutti gli altri bit. Si può, quindi, procedere come mostrano gli esempi riportati di seguito:

riporti	11110
$+3_{10}$	0011
-1_{10}	1111
$+2_{10}$	0010

riporti	10000
-6_{10}	1010
-8_{10}	1000
$+2_{10}$	0010

Come sempre, l'esempio a destra mostra un'anomalia, che in questo caso non è dovuta al riporto uscente dalla posizione più a sinistra (quella dei bit di segno): infatti, l'esempio di sinistra mostra anch'esso un riporto uscente a sinistra, ma si ottiene il risultato corretto ignorando semplicemente questo riporto. Il vero problema, messo in rilievo dall'esempio di destra, consiste nel fatto che sommando 2 operandi negativi si è ottenuto un risultato positivo. In complemento a 2, questa è la prova che è avvenuto un overflow; andando a controllare si può verificare che il risultato corretto è –14, mentre la rappresentazione su 4 bit utilizzata permette di manipolare solo i numeri interi fra –8 e +7. Analogamente, si ha overflow anche se sommando 2 numeri positivi si ottiene un risultato negativo. Si noti che, nelle rappresentazioni di numeri con segno, non è necessario trattare il caso della sottrazione, dal momento che questa viene effettuata sommando l'opposto del secondo operando, anziché procedere con la sottrazione diretta.
Una possibile alternativa al complemento a 2 è data dalla rappresentazione in modulo e segno. Bisogna, innanzitutto, riconoscere che questo tipo di rappresentazione è esattamente la stessa (a parte il cambio di base) che utilizziamo noi umani per la base 10. In binario, tutte le cifre hanno 2 valori come pure il segno (+ o –), per cui è venuto naturale rappresentare gli stessi segni mediante dei bit (0 per +, 1 per –);

invece, nel caso dei numeri decimali, le cifre hanno 10 valori e il segno solo 2 per cui sono stati adottati simboli diversi.
Anche le regole per la somma sono quelle che si adottano normalmente nei calcoli manuali: se i segni degli operandi sono concordi, bisogna sommare i moduli e il segno del risultato è quello degli operandi, se i segni sono discordi, bisogna sottrarre il modulo inferiore da quello superiore e il segno è determinato dall'addendo con modulo superiore.
L'automazione di queste regole fa sorgere, però, un problema: determinare l'operando con modulo maggiore comporta già di per sé una sottrazione. Per semplificare l'algoritmo nel caso di segni discordi, si adottano le seguenti regole: si sottrae il modulo del secondo operando dal primo; se il prestito uscente a sinistra è 0, il primo operando aveva effettivamente modulo maggiore, altrimenti è il contrario, e il modulo corretto del risultato viene ottenuto complementando a 2 il risultato ottenuto (quest'ultima operazione si effettua invertendo tutti i bit e sommando al risultato 1 nella posizione più a destra).
Situazioni di overflow possono presentarsi solo nel caso di somma dei moduli (operandi con segni concordi) e vengono rivelate dal fatto che questa somma produce un riporto uscente a sinistra. Per quanto riguarda la lunghezza minima della rappresentazione del risultato che garantisce contro possibili overflow, si applicano le stesse formule valide per i numeri senza segno.

3. Moltiplicazione

Anche per la moltiplicazione è possibile partire da una rivisitazione dell'algoritmo adottato nelle operazioni manuali con numeri decimali, iniziando a considerare il caso dei numeri senza segno. Il meccanismo di moltiplicazione consiste nel partire dalla cifra meno significativa del moltiplicatore, moltiplicare tutto il moltiplicando per quella cifra, ottenendo un primo risultato parziale; si passa a moltiplicare tutto il moltiplicando per la successiva cifra del moltiplicatore, il risultato viene sommato a quello parziale spostandolo di una posizione a sinistra. Si continua così a moltiplicare ogni cifra del moltiplicatore per il moltiplicando, e a sommare il risultato spostandosi ogni volta di una posizione a sinistra.
Per poter effettuare la moltiplicazione di una cifra del moltiplicatore per il moltiplicando, è necessario servirsi della famosa *tavola pitagorica*, che dice qual è il risultato della moltiplicazione per ciascuna delle possibili coppie di cifre della rappresentazione. Nel caso della base 2, la tavola pitagorica si riduce alla seguente forma:

	0	1
0	0	0
1	0	1

L'overflow è possibile anche nella moltiplicazione. Per il caso dei numeri senza segno, la lunghezza minima del prodotto, che garantisce di poter rappresentare qualsiasi risultato, è di $m + n$ bit, dove m e n sono le lunghezze in bit delle rappresentazioni dei due fattori.
La moltiplicazione di numeri in modulo e segno è sostanzialmente identica a quella dei numeri senza segno, poiché il modulo del prodotto è il prodotto dei moduli (che sono numeri positivi senza segno), mentre il segno del risultato viene ottenuto dal semplice EXOR dei due bit di segno dei fattori. L'unica variazione da notare è che la moltiplicazione di un fattore lungo m bit per uno di lunghezza n bit, produce un prodotto che occupa al massimo $m + n - 1$ bit.
Per i numeri in complemento a 2, bisogna rifarsi alla formula data all'inizio di questo articolo. La moltiplicazione può avvenire come quella dei numeri senza segno, tenendo però presente la seguente particolarità della rappresentazione. Il bit più significativo del moltiplicatore ha un peso negativo (-2^{n-1}), per cui, se questo bit è a 1, bisogna sottrarre il moltiplicando (o sommare il suo complemento a 2), anziché sommarlo.
Nel caso di fattori di lunghezza m e n bit rispettivamente, il prodotto occuperà $n + m$ bit al massimo. È bene notare che non si può usare 1 bit in meno, per via di un'unica combinazione nei valori dei fattori che richiede il massimo di bit; questo unico caso corrisponde alla moltiplicazione $(-2^{n-1}) \times (-2^{m-1}) = +2^{m+n-2}$, che richiede, appunto, una rappresentazione di almeno $n + m$ bit.
L'esempio seguente mostra la moltiplicazione $(-5_{10}) \times (-3_{10})$, con fattori su 4 bit, che in complemento a 2 sono rappresentati come $X = 1011$ (moltiplicando) e $Y = 1101$ (moltiplicatore). Poiché il moltiplicando è negativo, bisognerà som-

mare una serie di addendi negativi con una rappresentazione su 8 bit, che è la massima necessaria per il prodotto dei due operandi da 4 bit. In complemento a 2, l'allungamento della rappresentazione si effettua ripetendo a sinistra il segno del numero (cioè il bit più a sinistra), per questo motivo i vari addendi negativi appaiono nell'esempio seguente in una rappresentazione allungata rispetto a quella originale del moltiplicando.

	1011 x
	1101
	11111011
	0000000
	111011
Sottrazione del moltiplicando W	00101
Risultato= 15_{10}	00001111

4. Divisione

Contrariamente alle altre operazioni aritmetiche viste in precedenza, non è sempre possibile calcolare esattamente il quoziente di una divisione. Il motivo risiede nel fatto che la divisione, anche se ci si limita a operandi interi, non produce necessariamente un risultato che ha una rappresentazione con numero finito di cifre.
In questo contesto, verrà esaminata la divisione intera, in quanto l'algoritmo per il caso generale è lo stesso, ma varia solo il momento in cui si decide di interrompere il calcolo dei bit del quoziente. Nella divisione intera, il dividendo *X*, il divisore *D*, il quoziente *Q* e il resto *W* sono legati dalla seguente relazione, dove tutte le quantità sono intere:

$$X = QD + W \qquad 0 \le |W| < |D|$$

Un ulteriore vincolo è rappresentato dal fatto che il resto abbia lo stesso segno del dividendo. Anche in questo caso, l'algoritmo utilizzato potrebbe essere quello ben noto della base 10, con gli opportuni adattamenti alla base 2. Questo algoritmo viene detto di tipo *restoring*, in quanto bisogna ripristinare il valore del resto parziale, qualora questo assuma valore negativo a seguito di una delle sottrazioni che vengono effettuate.
La velocità di esecuzione può essere migliorata se si permette di ottenere un risultato negativo dopo qualsiasi sottrazione; ovviamente, al passo successivo bisognerà tenere conto di questo fatto. Il nuovo algoritmo che si ottiene è detto di tipo *non-restoring*, in quanto non si disfa mai l'operazione di sottrazione/somma appena fatta. La divisione non-restoring per numeri positivi può essere descritta nel modo seguente, dove W_i è l'*i*-esimo valore assunto dal resto parziale e q_i è l'*i*-esimo bit del quoziente; si noti, inoltre, che vale l'ipotesi che il dividendo sia rappresentato su 2*n* bit e il divisore su *n* bit.

$W_0 = X$
$W_1 = 2\,W_0 - D\,2^n$
for $j = 1 \ldots . n-1$
  if $W_j \ge 0$ **then** $q_{n-j} = 1$; $W_{j+1} = 2\,W_j - D\,2^n$;
    else $q_{n-j} = 0$; $W_{j+1} = 2\,W_j + D\,2^n$;
endfor
if $W_n \ge 0$ **then** $q_0 = 1$;
  else $q_0 = 0$; $W_n = W_n + D\,2^n$;
    [correzione del segno del resto]

Nell'algoritmo sopra descritto il resto parziale viene scalato di una posizione a sinistra (moltiplicazione per 2) a ogni passo, anziché far scalare di una posizione a destra la quantità da sottrarre, come avviene quando la divisione viene effettuata manualmente.
Il numero di bit interi del quoziente può essere derivato dalla sottrazione del numero dei bit della rappresentazione del dividendo meno quelli della rappresentazione del divisore, come avviene anche in decimale; il numero ottenuto rappresenta un valore massimo. Nell'esempio seguente, viene mostrato il funzionamento dell'algoritmo non-restoring per la divisione di 11_{10} = 00001011_2 (dividendo) per 2_{10} = 0010_2 (divisore); si noti che viene effettuata una divisione di un numero da 8 bit per un altro da 4 bit, di conseguenza la parte intera del quoziente non richiederà più di 4 bit.
Nella colonna più a sinistra viene anche mostrato il valore del riporto uscente a sinistra, che indica il segno del risultato delle varie operazioni, si noti anche che le sottrazioni vengono effettuate sommando il complemento a 2 del divisore.

$W_0 =$ 0 0000 1011

$2\,W_0 =$ 0 0001 0110

$-D\,2^4 =$ 1 1110

$W_1 =$ 1 1111 0110 $q_3 = 0$

$2\,W_1 =$ 1 1110 1100

$+D\,2^4 =$ 0 0010

$W_2 =$ 0 0000 1100 $q_2 = 1$

$2\,W_2 =$ 0 0001 1000

$-D\,2^4 =$ 1 1110

$W_3 =$ 1 1111 1000 $q_1 = 0$

$2\,W_3 =$ 1 1111 0000

$+D\,2^4 =$ 0 0010

$W_4 =$ 0 0001 0000 $q_0 = 1$

resto = 0 0001

Il quoziente è dato da $0101_2 = 5_{10}$ mentre il resto finisce nella parte superiore di *W* ed è pari a 1.

Bibliografia

[1] Dadda L. : Fondamenti dell'aritmetica digitale: i codici numerici. *Mondo Digitale*, Anno III, n. 9, marzo 2004, p. 61-65.

[2] Ercegovac M.D., Lang T.: *Digital Arithmetic*. Morgan Kaufmann, 2004.

[3] Koren I.: *Computer Arithmetic Algorithms*. Prentice-Hall, 1993.

Luigi Ciminiera si è laureato in Ingegneria Elettronica nel 1977 presso il Politecnico di Torino. Dal 1979 ha iniziato a collaborare alle attività del Dipartimento di Automatica e Informatica, presso il quale ha ricoperto varie posizioni. Attualmente è professore ordinario di Sistemi di Elaborazione dell'Informazione. Dal 1989, è membro del Comitato di Programma di IEEE Symposium on Computer Arithmetic, di cui è anche stato Program Co-Chairman nell'edizione del 2001. Oltre all'aritmetica per elaboratori, i suoi interessi scientifici comprendono anche i sistemi peer-to-peer e le griglie computazionali. È stato co-autore di 2 libri a diffusione internazionale e di otre 100 contributi pubblicati in riviste e conferenze scientifiche. Egli ricopre attualmente la carica di Preside della II Facoltà di Ingeneria del Politecnico di Torino.
luigi.ciminiera@polito.it

ERRATA CORRIGE

Sul numero 7 di marzo 2003 di Mondo Digitale, nell'articolo "Aspettando Robot" sono state pubblicate, a pagina 15, due immagini con didascalie errate. Riportiamo, qui sotto le due didascalie corrette e complete di fonte.

Figura 9 (sopra)
La cella robotica sottomarina realizzata dal CNR-IAN-Reparto Robotica, Unige-Dist e Ansaldo, nell'ambito del Progetto Europeo 1996/1998 EC MAS3 CT950024 AMADEUS (Advanced Manipulation for DEep Underwater Sampling) (foto G. Veruggio).

Figura 9 (sotto)
Il robot sottomarino Romeo realizzato dal Reparto Robotica del CNR-IAN, durante le prove finali del Progetto Europeo 1997/2000 EC MAS3 CT970083 ARAMIS (Advanced Rov package for Automatic Mobile Investigation of Sediments) (foto G. Veruggio).

ICT E DIRITTO

Rubrica a cura di

Antonio Piva e David D'Agostini

Scopo di questa rubrica è di illustrare al lettore, in brevi articoli, le tematiche giuridiche più significative del settore ICT: dalla tutela del *domain name* al *copyright* nella rete, dalle licenze software alla *privacy* nell'era digitale. Ogni numero tratterà un argomento, inquadrandolo nel contesto normativo e focalizzandone gli aspetti di informatica giuridica.

L'accesso abusivo ai sistemi informatici e telematici:
Aspetti giuridici e informatici di un attacco hacker

1. Criminalità informatica

Il progresso dell'ICT, oltre a portare indiscutibili benefici in ogni settore della società, ha sollevato alcune problematiche legate alla sicurezza dei sistemi di comunicazione e alla loro integrazione; in un contesto caratterizzato da un intrinseco elevato tasso di vulnerabilità, ha avuto terreno fertile lo sviluppo dei così detti *computer crimes*. Con questa espressione si vuole indicare una categoria di illeciti penali, in cui l'elaboratore può essere, da un lato, lo strumento per la consumazione del reato (si pensi alle frodi informatiche), dall'altro l'oggetto sul quale il reato stesso viene commesso (per esempio, nel danneggiamento informatico).
L'esigenza di disciplinare in maniera sistematica il settore della criminalità informatica, ha portato all'approvazione della legge 23 dicembre 1993 n. 547, con la quale è stato modificato il Codice penale italiano mediante l'aggiornamento di alcuni reati già esistenti e l'inserimento di altri completamente nuovi. Tra questi ultimi, l'art. 615 *ter* c.p. rubricato "*Accesso abusivo a un sistema informatico o telematico*", prevede che chiunque si introduce abusivamente in un sistema informatico o telematico protetto da misure di sicurezza, ovvero vi si mantiene contro la volontà, espressa o tacita, di chi ha il diritto di escluderlo, sia punito con la reclusione fino a tre anni.

2. L'accesso abusivo

I fenomeni di accesso abusivo costituiscono sul piano criminologico una categoria piuttosto eterogenea che, secondo la finalità perseguita dall'intruso, possono essere classificate in ipotesi meramente ludiche (hackeraggio) o in ipotesi che precedono la consumazione di altri reati dove all'accesso segua la distruzione di dati o il compimento di frodi, spionaggio industriale ecc..
L'intrusione informatica può concretizzarsi tramite l'accesso fisico al locale in cui è ubicato l'elaboratore, ovvero può essere compiuta in modo autonomo, mediante accesso da remoto, costituendo, in ogni caso, un fatto separato e distinto rispetto all'accesso fisico.
Le statistiche evidenziano come l'accesso abusivo sia compiuto in misura maggiore da personale interno all'organizzazione piuttosto che da estranei, dato estremamente significativo di cui tener conto sia nella progettazione e realizzazione di una rete aziendale che nella stesura di un'adeguata *policy* di sicurezza e di riservatezza dei dati.
Dal punto di vista giuridico, l'art. 615 *ter* (si veda il Riquadro 1), punisce alternativamente due condotte diverse:
a. l'introduzione abusiva in un sistema informatico o telematico protetto da misure di si-

Riquadro 1
Accesso abusivo a un sistema informatico o telematico: *reclusione fino a 3 anni*.
Se il fatto è commesso con abuso della qualità di operatore di sistema: *reclusione da 1 a 5 anni*.
Se il colpevole usa violenza su cose o persone: *reclusione da 1 a 5 anni*.
Se ne deriva il danneggiamento del sistema, dei dati o dei programmi: *reclusione da 1 a 5 anni*.

curezza, vale a dire l'operazione con la quale un soggetto, connettendosi a un sistema di elaborazione, acquisisce la possibilità di prendere cognizione di dati, informazioni e programmi, alterarli in tutto o in parte, cancellarli, aggiungerne dei nuovi; l'accesso abusivo, ossia fraudolento, si prefigura anche nel caso di formale proibizione o di semplice mancanza di autorizzazione;

b. il mantenimento nel medesimo contro la volontà, espressa o tacita, di chi ha il diritto di escluderlo. Infatti, un soggetto pur essendo autorizzato a entrare in un sistema, potrebbe mantenervisi indebitamente compiendo operazioni non autorizzate; ciò può accadere nei luoghi di lavoro nel caso di utilizzo temporaneo della postazione di un collega.

3. Le misure di sicurezza

In via del tutto generale, per poter punire il colpevole ai sensi dell'art. 615 *ter*, appare necessario che il sistema informatico o telematico attaccato sia protetto da misure di sicurezza, ove, in tale termine, sono ricomprese le misure tecniche, informatiche, organizzative, logistiche e procedurali che si frappongono al libero utilizzo di un sistema da parte delle persone non autorizzate come, per esempio, meccanismi di *autentificazione* degli utenti tramite opportune *password* o dispositivi biometrici, profili di autorizzazione, programmi *software* dedicati, *firewall* o altri apparati *hardware*, *server* chiusi a chiave ecc. (misure previste anche nell'allegato B, contente il disciplinare tecnico in materia di misure minime di sicurezza, del "Codice in materia di protezione dei dati personali", decreto legislativo 30 giugno 2003 n.196).

Se in una LAN è attiva la condivisione di tutti i dischi rigidi, non può essere condannato per accesso abusivo il dipendente infedele che vada a esplorare le risorse dei colleghi; il discorso si fa ancora più significativo nel caso di connessione da remoto allorché nell'elaboratore (in genere, un *server*) siano attivi processi e servizi, liberamente accessibili dall'esterno, tramite connessione a determinate porte, o siano configurati in maniera da consentire determinate operazioni senza richiedere particolari *permission* (basti pensare al Netbios sulla 138 e 139 ovvero all'FTP sulla 21, si veda il Riquadro 2).

Riquadro 2

Misure di sicurezza: un hacker si è introdotto nel sito telematico del GR1, sostituendo il *file* contenente l'edizione del giornale radio con un proprio file. Le indagini chiarirono che l'imputato aveva sfruttato una caratteristica tipica del sistema operativo Windows 95 sul quale risulta di *default* attiva la condivisione file e stampanti sul protocollo Netbios, senza richiesta di alcuna password; il giudice, pertanto, lo ha assolto.

Le misure di sicurezza predisposte devono, inoltre, essere effettive: non sarebbe sufficiente, per esempio, la semplice richiesta di *user name* e *password* fatta all'avvio dal sistema operativo se tali campi sono vuoti e, quindi, per accedere si rivela sufficiente la pressione del tasto "invio" (o, peggio, un click su "annulla", si pensi al Windows 95/98).

Risulta, quindi, fondamentale che i responsabili della sicurezza dei sistemi prendano cognizione delle tecniche utilizzate dagli hacker (sintetizzate nei successivi paragrafi) per porre in essere un attacco telematico e/o per effettuare un accesso abusivo, in modo da rendere effettive le contromisure di ordine tecnico e organizzativo, minimizzando i rischi di introduzione non autorizzata ai sistemi.

È d'obbligo, comunque, segnalare che la Corte di Cassazione ha ritenuto che la condizione determinante per la configurazione del reato, sia non tanto la presenza di misure di protezione interne o esterne al sistema, quanto l'aver agito contro la volontà di chi dispone legittimamente del sistema. Infatti, un sistema dovrebbe essere giuridicamente tutelato sempre e non solamente quando il titolare lo abbia dotato di *firewall* o di IDS (*Intrusion Detection System*).

4. Il footprinting

Il primo passo di un attacco telematico, posto in essere in maniera metodica, consiste nella ricerca del maggior numero possibile di informazioni sul potenziale obiettivo, al fine di determinarne il *footprinting*, ossia l'impronta, e quindi di ricostruire il suo profilo informatico e, in ultima analisi, il suo livello di protezione.

Tale fase ricognitiva viene condotta sistematicamente per assicurare la raccolta di tutti gli elementi utili, vale a dire: nome di dominio, blocchi della rete, indirizzi IP, tipo di connessio-

ne, servizi TCP e UDP in esecuzione su ciascuno dei sistemi identificati, architettura di sistema, meccanismi di controllo degli accessi, sistemi di intercettazione delle intrusioni, meccanismi di autenticazione.
Come punto di partenza usualmente viene sfruttato il sito Internet della vittima sul quale possono essere rinvenibili informazioni (identità del soggetto, sede fisica, numeri telefonici, indirizzi e-mail, e in alcuni casi si deduce la *privacy policy* ecc.); mediante un'interrogazione (cosiddetti, *whoi*) alla banca dati della *Registration Autority* si ottiene il nominativo della persona fisica o giuridica a cui il nome di dominio è registrato, nonché del *provider/manteiner* e ulteriori dati tecnici quali i *server* DNS primari e secondari, indirizzi IP della rete ed eventuali sottoreti.
Nel codice sorgente HTML delle pagine web possono essere contenuti informazioni utili ed eventuali commenti non visibili in quanto nascosti nei cosidetti *metatag*, pertanto non è infrequente la pratica di scaricare un intero sito per analizzarlo accuratamente *offline*, anche allo scopo di comprendere la struttura del sito. È possibile, inoltre, utilizzare i motori di ricerca, consultare *newsgroup* e database di file da cui ricavare notizie e indizi circa lo stato della struttura e il suo livello di protezione.
Una volta identificate le potenziali reti, viene cercato di stabilirne la topologia e individuare possibili percorsi di accesso; validissimo strumento è il programma *Neotrace* che, sfruttando la proprietà TTL (*time-to-live*) del pacchetto IP, permette di risalire al suo percorso nel passaggio da un *host* al successivo, riuscendo inoltre a identificare eventuali dispositivi di controllo degli accessi (*firewall, router* con filtri sui pacchetti, *switch*), il tutto in modalità grafica che integra tali informazioni con le interrogazioni *whoi*.
Per quanto la maggior parte delle informazioni ricavabili con questi metodi debba in forza di legge o per effettive esigenze pratiche essere resa pubblica, tuttavia è importante che il titolare effettui un'attenta e consapevole valutazione e classificazione del tipo di informazioni distribuite.

5. La scansione

Se il *footprinting* è una ricognizione per la raccolta di informazioni, con il passo successivo, la scansione, vengono stabiliti quali sistemi siano attivi e raggiungibili via Internet e, in base a tali informazioni, viene identificato quale tipo di attacco è attuabile, utilizzando una serie di strumenti e tecniche come il *ping sweep*, la scansione delle porte e il tool di ricerca automatica.
Il ping sweep consiste nell'esecuzione automatizzata di una serie di comandi ping su un intervallo di indirizzi IP e blocchi della rete che, oltre a restituire l'elenco dei sistemi in funzione, ove opportunamente impostato, può risolvere il nome degli host.
Si rivela, pertanto, di fondamentale importanza per la vittima riconoscere questo genere di attività valutando attentamente il tipo di traffico consentito sulla rete o in sistemi specifici.
I metodi principali per l'intercettazione di questo genere di attacchi consistono nell'adozione di programmi di IDS (*Intrusion Detection System*) funzionanti sulla rete; inoltre, sia router che i firewall, possono essere configurati in modo tale da non rispondere a eventuali ping inviati, dando in tal modo l'illusione all'esterno che la rete o le macchine in oggetto non siano collegate in rete.
Mediante interrogazioni ICMP (*Internet Control Messaging Protocol*) è possibile ottenere la *netmask* di una scheda di rete, informazione preziosa perché permette di identificare tutte le sottoreti utilizzate e, di conseguenza, di concentrare gli attacchi su una specifica porzione della rete, evitando per esempio gli indirizzi *broadcast*.
Uno dei rimedi più consigliati consiste nel bloccare sui *router* esterni i messaggi ICMP che forniscono informazioni (quanto meno la richiesta di *address mask*).
Ulteriore passo in avanti viene compiuto con la scansione delle porte, vale a dire la connessione alle porte TCP e UDP presenti nel sistema scelto come obiettivo per stabilire quali servizi siano in esecuzione e quali in stato di *listening*, cioè in ascolto. L'identificazione delle porte è fondamentale per risalire al tipo di sistema operativo alle applicazioni in uso; inoltre, eventuali servizi attivi potrebbero consentire a utenti non autorizzati di accedere a sistemi configurati in maniera non corretta o sui quali sia installato un *software* con vulnerabilità note.
In realtà, esistono diverse modalità di scansione a seconda del tipo di pacchetto inviato e del-

la risposta provocata; alcuni sono facilmente intercettati (TCP *connect scan*), altri garantiscono un funzionamento su tutti gli host (scansioni SYN e *connect*).
Si osserva che si possono facilmente reperire su Internet numerosi programmi che facilitano il compito dell'aspirante intrusore: *Strobe, Netcat, Nmap* per piattaforme Unix; *NetScan, WinScan, SuperScan* per quelle Windows sono i più noti in ragione della loro efficacia.
I principali metodi utilizzati per l'intercettazione di tentativi di scansione delle porte sono l'utilizzo di programmi di *Intrusion Detection System* (*Snort*) e l'attivazione di meccanismi di *reporting* delle anomalie che indichino i tentativi di scansione, configurando il sistema in modo che vengano inviati gli *alert* in tempo reale via e-mail e sms. È, inoltre, possibile ridurre al minimo i rischi disattivando tutti i servizi non strettamente necessari o quantomeno frapponendo un firewall software o hardware tra la macchina e la rete.
Il passo successivo consiste nel rilevamento del sistema operativo mediante il cosiddetto *fingerprinting* dello *stack* (attivo o passivo a seconda che vi sia l'invio di pacchetti al sistema *target* ovvero il mero esame del traffico di rete); tale tecnica, basandosi sulle differenze tra le diverse implementazioni dello *stack* IP, permette appunto di riconoscere con una certa sicurezza il sistema operativo in uso.

6. L'enumerazione

In seguito, mediante l'enumerazione si procede all'identificazione di *account* validi o di condivisioni di risorse poco protette; la differenza principale tra la raccolta di informazioni esaminate finora e l'enumerazione è il livello di invadenza: l'enumerazione richiede connessioni dirette ai sistemi e interrogazioni esplicite e dovrebbe di conseguenza essere "loggata", cioè registrata dalla vittima.
I dati che trapelano da questa attività potrebbero rilevarsi decisivi ai fini dell'intrusione, dal momento che, una volta enumerato un nome utente valido o una condivisione, risalire alla password corrispondente o a un punto debole del protocollo di condivisione è solo questione di tempo.
Le informazioni enumerate sono relative a risorse e condivisioni di rete, utenti e gruppi, applicazioni e *banner*; le tecniche di enumerazione si differenziano a seconda del sistema operativo (per questa ragione è fondamentale conoscere tale dato).
Le tecniche maggiormente diffuse sfruttano l'interfaccia NetBios, i sistemi SNMP (*Simple Network Managment Protocol*), i trasferimenti di zona DNS e altre funzioni integrate nel sistema operativo.
La cattura dei banner, invece, è l'operazione di connessione ad applicazioni remote e l'attenta osservazione dell'*output* da queste prodotto e può rivelarsi particolarmente istruttiva permettendo di identificare il software e la versione in esecuzione sul server, sufficienti a individuare le relative vulnerabilità.
La cattura di *login* e password spesso viene effettuata tramite lo *sniffing* dei pacchetti di trasmissione, tramite appositi software detti *sniffer* e facilmente reperibili su Internet, che si mettono in ascolto sulla rete, catturando tutti i pacchetti dati da e per una particolare macchina o rete, consentendo in pratica di vedere tutti i dati trasmessi e ricevuti, tra cui anche eventuali password di accesso ai sistemi.
Un altro modo per condurre l'enumerazione su applicazioni NT/2000 consiste nel visionare il contenuto del Registro di configurazione di Windows, dopo averlo copiato dal sistema *target*, nel quale vi si trovano dati relativi all'utente e alla configurazione.
In ambiente Unix, a tale scopo viene impiegato il ben noto comando *finger*.

7. L'attacco

Nel corso di un attacco informatico sono spesso utilizzate le tecniche di sniffing e anche di *spoofing*:
In particolare, lo *spoofing*, utilizzando una debolezza intrinseca del protocollo TCP/IP, permette il mascheramento della propria identità, facendo, in tal modo, credere alla vittima di essere un altro computer.
Se non è possibile individuare le password di accesso al sistema, l'alternativa più valida consiste nel cercare i difetti dell'architettura del sistema operativo e delle sue applicazioni.
Il più temuto è il *buffer overflow*: quando le applicazioni non verificano la lunghezza dell'*input*, tendono a verificarsi errori nelle applicazioni dovuti a sovraccarichi di dati che pos-

sono essere spinti nello *stack* di esecuzione della CPU, con la conseguenza che, se i dati immessi vengono scelti opportunamente, l'errore può comportare l'esecuzione di righe di codice. Poiché la programmazione genera inevitabilmente errori, l'unico rimedio efficace consiste nell'aggiornamento continuo dei programmi (come del resto previsto dalla già citata legge sulla *privacy*, il "Codice in materia di protezione dei dati personali"), mediante l'applicazione di *patch* che vanno letteralmente a tappare le falle.
Ulteriore tipologia di attacco è il DoS, ossia *Denial of Service*; in realtà, il rifiuto di servizio non è tecnicamente un accesso, ma sembra opportuno farne menzione, seppur brevemente, in ragione della sua diffusione, della semplicità con cui può essere condotto, nonché del danno che arreca alle imprese, dovuto al mancato guadagno, al tempo e al lavoro necessario per identificare e annullare questo disservizio, nonché per ripristinare il corretto funzionamento del sistema.
Gli attacchi DoS, in pratica, vengono attuati "bombardando" di richieste un particolare servizio attivo nella macchia (per esempio, il servizio Web); il risultato di tale attività di richieste dirette o più spesso amplificate, può portare all'esaurimento della larghezza di banda di una determinata rete e/o, in breve tempo, a un sovraccarico di lavoro e il conseguente esaurimento delle risorse di sistema, che si traduce in un blocco della macchina attaccata.
A tal fine, possono, ancora una volta, essere sfruttati difetti di programmazione che, impedendo a un'applicazione o al sistema operativo di gestire condizioni eccezionali, permettono l'invio di dati non previsti dall'elemento vulnerabile.

8. Il social engeenering

Un argomento sicuramente ben noto non solo a chi si occupa di sicurezza informatica, suscitandone timore, ma anche ai non addetti ai lavori è l'ingegneria sociale.
Tale espressione, affermatasi nel gergo hacker è ormai di uso comune; descrive l'insieme delle tecniche di persuasione e/o di inganno messe in campo per accedere a un sistema informatico; generalmente, si presenta con modalità simili alla conversazione umana, meglio se telefonica o tramite e-mail e consiste nell'ottenere direttamente dall'interlocutore le informazioni desiderate. La casistica è pressoché infinita e può consistere nello spacciarsi per un collega, un tecnico, un cliente ecc., e nel fingere di avere dimenticato la password, di necessitare urgentemente di un certo codice e così via.
Non è necessario che l'informazione sia sempre direttamente impiegabile, ben potendo essere utilizzata in un secondo momento; per esempio, può rivelarsi importante conoscere la data di nascita o il nome della moglie di un utente, posto che tale potrebbe essere la password ingenuamente assegnata dal medesimo.
Proprio dall'intima convinzione che l'anello debole della sicurezza informatica sia rappresentato dal fattore umano, si ritiene indispensabile adottare severe contromisure per prevenire gli attacchi basati sull'ingegneria sociale e in particolare: limitare rigorosamente la fuoriuscita di dati; stabilire una seria politica per le procedure di supporto tecnico; usare estrema cautela nella configurazione dei firewall, anche in uscita; utilizzare la posta elettronica in modo consapevole, e, soprattutto, istruire i dipendenti fornendo loro le nozioni base sulla sicurezza dei sistemi informatici.

9. Conclusioni

L'intrusione abusiva spesso si rivela propedeutica o successiva rispetto alla commissione di altri reati, pertanto può concorrere con la violazione di domicilio di cui agli artt. 614 e 615 c.p. (per esempio, accesso fisico nel centro di calcolo e successiva penetrazione nella banca dati), con la detenzione e la diffusione abusiva di codici di accesso a sistemi informatici e telematici, nonché con la diffusione o comunicazione di programmi informatici diretti a danneggiare o interrompere un sistema informatico ovvero dati o programmi in esso contenuti, delitti previsti e puniti rispettivamente dagli artt. 615 *quater* e *quinquies* c.p. (per esempio, ricerca e detenzione sul pc di password, programmi di *crack, exploit*, nonché di software finalizzati a rilevare e sfruttare le vulnerabilità del sistema *target* e il successivo utilizzo degli stessi nel corso dell'accesso).
L'art. 615 *ter* può, altresì, essere contestata in concorso con la frode informatica (art. 640 ter

c.p.), che comporta necessariamente l'alterazione di un sistema mediante manipolazione di dati, informazioni o programmi contenuti.
Si deve, invece, escludere il concorso con il reato di danneggiamento di sistemi informatici e telematici di cui all'art. 635 *bis* c.p., in quanto tale evento ne costituirebbe solo una circostanza aggravante.
In conclusione, non si può omettere di considerare che, per i motivi sopra illustrati, il fenomeno dell'accesso abusivo a un sistema informatico o telematico debba essere combattuto in primo luogo creando negli operatori un'adeguata e diffusa cultura della sicurezza.
Si ritiene, pertanto, auspicabile un'iniziativa di formazione permanente finalizzata alla sensibilizzazione degli utenti sul presupposto che solamente attraverso un utilizzo consapevole delle nuove tecnologie da parte di tutti si possa consentirne uno sviluppo armonico e un'ampia diffusione nel pieno rispetto della legalità.

Antonio Piva laureato in Scienze dell'Informazione, Presidente, per il Friuli - Venezia Giulia, dell'ALSI (*Associazione Nazionale Laureati in Scienze dell'Informazione ed Informatica*) e direttore responsabile della Rivista di Informatica Giuridica.
Docente a contratto di Informatica giuridica all'Università di Udine.
Consulente sistemi informatici, valutatore di sistemi di qualità ISO9000 e ispettore AICA per ECDL base e advanced.
antonio_piva@libero.it

David D'Agostini avvocato, ha conseguito il master in informatica giuridica e diritto delle nuove tecnologie, fornisce consulenza e assistenza giudiziale e stragiudiziale in materia di *software, privacy* e sicurezza, contratti informatici, *e-commerce*, nomi a dominio, computer crime, firma digitale. Ha rapporti di partnership con società del settore ITC nel Triveneto.
Collabora all'attività di ricerca scientifica dell'Università di Udine e di associazioni culturali.
david.dagostini@adriacom.it