

COGITO[®]

La tecnologia linguistica secondo Expert System

Presentazione

Gli studiosi della lingua inglese segnalano un cambiamento in corso: i gerundi stanno migrando in rete. *Shopping, banking, trading*, tutti, prima o poi, sono diventati "on line" e altri lo diventeranno.

Scherzi a parte, dietro una piccola evoluzione del lessico si nasconde una rivoluzione nella vita di milioni di persone.

In Italia stiamo ancora godendoci quello che Internet ci ha fatto guadagnare, soprattutto in comodità, ma negli Stati Uniti, e in altri paesi che ci hanno preceduto, si lavora già per recuperare quello che si è perso: il rapporto umano.

Il bacino d'utenza Internet del nostro Paese, già relativamente piccolo, cresce più lentamente che altrove e siamo convinti che questo dipenda anche dalla mancanza di un'interfaccia amichevole tra l'utente di tutti i giorni e il *mare magnum* dei siti Web.

Oggi una soluzione esiste, si chiama NLP (Natural Language Processing), e rende possibile il dialogo tra persona e computer nel linguaggio di tutti i giorni.

Grazie a dieci anni di leadership nel settore, possediamo la migliore tecnologia per gestire le lingue in tutti i loro aspetti peculiari. Solo gli strumenti Expert System, quindi, sono in grado di estrapolare *i concetti*, cioè capire "di cosa si sta parlando", requisito essenziale per ottenere il vero miglioramento rispetto ai sistemi attuali di interazione.

Se tante attività si spostano "on line", è vantaggioso per tutti ridurre la distanza tra chi ha e chi non ha la conoscenza necessaria ad affrontare il cambiamento: Internet deve perciò diventare sempre più *Natural* Internet.

Nelle pagine che seguono scoprirete le caratteristiche e le potenzialità di Cogito®, la tecnologia NLP su cui sono basate le nostre rivoluzionarie soluzioni.

Marco Varone
Chief Technical Officer
Expert System S.p.A.

Sommario

Il problema della gestione dell'informazione	1
La soluzione: COGITO [®]	2
Le caratteristiche esclusive di COGITO [®]	4
Il parser	4
Il lessico	5
La memoria	5
La conoscenza	5
Rappresentazione del contenuto	6
I limiti delle altre tecnologie	7
COGITO [®] e i sistemi full text	7
Parole = stringhe di caratteri	7
Nessuna gestione dei concetti	8
Operatori booleani per la ricerca	9
Ordinamento dei risultati	10
COGITO [®] e i sistemi di pattern matching (o basati su reti neurali)	10
Fuzzy logic	10
Pattern matching	11
Co-occorrenza di termini	11
Importanza delle parole meno frequenti	12
Conclusioni	13

Il problema della gestione dell'informazione

*"Knowledge Management?
Trasferire la conoscenza da chi
la possiede a chi la richiede."*

Con l'inarrestabile affermarsi di Internet e delle tecnologie di distribuzione di massa dei contenuti, la quantità di documenti disponibili all'utenza è letteralmente esplosa e sembra non esserci limite alla proliferazione dei dati potenzialmente interessanti; grazie al grande successo della Rete tutti hanno a disposizione una quantità immensa di testi in formato elettronico.

Allo stesso modo, all'interno delle Intranet aziendali e nei documenti e mail personali la quantità di materiale da gestire aumenta in modo esponenziale, rendendo sempre più difficile ritrovare i documenti che servono: *devono* essere da qualche parte, ma rintracciarli con i metodi tradizionali richiede troppo tempo.

I sistemi per la gestione delle informazioni testuali usati finora non sono in grado di risolvere in modo soddisfacente il problema di separare le informazioni utili dal rumore di fondo.

Gli utenti si trovano sempre più spesso in queste situazioni critiche:

- troppe risposte (*overload*): il sistema non è in grado di catalogare (in ingresso) e ordinare (in uscita) le informazioni in modo utile, quindi produce un grande quantità di risposte non rilevanti ("scoria"), all'interno delle quali, da qualche parte, si trova la risposta cercata ("la pepita"). Di solito non si ha il tempo di setacciare il materiale restituito, esaminando una per una tutte le risposte.
- poche o nessuna risposta (*underload*): per l'inefficacia della fase di catalogazione o la scarsa capacità di discriminare del motore di ricerca, il sistema non capisce la domanda o non è in grado di ricondurla ad alcuna risposta.

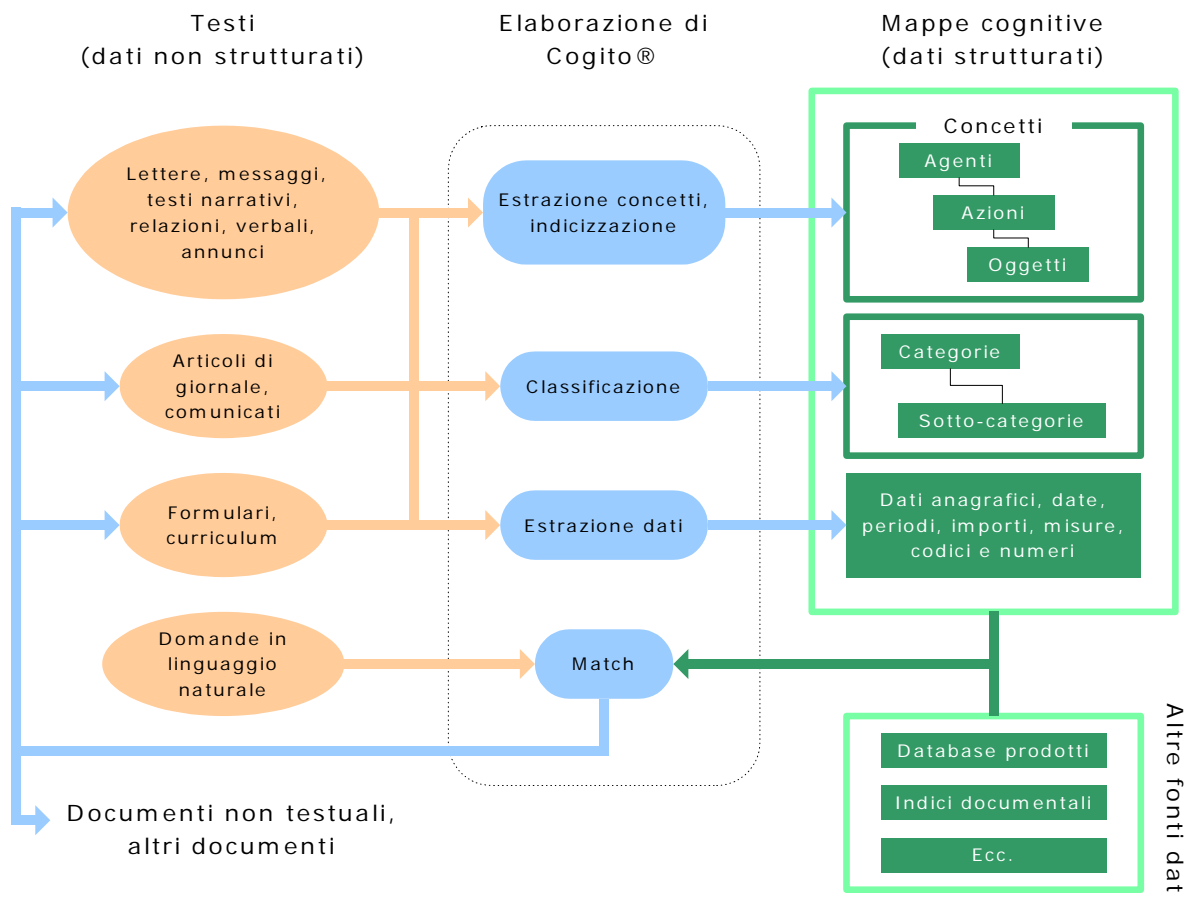
Internet e le tecnologie collegate non potranno esprimere al meglio le proprie potenzialità finché questi problemi non saranno risolti.

La soluzione: COGITO®

La nostra soluzione è **COGITO®**, risultato di dieci anni dedicati allo sviluppo di tecnologie linguistiche d'avanguardia impiegate per prodotti di grande successo.

COGITO® è un sistema software che "capisce" la lingua in un modo analogo a quanto fanno gli esseri umani, perché cattura tutti gli aspetti strutturali e lessicali del testo per comprenderne il significato.

Il risultato dell'elaborazione di **COGITO®** è una mappa cognitiva e concettuale, vale a dire una rappresentazione *strutturata* degli aspetti qualificanti del flusso di dati *non strutturati* in ingresso. La strutturazione dell'output consente ogni tipo di trattamento automatico degli elementi più significativi dei documenti.



COGITO® - Schema funzionale

I sistemi concorrenti si fermano ad un livello molto superficiale di elaborazione. Tanto per fare un esempio, questi prodotti ignorano completamente le cosiddette *stop words*, parole che (come gli articoli o le preposizioni) non hanno un significato proprio; eppure sono elementi fondamentali per dare senso compiuto al discorso. **COGITO®**, invece, prende in considerazione tutte le parole, così come fa una persona che legge o scrive.

Per capire quanto sia importante questa caratteristica, consideriamo due frasi:

(a) *Mi hanno fornito **una** carta **di** credito.*

(b) *Mi hanno fornito **della** carta **a** credito.*

Queste due frasi differiscono solo per un articolo ed una preposizione, ma hanno un significato completamente diverso. Nella frase (a) le parole "carta" e "credito" fanno riferimento ad un preciso strumento di pagamento, mentre nella frase (b) "carta" ha il suo significato più comune e "credito" si riferisce all'accordo intercorso tra il fornitore e chi ha ottenuto la carta. Per una persona, questa diversità di significato è ovvia, ma non è così per i programmi attuali di gestione del testo: dal loro punto di vista le due frasi sono identiche!

COGITO®, invece, riconosce le differenze ed individua due concetti differenti, concetti che potranno poi essere elaborati in modo distinto.

COGITO® è una tecnologia generale, scalabile e personalizzabile che può essere utilizzata per elaborare contenuti appartenenti a qualunque dominio semantico o settore di attività: documentazione tecnica specialistica, materiale enciclopedico, informazioni finanziarie, articoli di giornale, manualistica e così via.

Gli esempi che seguono e che illustrano il funzionamento di **COGITO®** rispetto agli altri metodi di elaborazione dei testi sono per la lingua italiana, seppure oggi la tecnologia linguistica di Expert System sia disponibile anche per l'inglese e il francese.

Le caratteristiche esclusive di COGITO®

COGITO® è uno strumento unico al mondo grazie alla potenza ed al livello d'integrazione dei suoi elementi, i principali dei quali sono:

- il *parser*: analizza la frase dal punto di vista morfologico, grammaticale e sintattico
- il *lessico*: risorse per il riconoscimento delle parole e dei possibili significati
- la *memoria*: storia delle analisi precedenti
- la *conoscenza*: risorse per rappresentare la conoscenza del mondo reale
- la *rappresentazione del contenuto*: il testo in forma cognitiva e strutturata

Il parser

Per comprendere il significato di una frase bisogna determinare innanzitutto il ruolo *grammaticale* di ogni parola. Nelle due frasi:

- (a) Alcune **scolare** si sono presentate in ritardo.
- (b) Ricordiamo di **scolare** molto bene la pasta prima di condirla.

appare la parola "scolare" con tipo grammaticale differente: nella frase (a) si tratta del plurale del sostantivo "scolara" mentre nella (b) è l'infinito del verbo "scolare".

I sistemi tradizionali considerano uguali le due parole, mentre COGITO® le riconosce come due significati diversi.

Altrettanto importante è riconoscere una parola indipendentemente dalla *forma* in cui è scritta; in italiano, sostantivi, aggettivi e verbi hanno diverse forme:

- (a) Marcello Mastroianni è stato l'**attore** italiano più conosciuto all'estero.
- (b) Oggi è difficile che **attrici** non più giovani abbiano ruoli da protagonista.

Nella frase (a) la forma usata è "attore", nella (b) è "attrici", ma la forma base è la stessa, come pure il concetto espresso è lo stesso (diverso è solo il genere - maschile/femminile - e il numero - singolare/plurale).

Con le forme flesse dei verbi, le possibilità sono molto più numerose. COGITO® riconduce correttamente le varie forme alla forma base invece di individuare semplicemente *n* parole diverse, come fanno gli altri sistemi.

Il parser di COGITO® gestisce in modo completo ed ottimale tutte le caratteristiche grammaticali della frase, effettua l'analisi logica e fornisce una base solida ai moduli che analizzano il contenuto.

Il lessico

Le informazioni sui significati possibili delle parole sono essenziali per la corretta interpretazione del contenuto di un testo.

Un esempio fa immediatamente capire come la ricchezza di significati sia fonte di problemi d'interpretazione:

- (a) *I due litiganti si sono scambiati **calci** e pugni.*
- (b) *Con la lente vide alcune piccole incisioni sul **calcio** della pistola.*
- (c) *Il campionato di **calcio** comincia la prima settimana di ottobre.*
- (d) *Il simbolo del **calcio** è Ca.*

Come si può vedere, una parola come "calcio" ha diversi significati e tutti devono essere identificati in modo preciso per consentire una corretta elaborazione concettuale dei contenuti.

All'interno di **COGITO**®, queste informazioni sono memorizzate in una serie di *reti semantiche* realizzate in modo specifico per l'elaborazione automatica dei testi: non semplici dizionari di termini, ma fitte reti di collegamenti e dati che consentono di rappresentare informazioni complesse, indispensabili per la disambiguazione. Grazie a queste informazioni, **COGITO**® sa che forme diverse (come "disastro aereo" e "sciagura aerea" oppure "motorino" e "ciclomotore") rappresentano in realtà lo stesso concetto, un'operazione impossibile per i sistemi che si limitano ad agire sulle parole e non sui concetti.

La memoria

Quando leggiamo, compiamo inconsciamente una serie di operazioni che ci consentono di giungere alla comprensione del testo; una di queste attività è la memorizzazione persistente dei concetti significativi delle frasi lette in precedenza.

Durante l'analisi dei documenti, **COGITO**® impiega una tecnica paragonabile per determinare il contesto semantico ideale per la disambiguazione e, successivamente, estrarre i concetti in modo preciso.

La conoscenza

La cultura è un elemento chiave per capire ciò che si legge.

Quando una persona con una buona cultura generale legge un testo specialistico sulla teoria della relatività ristretta di Einstein, comprende senza difficoltà le parole e l'impostazione generale del discorso, ma non riesce a capire veramente la *sostanza* di quanto letto a causa della mancanza della conoscenza specifica.

COGITO® contiene una conoscenza vasta e bilanciata del mondo reale, implementata sotto forma di regole descrittive che sono applicate durante l'analisi, con un meccanismo assimilabile al "buon senso" umano.

COGITO® è anche in grado di riconoscere gli elementi "particolari" (le date e i termini temporali, importi in denaro, quantità, eccetera) e di gestirli a livello concettuale.

La base di conoscenza generica di COGITO® può essere arricchita, tramite un meccanismo di apprendimento guidato, con le conoscenze specifiche di particolari domini cognitivi.

Rappresentazione del contenuto

Il risultato dell'analisi di COGITO® è una mappa cognitiva del contenuto del testo, dove:

- ogni *concetto* è memorizzato in modo indipendente dalle parole usate per rappresentarlo
- ogni *agente* è associato all'azione compiuta
- ogni *oggetto* è collegato all'azione relativa

Non solo: anche l'argomento principale del documento (ad esempio "sport" oppure "finanza") e gli eventuali argomenti secondari ("tennis" oppure "borsa"), le informazioni temporali, ed altre eventuali informazioni significative sono memorizzate in questa rappresentazione.

La caratteristica fondamentale della mappa cognitiva è quella di essere un insieme *strutturato* di dati; per questo si presta facilmente ad ogni genere di elaborazione formale come ricerche, classificazione, sintesi, traduzioni ed altro ancora.

I limiti delle altre tecnologie

COGITO® è superiore a tutte le tecnologie attuali quando si tratta di creare soluzioni di vero *knowledge management* e non semplicemente di manipolare parole in modo superficiale.

Per evidenziare i vantaggi competitivi della nostra tecnologia, basata sul riconoscimento avanzato del linguaggio naturale, è utile confrontare **COGITO®** con le principali tecnologie attualmente impiegate per la gestione del testo libero.

Il confronto riguarderà principalmente i motori di ricerca, perché sono queste le applicazioni più conosciute e diffuse, ma **COGITO®** è stato pensato per essere utilizzato in molte altre applicazioni, come la classificazione dei contenuti, l'interfaccia conversazionale in linguaggio naturale (l'utente pone domande nel linguaggio di tutti i giorni e il sistema risponde in modo appropriato, mostrando il risultato di una ricerca in un database testuale oppure "portando" l'utente in un punto specifico all'interno di un sito di grandi dimensioni, eccetera) o la traduzione automatica.

COGITO® e i sistemi full text

I sistemi "full text indexing & retrieval" operano su una versione "ripulita" del testo, da cui, cioè, sono state cancellate tutte le parole ad altissima frequenza che non hanno un significato proprio, come gli articoli e le preposizioni. Per questo genere di sistemi, il documento così filtrato non è altro che un insieme di stringhe di caratteri che compaiono un certo numero di volte.

Un approccio di questo tipo ha il suo forte nella fase di indicizzazione che è semplice e veloce, ma presenta difetti che lo rendono poco adatto alla ricerca di informazioni in modo selettivo, manifestando al contempo problemi di *overload* e *underload*: prendiamo in esame i limiti principali.

Parole = stringhe di caratteri

Consideriamo quest'importante annuncio di un famoso industriale:

Quest'anno chiuderemo due stabilimenti in Italia e ne apriremo altrettanti in Polonia

Il sistema full text indicizza letteralmente le parole "chiuderemo" e "apriremo", senza riconoscere in esse forme dei verbi "chiudere" e "aprire". Nessuno, però, cercando notizie sulla chiusura di stabilimenti, penserà di utilizzare la parola "chiuderemo"; quindi la frase diventa in pratica introvabile.

Considerando le parole come semplici stringhe di caratteri da indicizzare così come sono, senza analizzarle in alcun modo, i sistemi full text perdono un patrimonio di informazioni contenute nei testi, pur impiegando tempo di CPU e consumando spazio disco per creare una gran quantità di "zavorra" inutile fatta di forme flesse.

Alcuni sistemi di questo tipo hanno una gestione delle forme che prevede una ricerca automatica di tutte le possibili varianti della parola, ma anche in questo caso permangono molti problemi.

Ad esempio, supponiamo di cercare informazioni su "colla epossidica"; un prodotto di questo genere troverà tutti i riferimenti alla parola "colla" e anche quelli al suo plurale "colle", ma, senza un riconoscimento del vero concetto, aggiungerà alla lista anche tutti i riferimenti a "colle" inteso come "collina": un classico caso di *overload*!

COGITO® non soffre di questi problemi perché gestisce in modo automatico e trasparente tutte le forme della lingua ed associa ad ogni termine il relativo concetto, non la sua forma ortografica.

Nessuna gestione dei concetti

Questi prodotti non sono in grado di capire il significato di un termine e, di conseguenza, non riescono a recuperare informazioni riguardanti un determinato concetto.

Per esempio, se si vogliono cercare informazioni sul "calcio" inteso come parte di un fucile o di una pistola, l'unico modo sarà di indicare al sistema di cercare la parola "calcio" insieme a "fucile" oppure insieme a "pistola", per cercare di restringere l'ambito della ricerca.

Si presenta allora un problema di *underload*, perché il sistema propone per primi i documenti che contengono entrambe le parole, relegando in fondo alla lista quei documenti potenzialmente interessanti dove la parola "calcio" è utilizzata nell'accezione giusta, ma non compare la parola "fucile".

Contemporaneamente si presenta *overload* per due motivi: primo, i documenti messi in fondo alla lista perché contengono solo la parola "calcio" possono riferirsi ad accezioni completamente diverse della parola stessa (calcio = elemento chimico, calcio = sport, eccetera); secondo, allo stesso livello di importanza nella lista, si trovano documenti non interessanti perché contengono solo la parola "fucile".

COGITO® rappresenta il contenuto di un documento sotto forma di concetti utilizzati ed evita alla radice tutti i problemi di questo tipo.

Alcuni prodotti cercano di aggirare questo problema utilizzando i sinonimi dei termini più comuni, evitando all'utente la necessità di ricordarli e scriverli quando specifica i criteri di ricerca. Questo metodo migliora l'*underload*, perché recupera più documenti potenzialmente attinenti, ma peggiora in modo rilevante l'*overload*, perché trova tutti i documenti che contengono certe parole, indipendentemente dal loro significato.

Basta un esempio per capire il problema. I sinonimi di "presa" sono:

*stretta, appiglio, conquista, espugnazione, occupazione, cattura,
selvaggina uccisa, pizzico, piccola quantità, piccola dose*

Se l'utente sta cercando "presa" nel senso di "conquista", l'utilizzo dei sinonimi aumenterà il numero dei documenti utili recuperati ma, allo stesso tempo, farà crescere in modo elevato il numero di documenti non pertinenti, dato che saranno trovati anche tutti i testi contenenti termini che non hanno nulla a che fare ("stretta", "appiglio", "pizzico" e così via).

Ancora peggiore è il caso di una ricerca di "presa" nel significato di "presa elettrica"; tutti i documenti aggiuntivi recuperati tramite i sinonimi di "presa" non sono pertinenti e l'overload diventa preponderante. COGITO® memorizza i concetti e non ha problemi a gestire correttamente questi casi in modo automatico e trasparente per l'utente.

Operatori booleani per la ricerca

COGITO® può essere utilizzato per implementare sistemi di ricerca che consentano all'utente di formulare domande nel linguaggio di tutti i giorni, come se si rivolgesse ad un'altra persona.

Ad esempio, volendo fare ricerche in un database di informazioni amministrative gestito da COGITO®, sarà possibile fare domande del tipo:

Per quali documenti è necessaria la carta da bollo?

ed ottenere in risposta le informazioni desiderate.

Con un sistema full text l'utente potrà tentare usando gli *operatori booleani* AND e OR:

(documento OR documenti) AND ("carta da bollo" OR "carta bollata" OR "carta legale")

In realtà, questa domanda non è equivalente alla prima, perché il concetto di documento non si esaurisce nelle forme "documento" e "documenti", ma ne prevede tante altre: "attestato", "attestati", "carta d'identità", "atto di nascita", "passaporto", eccetera. Occorrerà quasi sicuramente un problema di underload, ma, all'opposto, potrebbe esserci overload, dato che potrebbero essere recuperati anche i riferimenti alle parole "documento" e "documenti" come forme del verbo "documentare".

Va inoltre notato che, nell'esempio, l'utente ha inserito tutti i sinonimi in modo esplicito, un'eventualità improbabile data la difficoltà di ricordare tutte le forme possibili.

Scrivere criteri di ricerca con gli operatori booleani non è facile per gli utenti comuni, perché si tratta di una rappresentazione innaturale e anche gli esperti hanno spesso bisogno di più tentativi prima di ottenere un risultato soddisfacente: in pratica si tratta di *abbassarsi* al livello della macchina ed utilizzare il suo linguaggio.

Alcuni produttori vantano, per i propri sistemi full-text, la capacità di gestire domande in linguaggio naturale. In realtà si tratta semplicemente di ricerche con operatori booleani "mascherate". Questi sistemi si limitano a trasformare la domanda scritta nel linguaggio normale in un criterio di ricerca con operatori booleani, cancellando le parole che non hanno un significato proprio (nel nostro caso "per", "la", "da") e quelle molto comuni ("è", "e", "quali"), ottenendo criteri come (le possibilità sono molte di più):

documenti AND necessaria AND carta AND bollo
(documenti OR necessaria) AND (carta OR bollo)
documenti OR necessaria OR carta OR bollo

Il risultato è di solito pessimo, perché verrà trovata solo una parte delle informazioni desiderate (forte underload) e verrà recuperato molto materiale non attinente ("documenti" e "carta" sono termini molto comuni) con un forte overload.

Ordinamento dei risultati

Un problema da risolvere quando si cercano informazioni all'interno di un insieme molto ricco di documenti è l'ordinamento dei risultati: idealmente, dovrebbero comparire in ordine di rilevanza del contenuto, con i documenti più attinenti in cima alla lista e gli altri a seguire.

I sistemi full text utilizzano un semplice approccio statistico per ordinare il risultato: più volte compare nel documento un termine ricercato, più è giudicato rilevante il documento stesso. Questo metodo fornisce risultati accettabili in alcuni casi, ma, molto spesso, non riesce a fornire un ordine corretto, obbligando l'utente a dedicare molto tempo alla ricerca del documento più adatto tra i tanti trovati.

Un esempio può far capire perché questo approccio non è sempre adatto. In un testo che parla dell'acquisto di prodotti alimentari tipici tramite Internet, la parola "pagamento" comparirà diverse volte ed una ricerca con la parola pagamento metterà questo documento tra quelli più importanti anche se, in realtà, l'argomento vero non è questo, ma i prodotti descritti.

COGITO® risolve questo problema grazie alla possibilità di esprimere in modo più preciso la ricerca da fare (grazie alla comprensione del linguaggio naturale) e all'utilizzo delle mappe cognitive, che non fanno riferimento alle parole, ma ad un ricco insieme di informazioni di sintesi correlate tra loro.

COGITO® e i sistemi di pattern matching (o basati su reti neurali)

I sistemi di questo tipo si basano sulle stesse tecniche dei prodotti full text, ma utilizzano un'analisi statistica più avanzata dei contenuti da gestire.

Le idee adottate per migliorare il funzionamento del sistema sono:

- fuzzy logic
- pattern matching
- co-occorrenza di termini
- importanza delle parole meno frequenti

Fuzzy logic

Questo nome fa riferimento all'uso di una logica volutamente "imprecisa" per affrontare i problemi derivanti dal trattamento delle parole come semplici stringhe di caratteri.

L'idea è di usare solo la parte iniziale della parola come elemento indicativo del termine e di includere nel criterio di ricerca tutte le parole di lunghezza simile che iniziano in quel modo: per esempio, nella parola "cameriere", si considera "camerier-", per "pesca", "pesc-", per "simpatico", "simpatic-", per "accendere", "accend-".

A prima vista può sembrare una scorciatoia accettabile per gestire le forme plurali ed i verbi, ma, in realtà, i risultati sono peggiori e si ha un forte aumento dell'overload. Il motivo è semplice: l'idea è stata sviluppata pensando alla lingua inglese dove, per ottenere la maggior parte delle forme flesse, è sufficiente aggiungere la lettera "s". Nel caso della lingua italiana e della maggior parte delle altre lingue, invece, la situazione è molto più complessa e il sistema diventa controproducente.

Considerando solo una parte della parola si va incontro a moltissimi casi di ambiguità che generano overload: ad esempio, "camerier-" *funziona* e anche "simpatic-" può andare, ma già con "accend-" si manifestano problemi, perché alcune forme del verbo hanno una radice diversa ("[io] accesi", "[essi] accesero") e, ancora peggio, altre parole hanno la stessa parte iniziale e vengono perciò incluse nella ricerca ("accendigas", "accendino"). Il prefisso "pesc-" è ugualmente critico, perché sono considerati molti termini non attinenti ("pesce", "pescivendolo", "[egli] pescava").

Pattern matching

Qui l'idea di base è l'allargamento dell'*unità di riferimento* dalla parola al gruppo di parole: invece di considerare ogni parola come oggetto a sé, questa funzione considera come oggetti unici i gruppi di parole che si ripetono con una certa frequenza.

In questo modo, *collocazioni* come "carta di credito" o "ferro da stiro" sono gestite come forme uniche e non come parole spezzate, migliorando le potenzialità di ricerca.

In realtà, il sistema memorizza la coppia carta/credito come oggetto unico e, di conseguenza, trova questo gruppo anche in una frase (come la già citata "Mi hanno fornito della carta a credito.") dove il significato è completamente diverso.

Inoltre il sistema continua a trattare questi oggetti come semplici sequenze di caratteri e non come concetti e non riesce perciò a generalizzarne la gestione: "carte di credito" non viene riconosciuto come una forma di "carta di credito", ma come un oggetto diverso.

Co-occorrenza di termini

Con questo metodo si analizzano statisticamente le co-occorrenze di parole all'interno di un testo, seguendo il principio secondo cui i documenti nei quali molte parole di un gruppo compaiono insieme condividono uno stesso argomento: in questo modo si possono classificare automaticamente i documenti in base al loro contenuto. Si tratta, in effetti, di una metodologia più adatta a classificare contenuti che a migliorare la ricerca vera e propria.

Sicuramente esiste una tendenza statistica che lega la presenza di parole e la similitudine semantica, ma si tratta di una caratteristica non generalizzabile in modo affidabile e assoluto, quindi, di utilità limitata. Il motivo di questo limite è illustrato meglio da un esempio.

Supponiamo che il sistema utilizzi questa regola:

I documenti in cui compaiono assieme le parole "vecchia", "porta" e "legno" appartengono alla categoria "antiquariato", sotto-categoria "infissi".

Le parole, però, assumono un significato diverso in base al contesto in cui trovano: la parola "porta", ad esempio, presa da sola, non definisce una nozione univoca. Se cambia il contesto, così pure il significato:

- (a) *Quella vecchia **porta** è di legno pregiato.*
- (b) *La vecchia contadina olandese **porta** il legno per preparare gli zoccoli.*

Nella frase (a) la parola "porta" è un sostantivo, nella (b) è un verbo; "vecchia" è un aggettivo in (a) mentre in (b) è un sostantivo. Per questo nessun sistema che si basa solo sulla presenza congiunta di stringhe come "vecchia", "porta" e "legno" non sarà mai in grado di distinguere tra (a) e (b) e le proporrà entrambe come risultato di una ricerca per categoria, con la frase (b) che rappresenta overload.

In sostanza, la co-occorrenza di termini può essere in qualche modo utile nella ricerca di documenti simili ad altri dal punto di vista statistico, ma non consente di cercare informazioni in modo più selettivo rispetto ad una normale ricerca full text.

Importanza delle parole meno frequenti

Un'altra funzionalità implementata in questi sistemi è la gestione speciale delle parole statisticamente meno frequenti nei documenti presi in esame; la teoria sottostante sostiene che le parole più informative di un testo siano quelle meno comuni e che, di conseguenza, a tali parole debba essere attribuita un'importanza maggiore.

Quest'impostazione fornisce un vantaggio per la classificazione dei documenti, perché riflette una proprietà statisticamente osservata, ma non aggiunge valore alla ricerca di informazioni.

L'informazione contenuta in un documento è espressa *anche* attraverso le parole più comuni, perché proprio queste contribuiscono a dare senso compiuto alle frasi. Un sistema che si limiti ad utilizzare i termini meno frequenti come indicativi del contenuto non potrà mai essere uno strumento veramente efficace di gestione delle informazioni. Se chi scrive impiega *tutte* quelle parole (neppure una di meno), un motivo c'è: è perché ritiene che tutte siano necessarie.

Dato che il testo è normalmente destinato ad essere letto e compreso da altre persone, solo un approccio analitico ad imitazione di quello umano può ottenere i migliori risultati.

Conclusioni

COGITO® è la più avanzata tecnologia di Natural Language Processing disponibile per la lingua italiana e tra le più potenti in assoluto per l'inglese e il francese. **COGITO®**, inoltre, sta crescendo applicato anche ad altre lingue alle quali sono trasferibili i risultati raggiunti grazie alla raffinata ricerca e sviluppo compiuta sull'italiano.

L'interpretazione del linguaggio naturale ed il riconoscimento dei concetti che **COGITO®** rende possibili hanno innumerevoli applicazioni pratiche, eccone alcune:

<i>settore</i>	<i>applicazioni</i>
Internet customer support CRM (Customer Relationship Management)	<ul style="list-style-type: none"> - sistemi di self help per Call Center ed help desk - sistemi di risposta automatica a FAQ - gestione non presidiata della posta elettronica (classificazione, risposta/smistamento) - navigazione web dinamica con interfaccia in linguaggio naturale - advertising personalizzato - customer profiling basata sull'analisi dei quesiti
e-commerce Customer Sales Representatives	<ul style="list-style-type: none"> - assistenza all'acquisto (commesso virtuale) - personalizzazione offerta prodotti
Knowledge management Motori di ricerca	<ul style="list-style-type: none"> - sistemi per la classificazione intelligente - search engine <i>concept sensitive</i> - front end in linguaggio naturale per sistemi di ricerca tradizionali - navigazione e ricerca immediata "Point&Go"
Elaborazione testi Traduzione	<ul style="list-style-type: none"> - sistemi di correzione automatica - sistemi qualitativamente superiori di supporto alla traduzione

COGITO® language technology ergo... Semantic Intelligence

COGITO® è l'esclusiva piattaforma semantica di Expert System per un'efficace gestione della conoscenza: dalla ricerca ed estrazione all'analisi, classificazione e trasformazione delle informazioni non strutturate. Insieme di tecnologie e risorse frutto di centinaia di anni/uomo di ricerca e sviluppo, COGITO® è la migliore tecnologia per la comprensione semantica disponibile sul mercato perché supera qualsiasi altro approccio di trattamento automatico della lingua ed è dotato di una rete semantica - chiamata Sensigrafo® - che non teme confronti:

- più di 520.000 concetti
- più di 2.800.000 connessioni (in riferimento a contesti, tipi di costruzione, argomenti e domini, locuzioni, fraseologie...)
- più di 400.000 collegamenti di iponimia e iperonimia
- più di 55.000 collegamenti di iperonimia e troponimia
- più di 370.000 collegamenti per il corpus e decine di migliaia di collegamenti tra soggetti, oggetti, meronimi...

Expert System S.p.A.

Via Virgilio 56/Q
41100 Modena - Italy

tel: +39 059 894011
fax: +39 059 894099

Via Machiavelli 47
00185 Roma - Italy

www.expertsystem.it
info@expertsystem.it

Expert System S.p.A. è leader di mercato nella realizzazione di soluzioni avanzate di Semantic Intelligence per aziende ed enti governativi.

Unica realtà italiana a fornire a Microsoft™ tecnologie avanzate integrate in tutti i suoi principali prodotti, Expert System fa della **gestione "intelligente" delle informazioni non strutturate** il proprio core business.