Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Tecniche LSA e Chat-bot per il recupero automatico di informazioni

Agnese Augello, Giovanni Pilato, Giorgio Vassallo, Salvatore Gaglio

RT-ICAR-PA-04-09 **luglio 2004**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.na.icar.cnr.it*
– Sezione di Palermo, Viale delle Scienze, 90128 Palermo, URL: *www.pa.icar.cnr.it*

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Tecniche LSA e Chat-bot per il recupero automatico di informazioni

Agnese Augello[2], Giovanni Pilato[1] , Giorgio Vassallo[2],
Salvatore Gaglio[1,2]

Rapporto Tecnico N.9: ***Data:***
RT-ICAR-PA-04-09 **luglio 2004**

[1] Istituto di Calcolo e Reti ad Alte Prestazioni, ICAR-CNR, Sezione di Palermo Viale delle Scienze edificio 11 90128 Palermo
[2] Università degli Studi di Palermo Dipartimento di Ingegneria Informatica Viale delle Scienze 90128 Palermo

Sommario

L'obiettivo del lavoro è la realizzazione di un sistema di reperimento dell'informazione basato sul dialogo in linguaggio naturale. Il sistema permette agli utenti di interagire con una comunità di chat-bot competenti in determinati argomenti, con lo scopo di navigare in uno spazio concettuale generato automaticamente con la metodologia di analisi della semantica latente (LSA). La base di conoscenza di ciascun chat-bot è stata codificata nello spazio semantico creato con la LSA. Grazie a questo approccio, i chat-bot sono in grado di stimare la propria competenza semantica relativamente alle domande dell'utente.

1. Introduzione

I sistemi di recupero dell'informazione si occupano della rappresentazione, memorizzazione, organizzazione ed accesso ai contenuti informativi di una collezione di documenti[1].

La rappresentazione e la memorizzazione di tali contenuti deve essere effettuata in modo da permettere ad un utente di accedere, mediante un sistema interattivo, ai documenti più rilevanti per le sue esigenze informative. I sistemi più diffusi restituiscono, a partire da una richiesta formulata mediante un insieme di parole chiave, i documenti contenenti tali parole.

Tali sistemi realizzano interfacce poco usabili poiché obbligano gli utenti a sintetizzare i contenuti ricercati in un elenco di parole spesso non sufficienti a rappresentare il contenuto semantico dell'informazione ricercata. Inoltre con tale strategia di reperimento insorgono problemi legati alle proprietà di sinonimia e polisemia dei termini, con la conseguente distorsione dei risultati ottenuti.

I risultati generalmente vengono valutati mediante le misure di Precisione e Richiamo; la Precisione è definita come la frazione di documenti reperiti che è rilevante, mentre il Richiamo è definito come la frazione di documenti rilevanti che viene recuperata[1]. Tali limiti divengono ancora più evidenti nel caso si voglia realizzare un sistema di dialogo capace di rispondere alle interrogazioni dell'utente espresse in linguaggio naturale con un contenuto informativo coerente.

Il modo più semplice per l'implementare un sistema di dialogo consiste nell'utilizzare agenti software denominati chat-bot che interagiscono con l'utente mediante un meccanismo simile a quello utilizzato dai sistemi di ricerca tradizionali. Infatti tali sistemi cercano una corrispondenza lessicale tra la domanda dell'utente e i moduli domanda-risposta della propria base di conoscenza e pertanto soffrono degli stessi limiti indicati precedentemente per i sistemi di recupero delle informazioni.

L'obiettivo di questo lavoro è la realizzazione di un sistema di interazione con una comunità di chat-bot, aventi competenze specifiche, in grado di superare tali limiti mediante l'applicazione della metodologia di analisi della semantica latente (LSA). Tale metodologia permette di inferire le relazioni latenti tra le parole appartenenti ad un grande insieme di testi; le parole vengono rappresentate da vettori all'interno di uno spazio semantico di grandi dimensioni e per stabilire la similarità tra le parole, viene applicata una misura di distanza tra i vettori che le codificano.
Gli studi sull'analisi semantica latente hanno dimostrato che i vettori che codificano parole che tendono ad apparire negli stessi contesti, distano poco all'interno dello spazio semantico[2]. Nello specifico l'applicazione della metodologia di analisi della semantica latente unita all'utilizzo della tecnologia dei chat-bot ALICE ha consentito la realizzazione di chat-bot in grado di verificare la propria competenza 'semantica' rispetto alle domande formulate dall'utente aumentando in tal modo l'efficacia del sistema. L'analisi semantica permette nel corso del dialogo di inferire il contesto nel quale rientra la domanda espressa dall'utente facendo intervenire nella conversazione il chat-bot (o i chat-bot) più preparati sull'argomento. Il sistema grazie all'integrazione delle due tecnologie ha permesso di incrementare in maniera significativa l'efficacia della ricerca valutata in termini di Precisione e Richiamo nell'informazione reperita e la qualità dell'interfaccia valutata in termini di usabilità.

2. I Chat-bot ALICE

Chat-bot è un termine relativamente recente nato per indicare agenti software in grado di sostenere attraverso interfacce multimediali una conversazione in linguaggio naturale (to chat), con un interlocutore umano.
I chat-bot rappresentano un approccio concreto al problema dell'interazione uomo-macchina e come tali, un elemento fondamentale nel processo di creazione di software intelligenti in grado di emulare il comportamento umano.
Il funzionamento dei chat-bot si basa essenzialmente sul riconoscimento, all'interno delle frasi dell'utente, di schemi fissi o parole chiave, a partire dalle quali si costruiscono le risposte.
L'interlocutore mediante un terminale fornisce la domanda ed il chat-bot, analizzando la sua base di conoscenza, restituisce la risposta più adatta, ovvero quella che realizza la corrispondenza (to match) migliore con la richiesta dell'utente.
Fra i software realizzati per la creazione di chat-bot particolare interesse viene riservato al progetto ALICE (*Artificial Linguistic Computer Entity*) [3].
ALICE fa uso di una base di conoscenza costituita a partire da una collezione di documenti contenenti moduli domanda-risposta. Tali moduli sono descritti mediante *l'Artificial Intelligence*

Mark-up Language (AIML)[4] un linguaggio di mark-up basato su XML. Il generico modulo domanda-risposta in AIML viene chiamata *categoria* ed il tag utilizzato per descriverla è <*category*>.

Ogni categoria è composta da una domanda, da una risposta e da un contesto opzionale. La domanda è chiamata *pattern*, il tag usato per descriverla è <*pattern*> e contiene un'espressione che può consistere di frasi o porzioni di frasi in linguaggio naturale, singole parole e simboli wildcard (*, _) utilizzati per indicare una corrispondenza con una qualsiasi parola[1].

La risposta è detta *template*, il tag usato per descriverla è <*template*> e contiene una espressione che può consistere di parole in linguaggio naturale e tag AIML.

I tag AIML possono trasformare la risposta in un piccolo software che può salvare dati, attivare altri programmi, dare risposte condizionali o richiamare ricorsivamente altre categorie.

I due tipi di contesto opzionale sono i tag <*that*> e <*topic*>, che consentono di spostare la conversazione su un determinato argomento (<*topic*>) o di tenere traccia di quanto detto durante la conversazione (<*that*>).

Le categorie sono essenzialmente di tre tipi:

- atomiche
- predefinite
- ricorsive

Le categorie atomiche costituiscono il tipo più semplice di categoria; il pattern non contiene wild-card ed il template è una semplice frase in linguaggio naturale.

Le categorie predefinite permettono al chat-bot di rispondere quando il pattern corrisponde parzialmente alla domanda; ciò viene ottenuto con l'inserimento dei simboli di wild-card all'interno del pattern.

Infine le categorie ricorsive sono caratterizzate dall'avere il tag <*srai*> all'interno del template, che esegue la chiamata ricorsiva ad un'altra categoria. Tali categorie permettono di rendere verosimile il dialogo con l'utente implementando forme di:

- sinonimia: associando la stessa risposta a pattern diversi che esprimono lo stesso concetto;
- riduzione simbolica: riconducendo forme grammaticali complesse a forme più semplici;
- correzione grammaticale: riconducendo le espressioni inserite scorrettamente dall'utente ad espressioni grammaticalmente corrette.

Per migliorare l'efficienza della ricerca dei pattern e per avere una rappresentazione compatta in memoria, il software AIML memorizza tutte le categorie in un grafo gestito da un elemento del

[1] I simboli wildcard (*,_) indicano entrambi una generica parola ma possiedono un diverso livello di precedenza nell'algoritmo di pattern-matching utilizzato.

sistema denominato Graphmaster. I rami del grafo corrispondono alle parole che costituiscono il bagaglio lessicale del chat-bot, i percorsi dalla radice ad un nodo foglia corrispondono ad una domanda dell'utente e puntano alla risposta corrispondente alla domanda.
Il meccanismo di dialogo di ALICE si basa su un algoritmo di ricerca chiamato *pattern matching* che cerca la corrispondenza (*matching*) della domanda con i pattern della base di conoscenza del chat-bot. La base di conoscenza dei chat-bot inoltre è realizzata attraverso un processo ciclico di apprendimento supervisionato chiamato *targeting* che coinvolge utente, chat-bot e supervisore del chat-bot (*botmaster*). Il procedimento consiste nello scorrere i documenti che memorizzano le conversazioni del chat-bot; le domande dell'utente che hanno attivato risposte errate o incomplete costituiscono degli obiettivi (targets) che offrono lo spunto per la creazione di nuove categorie. Il supervisore stabilisce quali tra questi targets diventeranno nuove categorie.

3. LSA

L'analisi della semantica latente è una teoria ed un metodo che permette, per mezzo di computazioni statistiche applicate ad un grande insieme di documenti, di inferire e rappresentare il significato delle parole in essi contenute, in base al contesto in cui le parole vengono utilizzate [2].
L'analisi della semantica latente si basa su un principio: il significato di una parola può essere definito statisticamente a partire da un grande insieme di contesti in cui è presente la parola [5]. Con il termine contesto ci si può riferire ad una frase, ad un paragrafo, o all'intera pagina di un documento. L'insieme dei contesti in cui una data parola può o non può comparire fornisce un insieme di vincoli mutui che determina la similarità di significato della parola rispetto ad altre parole o ad insiemi di documenti.
Le parole ed i documenti analizzati, vengono rappresentati da vettori all'interno di uno spazio semantico parole-documenti; la procedura matematica utilizzata per questo scopo è la decomposizione ai valori singolari (SVD: *Singular Value Decomposition*), una tecnica dell'algebra matriciale che permette di rappresentare le parole ed i documenti come vettori all'interno di uno spazio di grandi dimensioni [6].
La dimensione dello spazio viene scelta opportunamente in modo da eliminare il dettaglio informativo che viene considerato come 'rumore' e catturare le relazioni nascoste (da qui il termine 'latente') tra le parole e tra le parole e gli stessi documenti; la dimensione massima che può essere scelta è determinata dal numero dei documenti analizzati.
La scelta della dimensione ottimale è il punto cruciale dell'applicazione di questo metodo: una dimensione troppo piccola può far perdere informazione utile, mentre una dimensione troppo grande può trattenere l'informazione inutile.

E' stato dimostrato che le rappresentazioni del significato delle parole e dei documenti che possono essere ottenute con la tecnica di analisi della semantica latente, possono simulare una varietà di fenomeni cognitivi umani quali ad esempio l'acquisizione della conoscenza, la classificazione delle parole e la comprensione dell'informazione contenuta in un testo [7].

E' importante notare che per applicare questa tecnica non vengono usati dizionari o basi di conoscenza, non viene analizzato l'ordine delle parole, e non vengono nemmeno usati analizzatori sintattici o morfologici.

Un aspetto che differenzia l'analisi della semantica latente da altri metodi riguarda i dati in ingresso: questa tecnica infatti, riesce ad indurre le rappresentazioni delle parole e dei documenti a partire da associazioni tra espressioni 'unitarie' di significato, come le parole ed i documenti in cui esse compaiono e non da associazioni tra parole successive; l'efficienza della tecnica dipende dalla potente analisi matematica su cui essa si basa.

I sistemi classici di recupero dell'informazione riassumono il contenuto informativo dei documenti e delle richieste d'informazione dell'utente in un insieme di termini indice e recuperano quei documenti che effettuano una corrispondenza lessicale con la richiesta d'informazione, ossia quei documenti che contengono i termini indice contenuti nella richiesta d'informazione.

Tale approccio è semplice ma solleva molte questioni; la maggior parte della semantica di un documento o di una richiesta d'informazione viene persa nel momento stesso in cui il testo che esprime tale semantica viene rimpiazzato da un insieme di termini. Ciò è dovuto principalmente alla varietà di parole che normalmente possono essere usate per descrivere un concetto.

Tra i problemi che un sistema di recupero d'informazione deve affrontare nell'analizzare testi in linguaggio naturale, risultano evidenti quelli che derivano dalla sinonimia, ossia la possibilità di esprimere lo stesso concetto con diversi termini, e dalla polisemia, ossia la possibilità di utilizzare uno stesso termine con differenti significati in contesti differenti.

A causa di questi problemi, con una strategia di reperimento basata solo sulla corrispondenza lessicale tra il documento e la richiesta d'informazione, si rischia di ottenere un numero troppo alto o basso di risultati: a causa della sinonimia molti documenti rilevanti possono essere scartati perché non contengono gli stessi termini della richiesta d'informazione, a causa della polisemia possono essere restituiti documenti che, pur contenendo gli stessi termini della richiesta d'informazione, non sono in realtà pertinenti alla richiesta stessa.

Il significato di un testo si deduce dai concetti che in esso vengono descritti piuttosto che dalle parole utilizzate. Per questo motivo nasce la necessità di accostarsi al problema con un approccio diverso. L'analisi semantica latente può essere applicata per questo scopo, e l'applicazione della

tecnica finalizzata al recupero d'informazione viene chiamata indicizzazione basata sulla semantica latente (LSI: *Latent Semantic Indexing*).

L'indicizzazione basata sulla analisi della semantica latente utilizza la tecnica di decomposizione ai valori singolari (SVD) per ridurre le dimensioni dello spazio semantico parole-documenti e per cercare di risolvere i problemi dovuti alla polisemia e alla sinonimia che costituiscono un grosso ostacolo ai sistemi automatici di reperimento d'informazione.

L'indicizzazione basata sulla analisi della semantica latente si basa sull'assunzione che la variabilità della scelta delle parole nel descrivere un concetto oscura parzialmente la struttura semantica del documento. Riducendo la dimensione dello spazio termini-documenti viene eliminato il dettaglio informativo e possono essere rivelate le relazioni semantiche nascoste.

Tale metodo analizza statisticamente l'utilizzo delle parole nell'intera collezione di documenti e riesce a posizionare vicini nello spazio quei documenti che pur non avendo termini in comune sono correlati semanticamente.

L'applicazione della tecnica coinvolge l'esecuzione delle seguenti attività:

- Raccolta e pre-elaborazione dei documenti;
- Costruzione della matrice di co-occorrenze parole-documenti;
- Pesatura della matrice;
- Decomposizione ai valori singolari della matrice.

L'applicazione della indicizzazione basata sulla semantica latente richiede la raccolta di un insieme sufficientemente ampio di documenti, dove per documento si intende l'intero testo, un paragrafo, o anche una singola frase.

Un numero elevato di documenti permette di ottenere una maggiore completezza, poiché aumenta il grado di copertura dello spazio dei termini, ed una maggiore precisione, poiché se aumenta il numero di documenti aumenta la possibilità di trovare le stesse parole in contesti diversi; una conseguenza si riscontra nel fatto che le relazioni indotte sono più significative.

I documenti raccolti vengono sottoposti ad una fase di pre-elaborazione che permette di eliminare la punteggiatura, gli spazi e le parole comuni e quindi di ottenere i termini significativi per i documenti.

Successivamente viene costruita una matrice A di dimensioni $m \times n$, in cui le m righe corrispondono agli m termini indice estratti dall'insieme di documenti e le n colonne corrispondono agli n documenti; il generico elemento $a_{i,j}$ della matrice indica il numero di occorrenze del termine i-esimo all'interno del documento j-esimo.

La matrice A è molto sparsa, poiché il numero di termini contenuti in un generico documento è generalmente di gran lunga minore del numero di termini presenti nell'intera collezione di documenti.

Per incrementare o decrementare l'importanza relativa dei termini all'interno del singolo documento e lungo l'intera collezione di documenti si utilizza uno schema di pesatura; ogni valore di occorrenza viene sostituito da un valore che rappresenta il peso di ciascun termine[8], dato da:

$$d_{ij} = L_{ij} \cdot G_i \cdot N_j$$

dove:

- L_{ij} è il 'Peso Locale', ossia il peso del termine i-esimo nel documento j-esimo;
- G_i è il 'Peso Globale', ossia il peso del termine i-esimo nell'insieme dei documenti;
- N_j è il 'Fattore di Normalizzazione'.

In letteratura esistono vari schemi di pesatura che combinati opportunamente tra loro influiscono in modo determinante sull'efficacia della tecnica di indicizzazione con l'analisi semantica latente[9].

La matrice ottenuta viene sottoposta ad una delle più importanti decomposizioni ortogonali derivanti dall'algebra lineare numerica: la decomposizione ai valori singolari (SVD). Tale decomposizione è comunemente usata nella soluzione dei problemi lineari ai minimi quadrati vincolati e non vincolati, per la stima del rango di una matrice, per stimare i coefficienti nell'analisi di correlazione canonica ed in un vasto insieme di applicazioni quali ad esempio il reperimento di informazione e la tomografia di riflessione sismica[10].

In particolare la SVD troncata ad una dimensione minore del rango della matrice può essere utilizzata per generare una approssimazione di rango $k<<r$, dove r è il rango della matrice.

La decomposizione ai valori singolari della matrice A troncata ad una dimensione k, con $k<<r$ è data da:

$$A_k = U \cdot \Sigma \cdot V^T$$

con U di dimensioni $m \times k$, V di dimensioni $n \times k$ e Σ di dimensioni $k \times k$. A_k è una approssimazione della matrice A che permette di evidenziare informazioni importanti sulla struttura della matrice di partenza.

4. Tecniche LSA e Chat-bot per il recupero automatico di informazioni

In questo lavoro è stato implementato un sistema di reperimento dell'informazione accessibile mediante un'interfaccia in linguaggio naturale. L'obiettivo perseguito è stato quello di permettere ad un utente di cercare l'informazione desiderata esprimendo le proprie richieste in modo informale, mediante la formulazione di domande in linguaggio naturale.

Il sistema è stato realizzato con la combinazione di due tecnologie: la tecnologia dei chat-bot ALICE e la tecnica di analisi della semantica latente; l'utente dialoga con una comunità di chat-bot per ottenere l'informazione desiderata. I chat-bot implementano la conversazione mediante un meccanismo simile a quello utilizzato dai sistemi di ricerca tradizionali, ovvero cercando una corrispondenza lessicale tra la domanda dell'utente e le categorie della propria base di conoscenza.

I limiti dovuti ad un simile approccio possono essere superati potenziando il meccanismo di risposta dei chat-bot con un sistema di ricerca realizzato con la tecnica di analisi della semantica latente; i chat-bot in questo modo possono rispondere adeguatamente se esiste almeno una categoria che, pur non corrispondendo lessicalmente alla domanda, risulta semanticamente correlata ad essa.

Una proprietà fondamentale dell'analisi della semantica latente è che tale metodologia permette di inferire le relazioni sottostanti tra le parole appartenenti ad un vasto insieme di testi; le parole vengono rappresentate da vettori all'interno di uno spazio semantico di grandi dimensioni e per stabilire la similarità tra le parole, viene applicata una misura di distanza tra i vettori che le codificano. Gli studi sull'analisi della semantica latente hanno dimostrato che i vettori che codificano parole che tendono ad apparire negli stessi contesti distano poco all'interno dello spazio semantico[2]. In questo modo è possibile effettuare una discriminazione tra i diversi argomenti di conversazione.

Tale proprietà ha suggerito la realizzazione dell'interfaccia al sistema di reperimento dell'informazione con una comunità di chat-bot esperti in determinati argomenti; ciascun chat-bot viene precedentemente addestrato per un particolare dominio di conoscenza, con un insieme di file contenenti le categorie AIML opportune. L'analisi semantica permette nel corso del dialogo di inferire il contesto nel quale rientra la domanda espressa dall'utente facendo intervenire nella conversazione il chat-bot (o i chat-bot) più preparati sull'argomento. La scelta di combinare le due tecnologie è stata effettuata al fine di incrementare due parametri fondamentali nella valutazione di un sistema di ricerca: l'efficacia della ricerca che può essere misurata in termini di Precisione e Richiamo nell'informazione reperita[1], e la qualità dell'interfaccia che può essere valutata in termini di usabilità.

4.1 Architettura del sistema

Il sistema realizzato permette di ottenere informazioni su un insieme di argomenti specifici ed è composto da una comunità di *n* chat-bot, di cui $n - 1$ sono esperti in argomenti specifici e l'ultimo è un chat-bot in grado di rispondere ad argomenti di carattere generale.

I chat-bot si alternano nel dialogo con l'utente intervenendo nella conversazione quando il proprio livello di competenza sull'argomento trattato dalla domanda raggiunge un valore maggiore o uguale rispetto ai livelli di competenza degli altri chat-bot; questo vuol dire che più chat-bot potranno intervenire contemporaneamente nel corso del dialogo se sono preparati in egual misura sull'argomento della domanda. Per valutare il livello di competenza di ciascun chat-bot rispetto ad ogni domanda, e per implementare il meccanismo di risposta dei chat-bot è stata applicata la tecnica di analisi della semantica latente.

Per la realizzazione del sistema è risultato necessario implementare:

- un modulo per l'applicazione dell'analisi semantica latente;
- un modulo per la creazione e l'addestramento dei chat-bot;
- un modulo per l'implementazione del dialogo con i chat-bot;
- un modulo per la valutazione del sistema di reperimento dell'informazione realizzato.

4.2 Applicazione dell'analisi della semantica latente

L'applicazione della metodologia di analisi della semantica latente permette di creare uno spazio semantico in cui ad ogni parola appartenente ad una collezione di documenti viene associata uan codifica sub-simbolica data da un vettore di numeri reali che ne rappresenta il contenuto semantico. Per potere ottenere questa rappresentazione devono essere effettuate le seguenti attività:

- Costruzione della matrice di co-occorrenze parole-documenti;
- Pesatura della matrice di co-occorrenze;
- Decomposizione ai valori singolari della matrice di co-occorrenze.

Per ottenere la matrice di co-occorrenze viene considerato un elevato numero di documenti, riguardanti gli argomenti per i quali devono essere addestrati i chat-bot. I documenti vengono analizzati ed elaborati mediante l'eliminazione di punteggiatura e parole-comuni in modo da estrarne le parole rappresentative. La matrice viene costruita memorizzando le informazioni riguardanti le occorrenze delle parole estratte dai diversi documenti in una struttura di indicizzazione.

Si indichi con $A(m \times n)$ la matrice ottenuta. Le *m* righe della matrice corrispondono alle *m* parole distinte estratte dai documenti, le *n* righe corrispondono ai documenti. L'elemento *a*(*i, j*) indica il numero di occorrenze della parola i-esima nel documento j-esimo.

I passi effettuati per la costruzione della matrice di co-occorrenze vengono descritti mediante la Figura1.

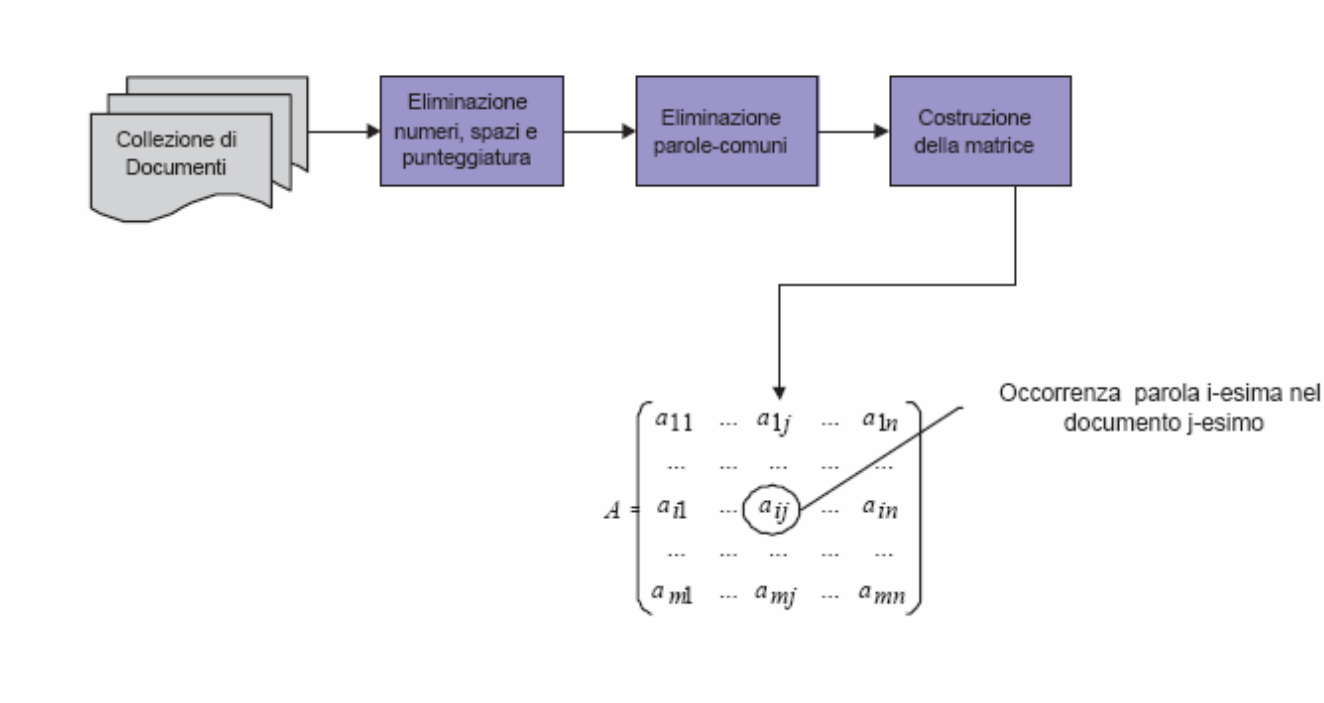

Figura 1: Costruzione della matrice delle occorrenze dei termini all'interno della collezione di documenti

La matrice ottenuta viene pesata opportunamente per tenere conto dell'importanza di ciascuna parola all'interno dello specifico documento e all'interno dell'intera collezione di documenti. Il valore di occorrenze dell'elemento della matrice $m(i, j)$ viene sostituito dal valore $d(i, j)$ dato dal prodotto:

$$d_{ij} = L_{ij} \cdot G_i \cdot N_j = SQRT_{ij} \cdot IGFF_i \cdot COSN_j$$

dove L_{ij} è il peso del termine i-esimo all'interno del documento j-esimo, G_i è il peso del termine i-esimo all'interno della collezione di documenti e N_j è il fattore di normalizzazione. Indicando con f_{ij} il numero di occorrenze dell'i-esimo termine nel j-esimo documento, n_i il numero di documenti in cui compare l'i-esimo termine, F_i la frequenza dell'i-esimo termine nella collezione di documenti, allora:

- $SQRT_{ij} = \begin{cases} \sqrt{f_{ij} - 0.5} + 1 & \text{se } f_{ij} \rangle 0 \\ 0 & \text{se } f_{ij} = 0 \end{cases}$
- $IGFF_i = \frac{F_i}{n_i}$
- $COSN_j = \frac{1}{\sqrt{\sum_0^m \left(IGFF_i \cdot SQRT_{ij}\right)^2}}$

Successivamente la matrice viene decomposta nel prodotto di tre matrici mediante la tecnica di decomposizione ai valori singolari.

La decomposizione ai valori singolari della matrice *A* con un numero *k* di valori singolari è data da:

$$A_k = U \cdot \Sigma \cdot V^T$$

con *U* di dimensioni $m \times k$, *V* di dimensioni $n \times k$ e Σ di dimensioni $k \times k$. Con la decomposizione della matrice si ottiene una nuova rappresentazione vettoriale per le parole ed i documenti in uno spazio di dimensioni pari a *k*. Per codificare le parole vengono prese in considerazione le righe della matrice $U \cdot \Sigma$. Il vettore rappresentato dalla riga i-esima della matrice $U \cdot \Sigma$ caratterizza la posizione della parola i-esima nello spazio a *k* dimensioni. Analogamente, i vettori rappresentati dalle colonne della matrice $\Sigma \cdot V^T$ caratterizzano la posizione del documento j-esimo nello spazio a *k* dimensioni.

Si consideri la i-esima parola corrispondente alla i-esima riga della matrice; il vettore $\vec{w}_i$ di *k* componenti date da $w_{1,i}, w_{2,i} \ldots w_{k,i}$ rappresenta il contenuto semantico della parola all'interno dello spazio. Ogni parola quindi viene rappresentata all'interno dello spazio semantico da un vettore.

Per valutare la similitudine tra le parole all'interno dello spazio semantico è stata considerata come parametro di valutazione il prodotto scalare, che misura la distanza euclidea tra due vettori.

La similitudine tra le parole è definita come:

$$sim\left(w_i, w_j\right) = \vec{w}_i \cdot \vec{w}_j$$

4.3 Creazione ed addestramento dei chat-bot

L'addestramento dei chat-bot esperti avviene mediante la scrittura di file AIML contenenti un insieme di categorie riguardanti l'argomento di competenza. Ogni categoria, formata da una coppia domanda-risposta, viene codificata per mezzo di un vettore all'interno dello spazio semantico costruito in precedenza. Da ogni categoria vengono estratte le parole, ottenute mediante

l'eliminazione della punteggiatura e delle parole-comuni. Per ciascuna parola contenuta nella categoria ed appartenente allo spazio semantico costruito in precedenza viene considerata la codifica relativa, ovvero il vettore memorizzato corrispondente. Il vettore relativo alla categoria è quindi ottenuto mediante la somma pesata dei vettori corrispondenti alle singole parole.

4.4 Dialogo con i chat-bot

Il dialogo che avviene con i chat-bot si basa sulla possibilità di confrontare di volta in volta la domanda effettuata dall'utente con le categorie di ciascun chat-bot.

Per potere effettuare questo confronto l'interrogazione dell'utente viene sottoposta, come viene fatto per ogni categoria, ad un insieme di operazioni al fine di estrarne le parole che la rappresentano. Per ogni parola contenuta nell'interrogazione ed appartenente allo spazio semantico costruito in precedenza viene selezionata la sua codifica vettoriale. Il vettore relativo all'interrogazione è quindi ottenuto mediante la somma pesata dei vettori corrispondenti alle singole parole.

Ottenuta la rappresentazione vettoriale dell'interrogazione è possibile confrontarla con qualsiasi categoria dei chat-bot. Per valutare la similarità dell'interrogazione rispetto ad una categoria viene utilizzata come misura di similarità il coseno dell'angolo tra i due vettori; il coseno in questo caso viene preferito al prodotto scalare poiché permette di valutare la similarità tra documenti di diversa lunghezza. Indicando con q l'interrogazione dell'utente e $\vec{q}$ il vettore corrispondente, con $c_{i,j}$ la categoria j-esima dell'i-esimo chat-bot e $\vec{c}_{i,j}$ il vettore corrispondente, la misura di similarità tra l'interrogazione e la categoria è ottenuta come:

$$sim(c_{i,j}, q) = \frac{\vec{c}_{i,j} \cdot \vec{q}}{|\vec{c}_{i,j}| \cdot |\vec{q}|}$$

Ogni qualvolta l'utente effettua una domanda ciascun chat-bot valuta la propria competenza per potere rispondere.

Sia j l'indice che indica un determinato chat-bot, sia nc il numero di categorie per il chat-bot, sia c_{ij} la i-esima categoria del chat-bot, q l'interrogazione dell'utente e sia σ un valore di soglia determinato sperimentalmente per $sim(c_{ij}, q)$ oltre il quale la categoria c_{ij} può essere considerata pertinente all'interrogazione. Viene definito Fattore di Competenza del j-esimo chat-bot fc_j come:

$$fc_j = \max\{sim(c_{ij}, q), \text{ con } i = 1 \ldots nc | sim(c_{ij}, q) \rangle \sigma$$

Ciascun chat-bot confronta il proprio Fattore di Competenza fc relativo alla query con quello degli altri chat-bot. Il chat-bot con Fattore di Competenza maggiore interviene nella conversazione e

prova a rispondere mediante il modulo di risposta di ALICE. Tale modulo si basa su un algoritmo di *pattern matching* che cerca la corrispondenza lessicale tra la domanda e le categorie del chat-bot. Se l'esecuzione del modulo non va a buon fine, il chat-bot risponderà con la categoria che ha determinato la sua attivazione nella conversazione. Se nessuno dei chat-bot esperti è in grado di rispondere alla domanda, interviene nella conversazione il chat-bot generico.

4.5 Risultati

Per valutare i risultati ottenuti con l'approccio descritto, è stata realizzata una comunità di quattro chat-bot, uno dei quali è un chat-bot generico con base di conoscenza fornita dalla comunità free-software di ALICE; gli altri chat-bot sono stati addestrati sui seguenti argomenti:

- Architettura greca classica;
- Pittura del Rinascimento e del Manierismo;
- Scultura del Rinascimento e del Manierismo.

Per generare lo spazio semantico con la tecnica di analisi della semantica latente, sono stati selezionati dal web 850 documenti; per creare uno spazio coerente con gli argomenti di specializzazione dei chat-bot sono stati scelti documenti inerenti ai settori di competenza dei chat-bot. In particolare come documenti sono stati considerati i paragrafi delle pagine HTML o, nel caso di specifiche opere d'arte il titolo dell'opera seguito da una breve descrizione.

Da questa collezione di documenti è stata ottenuta una matrice termini-documenti di dimensioni (7965 × 850), la tecnica di SVD applicata a questa matrice ha permesso di generare uno spazio semantico con una dimensione k=100.

Successivamente sono state create le categorie AIML dei chat-bot. La Tabella1 indica il numero di file e categorie AIML create per ciascun chat-bot.

Tabella 1: Numero di File e di categorie AIML creati per ciascun chat-bot esperto

	N. File AIML	N. Categorie AIML
Esperto di Arte Greca	14	237
Esperto di Pittura	10	137
Esperto di Scultura	7	158

Tali categorie sono state codificate con vettori all'interno dello spazio semantico costruito precedentemente con la tecnica di LSA.

Sia N il numero di interrogazioni relative ad uno specifico argomento a, sia N_a il numero di volte in cui ha risposto il chat-bot esperto per l'argomento e sia N_s il numero di volte in cui ha risposto un chat-bot non esperto. Il sistema è stato sottoposto a 300 interrogazioni, 100 per ciascun settore di

competenza dei chat-bot. Per ogni gruppo di interrogazioni sono state valutate la Precisione nell'argomento P_a e lo Scarto nell'Argomento S_a, definiti come:

$$P_a = \frac{N_a}{N}$$

e:

$$S_a = \frac{N_s}{N}$$

La soglia per determinare il fattore di Competenza dei chat-bot è stata fissata sperimentalmente al valore 0.5. Nelle colonne della Tabella2 per ciascun argomento sono indicati rispettivamente:

- Il numero di volte in cui ha risposto l'esperto di Pittura;
- Il numero di volte in cui ha risposto l'esperto di Scultura;
- Il numero di volte in cui ha risposto l'esperto di Arte Greca;
- La Precisione nell'Argomento P_a;
- Lo Scarto nell'Argomento S_a.

Tabella 2: Numero di risposte ottenute da ciascun chat-bot e valori di Precisione e Scarto nell'Argomento ottenuti per ciascun gruppo di interrogazioni

	Arte Greca	Pittura	Scultura	Bot Generico	P_a	S_a
Arte Greca	84	2	7	7	0.84	0.16
Pittura	2	76	16	6	0.76	0.24
Scultura	1	16	75	8	0.75	0.25

Per valutare, per completezza, l'efficacia dell' intero sistema nel reperire correttamente l'informazione, il sistema è stato sottoposto a 63 interrogazioni per ciascun argomento. Per ogni interrogazione sono state selezionate le categorie pertinenti e sono state valutate le misure di Precisione e Richiamo ottenute[1]. I risultati ottenuti sono stati confrontati con i risultati ottenuti a partire dalle stesse interrogazioni con Lucene[11], un motore di ricerca tradizionale basato sul *matching* lessicale tra l'interrogazione ed i documenti. Per entrambi i sistemi sono state valutati i livelli standard $[0\%,10\%,\ldots 100\%]$di Precisione e Richiamo. In Figura2 è mostrata la curva di Precisione-Richiamo ottenuta. Nella Tabella3 viene riportato un esempio di conversazione tra un utente ed i chat-bot esperti.

Tabella 3: Esempio di dialogo con la comunità di chat-bot esperti

User: Who was Leonardo da Vinci?

Painting Expert: Leonardo da Vinci was a Florentine artist.

User: Who was his teacher?

Painting Expert: In Verrocchio's workshop Leonardo was introduced to many activities.

User: Could you tell me about the exponents of mannerism?

Painting Expert: Leading Mannerists include Pontormo, Bronzino, Parmigianino, El Greco ,Tintoretto, Bernini, Brunelleschi, Benvenuto Cellini, Caravaggio and Tiziano.

Sculpture Expert: Leading Mannerists include Pontormo, Bronzino, Parmigianino, El Greco, Tintoretto, Bernini, Brunelleschi, Benvenuto Cellini, Caravaggio and Tiziano.

User: I'm interested in David of Michelangelo

Sculpture Expert: David / Apollo 1530.Marble, height: 146 cm. Museo Nazionale del Bargello, Florence. The statue was commissioned by the hated Papal Governor of Florence, Baccio Valori…

User: Tell me about greek colonization in Sicily.

Greek Art Expert: The Greek colonies in Sicily preserve the clearest traces of the pre-citystate...

User: Please, speak about Agrigento's valley.

Greek Art Expert: The principal temples at the Acragas's valley are Concordia, Castor and Pollux, Hera, Hercules, Vulcan or Hephaistos,Zeus Jupiter.

User: Which is the oldest of these temples?

Greek Art Expert: The temple of Hercules is Agrigento's oldest temple and was constructed with 38 columns, of which only nine are still standing.

User: I want now to speak about generic arguments.

Generic Chat-Bot: Interesting. I want to find out more about you.

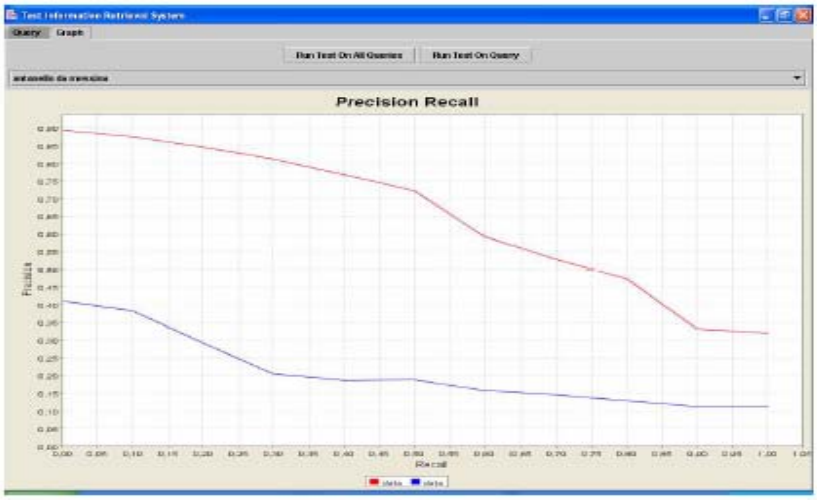

Figura 2: Curve di Precisione-Richiamo. Confronto tra Lucene ed il sistema basato sulla LSA

4.6 Conclusioni

L'obiettivo di questo lavoro è stato quello di realizzare un sistema che permettesse di interagire con una comunità di chat-bot aventi specifiche competenze. I chat-bot realizzati con tecnologia ALICE, sono in grado di stimare la propria competenza riguardo le domande formulate dall'utente; questa capacità è ottenuta mediante la metodologia di LSA, che permette di superare i limiti tradizionali delle rappresentazioni delle basi di conoscenza dei chat-bot. I risultati ottenuti mostrano che per la maggior parte delle interrogazioni viene effettuata una scelta del chat-bot specializzato coerente all'argomento della conversazione corrente. Per completezza è stata inoltre valutata l'efficacia del sistema di reperimento realizzato; sono state utilizzate a questo scopo 62 interrogazioni; per ciascuna di esse sono state selezionate le categorie pertinenti e sono state valutate le misure di Precisione e Richiamo confrontando I risultati con quelli ottenuti con un motore di ricerca tradizionale, ottenendo risultati soddisfacenti.

5. Ringraziamenti

Si ringraziano la Engineering Ingegneria Informatica S.p.A. e l'Ing. Maria Vasile per il contributo apportato.

6. Bibliografia

1. B.Ribeiro-Neto R.Baeza-Yates. *Modern Information Retrieval*. Addison Wesley, 1999.
2. S.T. Dumais, T.K. Landauer. A solution to plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 1997.
3. Artificial linguistic computer entity (A.L.I.C.E.) http://alice.sunlitsurf.com/alice/about.html
4. Artificial intelligence markup language (A.I.M.L.) http://alice.sunlitsurf.com/alice/aiml.html
5. Z. Harris. Structure distributionnelle. La grammaire,lectures, 1975.
6. F.T. Luk, G.H. Golub and M.L.Overton. A block lanczoz method for computing the singular values and corresponding singular vectors of a matrix. ACM Transactions on Mathematical Software, 7:149–169, 1981.
7. P.W. Foltz T.K.Landauer and D. Laham. An introduction to latent semantic analysis. Discourse Processes, 25:259–284, 1998.
8. G.W. Furnas; S.Deerwester; S.T.Dumais; T.K.Landauer and R.A.Harshman. Indexing by latent semantic analysis. The American Society for Information Science, 1990.
9. T.G. Kolda E.Chisholm. New term weighting formulas for the vector space method in information retrieval, 1999.
10. M.W.Berry. Large scale singular value computations. International Journal of Supercomputer Application, 1992.
11. Jakarta Lucene. http://jakarta.apache.org/lucene/