UNIVERSITÀ DEGLI STUDI DI PALERMO

FACOLTÀ DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA INFORMATICA

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

Sistema di interrogazione intelligente di Documenti ufficiali del Parlamento Europeo

TESI DI LAUREA DI:

Salvatore La Bua

RELATORE:

Ch.mo Prof. Ing. **Salvatore Gaglio**

CORRELATORI:

Ing. **Giovanni Pilato**

Dott. **Giorgio Vassallo**

ANNO ACCADEMICO 2003 - 2004

Indice

Indice dettagliato

Introduzione

Argomento principale della tesi in oggetto, e´ riuscire a sfruttare l'analisi della semantica latente [2] (cfr. Cap 2) per poter rappresentare le parole secondo il significato assunto nel contesto in cui si trovano e fornire all'utente un'interfaccia amichevole per il recupero di informazioni. Tale risultato si puo´ ottenere grazie all'analisi di grandi quantita´ di documenti, da cui si possono estrarre relazioni semantiche tra i termini che li compongono, effettuando calcoli statistici sulla frequenza di occorrenza delle singole parole nei documenti al fine di poter rappresentare (cfr. §2.4 e segg.) tutte le parole in uno spazio semantico. E´ possibile inoltre avere una conoscenza generale su un argomento specifico, in base ai documenti successivamente codificati.

Nel caso preso in esame, si e´ realizzata la rappresentazione in uno spazio semantico delle parole dei documenti in lingua inglese facenti parte dell' archivio del parlamento europeo [1].

In tal modo l'utente del sistema e´ in grado di porre all'applicazione molteplici domande circa un argomento specifico, ricevendo una risposta semanticamente legata alla domanda e non effettuando - come i piu´ diffusi sistemi per il recupero di informazioni (cfr. Cap. 1) - una ricerca lessicale per corrispondenza di termini tra la richiesta dell'utente stesso e tutti i documenti presenti nell'insieme dei testi disponibili.

Attraverso la rappresentazione dei termini nello spazio semantico e´ possibile quindi ottenere sottoinsiemi di parole strettamente legate tra

loro dal punto di vista semantico identificando vettori vicini nello spazio di codifica.

A ciascuna parola dello spazio verra´ associata una propria codifica vettoriale che la rappresenta nello spazio semantico considerato, tale codifica e´ necessaria per poter effettuare ad esempio confronti di distanza in modo da trovare parole semanticamente vicine: parole a distanza minore nello spazio n-dimensionale saranno maggiormente legate dal punto di vista semantico rispetto a quanto non lo siano parole le cui rappresentazioni vettoriali si trovino rispettivamente l'uno dall'altro ad una distanza maggiore.

Il primo capitolo riporta un'introduzione al recupero automatico dell'informazione e ai sistemi di recupero di informazione (Information Retrieval - IR), enunciando alcune delle caratteristiche principali dei sistemi di ricerca.

Nel secondo capitolo verra´ esposta la tecnica di analisi della semantica latente per la codifica delle parole e per il recupero di informazioni. Si fara´ inoltre riferimento alla scomposizione ai valori singolari (Singular Value Decomposition - SVD) della matrice utilizzata come base di partenza per la codifica dei termini.

Il capitolo terzo tratta della soluzione proposta, approfondendo le varie parti dello sviluppo dell'applicazione ed illustrando inoltre le tecniche di misura innovative utilizzate per ottenere la risposta piu´ pertinente alle richieste degli utenti.

Nell'ultimo capitolo e´ possibile trovare informazioni di carattere generale sul sistema sviluppato, collezione dei documenti, riferimenti alle piu´ diffuse interfacce di interazione uomo-macchina con un'introduzione ai chat-bot e alle differenze che intercorrono tra tali agenti software - i chat-bot - ed il sistema sviluppato in questa tesi.

Capitolo Primo

Introduzione al Recupero automatico dell'Informazione

1.1 Introduzione

Grazie soprattutto al continuo sviluppo di Internet ed alla diffusione dei documenti in formato digitale, ci stiamo rendendo sempre piu´ conto che e´ possibile disporre di qualsiasi tipo di informazione, sia essa di nostro interesse o meno. Ma e´ proprio per questo motivo - la disponibilita´ di tutte queste informazioni spesso non catalogate - che fa nascere il problema del recupero automatico e intelligente dell'informazione stessa, in modo da poter ottenere soltanto le informazioni a noi necessarie, potendole distinguere efficacemente da quelle di cui non abbiamo bisogno.

Senza adeguati strumenti di ricerca, sarebbe impossibile riuscire ad ottenere risultati soddisfacenti poiche´, in generale, le informazioni disponibili ad esempio in rete, spesso non sono caratterizzate da un adeguato metodo di catalogazione, per cui sarebbe veramente diffcile trovare cio´ che serve nonostante la disponibilita´ di cosi´ tanta conoscenza.

1.2 Recupero automatico di informazioni

Con recupero di informazioni (Information Retrieval - IR) si intende l'insieme di azioni, metodi, procedure utilizzati per recuperare dati archiviati, allo scopo di fornire informazioni su un dato argomento [8].

Un sistema di recupero di informazioni deve poter rappresentare, memorizzare, organizzare ed accedere ai contenuti informativi di una collezione di documenti [7].

Essi hanno come ingresso due categorie di dati: la prima e´ costituita dalle richieste dell'utente a cui il sistema dovra´ rispondere, mentre la seconda e´ costituita dall'insieme dei documenti da cui verranno estratte le possibili risposte.

La richiesta dell'utente - la query - dovra´ essere analizzata dal sistema per poter effettuare successivamente le ricerche all'interno della collezione di documenti, al fine di presentare all'utente stesso una o piu´ risposte di cui necessita.

I piu´ diffusi sistemi di recupero informazioni si basano su ricerche lessicali, all'interno della collezione dei documenti, delle parole chiave immesse dall'utente. Esistono inoltre metodi molto interessanti che si basano sull'analisi del significato delle parole all'interno del testo in cui si trovano; si tratta appunto della tecnica di analisi della semantica latente esposta nel capitolo secondo.

1.2.1 Caratteristiche dei sistemi di Recupero Informazioni

Come precedentemente accennato, la fonte maggiore di informazioni attualmente disponibile potrebbe essere il world wide web e, data la vastita´ delle informazioni in esso contenute, esistono diverse

caratteristiche che i sistemi per il recupero di informazioni devono soddisfare [15]. Tra queste possiamo ricordare:

1. Perfezionamento della ricerca (Relevance Feedback):
   Processo con cui l'utente perfeziona la ricerca identificando le pagine piu´ rilevanti tra quelle restituite. Questo permette al sistema di presentare all'utente una nuova lista di risultati piu´ particolareggiata rispetto alla prima.

2. Estrazione dell'Informazione (Information Extraction):
   Capacita´ del sistema di recupero informazioni (IR) di estrarre informazione dal testo, come ad esempio l'estrazione di nome, prodotti, localita´, etc. Richieste che a volte risultano molto difficili da esaudire se non si possiede una piena comprensione a priori del testo in analisi.

3. Recupero di dati multimediali (Multimedia Retrieval):
   Tecniche di accesso ad archivi di immagini, video, suoni senza descrizione testuale. Soluzioni generali in ambito multimediale sono molto complesse, come ad esempio l'indicizzazione delle immagini attraverso la distribuzione di colore.

4. Recupero di Efficacia (Effective Retrieval):
   La necessita´ di una efficacia delle ricerche e´ uno dei requisiti fondamentali di un sistema di IR. Trovare un testo che soddisfi le richieste di un utente non é semplice pero´, tramite l'introduzione di strategie di ordinamento e valutazione (ranking) sempre piu´ raffinate - che permettono di ordinare i documenti in base ad un peso di rilevanza - e´

possibile ottenere risultati migliori. E´ possibile inoltre migliorare l'efficacia delle ricerche tramite l'operazione di *stemming*, che consiste nell'effettuare ricerche in base alla radice di ogni termine cosi´ da non scartare quei documenti in cui il termine non compare nella stessa forma presente nella query, come ad esempio termini maschili/femminili oppure singolari/plurali.

5. Filtraggio secondo profili utente (Collaborative Filtering):
   Processo di identificazione dei documenti rilevanti a partire da un profilo utente: ogni documento proveniente da un flusso di dati (stream) viene analizzato per verificare se puo´ essere importante rispetto al profilo dell'utente corrente ed eventualmente gli viene mostrato.

6. Interfacciamento e navigabilita´ (Interfaces and Browsing):
   Molto importante e´ anche l'integrazione con i sistemi esistenti. Una parte altrettanto importante e che non deve essere trascurata e´ l'interfaccia del sistema stesso - che e´ cio´ con cui l'utente dovra´ dialogare - cosi´ come occorre rendere il sistema di recupero informazioni il piu´ performante e preciso possibile nella ricerca, e´ inoltre necessario progettare delle interfacce semplici, intuitive ma allo stesso tempo complete e adattabili alle esigenze dell'utente, come ad esempio permettere la formulazione di query, visualizzare messaggi di risposta all'utente, nonche´ fornire una facile interfaccia per la consultazione dei risultati che, sicuramente, e´ uno dei punti fondamentali per l'elevata usabilita´ dell'interfaccia stessa.

7. Espansione delle parole chiave di ricerca (“Magic”):
   Una causa non indifferente del fallimento di un sistema di recupero di informazioni e´ senz’altro la presenza di errori nel vocabolario. L’informazione e´ spesso descritta utilizzando differenti termini che si trovano in documenti rilevanti. Occorre quindi espandere la richiesta dell’utente per cercare non solo il termine specificato ma anche eventuali suoi sinonimi. Questo e´ possibile utilizzando un dizionario dei termini (detto thesaurus) oppure attraverso metodi di ricerca basati su indicizzazione della semantica latente (Latent Semantic Indexing) che riescono a superare tale problema in modo automatico.

8. Indicizzazione e Recupero efficienti e flessibili (Efficient, Flexible Indexing and Retrieval):
   Una caratteristica che certamente non puo´ mancare e´ l’efficienza; l’uso sempre piu´ frequente di motori di ricerca nel web ha reso essenziale avere tempi di risposta minimi a fronte di una query immessa dall’utente, specialmente a causa dell’enorme quantita´ di dati che e´ possibile reperire. Tecniche di compressione possono essere utilizzate al fine di ridurre lo spazio necessario per la memorizzazione e di conseguenza il tempo di recupero dell’informazione.

9. Sistemi distribuiti di recupero informazioni (Distributed IR):
   Grazie alla vertiginosa espansione di Internet si e´ verificata una crescita delle richieste verso i motori di ricerca e, affinche´ le risposte ad una query siano rapide, risulta necessario realizzare sistemi distribuiti per il recupero di informazioni, cosi´ da parallelizzare gli accessi ai dati.

Questo approccio comporta nuovi problemi di sincronizzazione tra le vare basi di dati (DataBase - DB), considerando l'eventualita´ di effettuare un'integrazione (merging) dei risultati provenienti dai differenti DB.

10. Soluzioni Integrate (Integrated Solutions):
L'integrazione del sistema di IR con altri sistemi gia´ esistenti e´ un altro punto da non trascurare. Esistono differenti strumenti (tools) che vengono utilizzati per per risolvere parte dei problemi di organizzazione delle informazioni. Certamente un'effettiva integrazione con altri tools si rende necessaria affinche´ le funzionalita´ di ricerca siano realmente disponibili all'utente in base a cio´ che desidera trovare.

1.3 Precisione e Richiamo

Una delle caratteristiche principali in una ricerca di informazioni, e´ la misura qualitativa e quantitativa dei risultati tramite i valori di Precisione e Richiamo [9], [10].

Con Precisione viene indicata la frazione dei documenti realmente di interesse rispetto a quelli recuperati dalla ricerca; mentre con Richiamo si intende la frazione dei documenti rilevanti che viene recuperata [Fig. 1].

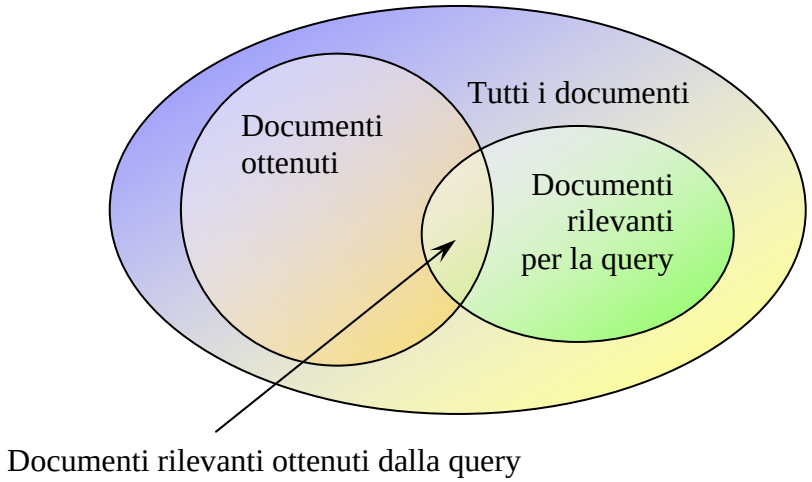


Fig. 1 - Precisione e Richiamo

Chiamando:

Precisione	→	P
Richiamo	→	R
N° Documenti rilevanti ottenuti	→	Dro
N° Documenti ottenuti	→	Do
N° Documenti rilevanti	→	Dr

Risultano:

$$P = \frac{Dro}{Do} \quad ; \quad R = \frac{Dro}{Dr}$$

La Precisione indica quanto il sistema e´ in grado di filtrare materiale inutile, mentre il Richiamo indica quanto il sistema e´ in grado di reperire informazione utile. Dove con “utile” si intende ovviamente un sottoinsieme di tutti i documenti che viene ritenuto semanticamente attinente, di volta in volta, alla query immessa dall’utente.

1.4 Metodi di ricerca: recupero dati e informazioni

I piu´ diffusi metodi per la ricerca di informazioni si basano sull'utilizzo di parole chiave immesse dall'utente che dovranno essere analizzate per poter trovare il documento che maggiormente possiede tali caratteristiche. La ricerca per parole chiave, pero´, non porta a risultati ottimali nel caso in cui l'utente cerchi informazioni e non dati, in quanto la risposta del sistema si basa rigorosamente su cio´ che l'utente specifica come parole chiave e quindi, ad esempio, una digitazione errata o l'utilizzo di parole poco attinenti a cio´ che realmente si sta cercando, porterebbe ad un risultato anche abbastanza lontano rispetto alle esigenze dell'utente stesso.

Si identificano quindi due principali categorie nei sistemi di ricerca: sistemi di recupero dati e sistemi di recupero informazioni.

1.4.1 Sistemi di recupero dati

I sistemi di recupero dati si preoccupano di dover effettuare ricerche all'interno di dati strutturati, che dovranno quindi attenersi a regole strutturali ben precise, e fornire all'utente un insieme di dati che soddisfi alle sue esigenze.

1.4.2 Sistemi di recupero delle informazioni

I sistemi di recupero informazioni, invece, effettuano ricerche all'interno di documenti che spesso non possiedono alcuna struttura, ad esempio all'interno di documenti di testo in linguaggio naturale, in cui cio´ che e´ realmente importante per l'utente non e´ tanto l'insieme dei

singoli termini, bensi´ il significato che si attribuisce agli stessi, nel testo in cui si trovano.

Affinche´ sia possibile realizzare sistemi di ricerca di informazione, e´ necessario sfruttare delle metodologie per poter estrarre dal testo particolari strutture semantiche: ciascun testo o documento verra´ rappresentato dall'insieme dei termini che maggiormente lo rappresentano, in modo da poter realizzare una struttura che consenta, tramite l'utilizzo di parole chiave ma soprattutto considerandone il loro significato, di ottenere uno o piu´ documenti da una ricerca.

Capitolo Secondo

Analisi della Semantica Latente e Decomposizione ai Valori Singolari

2.1 Introduzione

Attraverso l' Analisi della Semantica Latente (Latent Semantic Analysis - LSA) [2], [4], [6] e´ possibile rappresentare le parole tramite il loro significato analizzandone proprieta´ statistiche, come ad esempio valori di co-occorrenza dei termini nel contesto di una grande quantita´ di documenti. Si possono cosi´ identificare sottoinsiemi di parole fortemente legate fra loro dal punto di vista semantico.

L'analisi della semantica latente puo´ anche essere vista come un modello dei processi di calcolo e delle rappresentazioni che sono alla base di porzioni sostanziali dell'acquisizione e dell'utilizzo della conoscenza [5].

2.2 Scopo dell'analisi semantica dei documenti

L' Analisi della Semantica Latente e´ una tecnica statistico-matematica completamente automatica per l'estrazione di relazioni di utilizzo delle parole nelle frasi di un discorso. Tale tecnica non si basa

su alcun dizionario ma, come gia´ detto, soltanto sulla disposizione statistica delle parole nell'insieme dei documenti.

2.3 Analisi semantica nel recupero dell'informazione

L'analisi della semantica latente e´ una tecnica molto importante sfruttata proprio per ottenere una struttura semantica da un insieme di documenti di testo.

L'analisi della semantica latente e´ una teoria e un metodo per estrarre e rappresentare le parole secondo il loro significato contestuale tramite computazioni statistiche applicate ad un grande insieme di documenti [2], essa e´ quindi una tecnica per il recupero di informazioni basata unicamente sul testo, indipendentemente dalla sua struttura logica. Si afferma quindi che, in un documento di testo, l'argomento trattato non sia tanto associato ai termini realmente impiegati nel testo stesso, ma piu´ profondamente ai concetti che si utilizzano per descriverlo.

E´ possibile quindi, tramite l'analisi semantica, effettuare ricerche su documenti non piu´ basate sui singoli termini immessi dall'utente che dovranno comparire obbligatoriamente all'interno dei documenti da trovare, come in una semplice ricerca per parole chiave, ma le parole di ricerca verranno interpretate secondo il loro significato, riuscendo a recuperare risultati che, pur non contenendo le parole di partenza, figurano nell'insieme delle risposte e che non sarebbe stato possibile trovare tramite una ricerca semplicemente basata sui termini [4].

2.3.1 Concetto di Indicizzazione

Tornando ai sistemi di recupero dell'informazione e alla rappresentazione del testo mediante parole significative, esiste un particolare procedimento che vale la pena menzionare e che si rivelera´ fondamentale per l'ambito di pertinenza della tesi in esame: Il processo con cui si identificano le parole piu´ importanti che dovranno rappresentare il documento e´ noto come *Indicizzazione* (cfr. Latent Semantic Indexing - [25], [27]). L'indicizzazione e´ quindi un procedimento per estrarre, da una collezione di documenti, le parole maggiormente rappresentative per ciascun documento, che verranno successivamente identificate in strutture logiche ben precise. Tali parole sono dette "termini indice".

Un documento puo´ quindi essere rappresentato tramite un numero variabile di termini indice, in funzione ad esempio della lunghezza e del contenuto dei documenti stessi.

2.3.2 Fasi principali dell'indicizzazione

Affinche´ un testo possa essere indicizzato, e´ necessario che esso venga sottoposto a diverse fasi di elaborazione, di seguito trattate.

Un ottimo sito di riferimento e´ [26].

2.3.2.1 Analisi lessicale:

Fase preliminare di elaborazione, in cui i documenti vengono analizzati in modo da poter trascurare caratteri che non hanno importanza semantica, come ad esempio numeri e caratteri di punteggiatura.

2.3.2.2 Eliminazione dei termini senza valore semantico:

Si devono inoltre eliminare dal testo tutte quelle parole che non apportano significato al documento stesso, come parole frequenti, avverbi o altro che puo´ ovviamente variare da documento a documento. Tali parole sono chiamate stop-words.

2.3.2.3 Estrazione della radice dei termini:

Con tale procedimento, vengono eliminati prefissi e/o suffissi dalle parole, in modo da poter individuarne la radice. Questo e´ necessario ad esempio quando si vogliono considerare maschili/femminili, singolari/plurali e altre differenti forme di una stessa parola viste come un'unica entita´. Per la fase di estrazione della radice - stemming - si puo´ fare riferimento all'algoritmo di Porter [18].

2.3.2.4 Selezione dei termini indice:

Fatto quanto elencato fino ad ora, si devono scegliere i termini che dovranno successivamente rappresentare, nel modo migliore dal punto di vista semantico, il documento. Si potrebbe procedere identificando i gruppi di sostantivi che presentano un'elevata vicinanza sintattica.

Effettuati i suddetti passaggi, e´ adesso possibile rappresentare ciascun documento con un insieme di termini indice.

2.4 Utilizzo dei dati sotto forma di matrice

Per poter procedere all'analisi semantica di un insieme di documenti, e´ necessario rappresentare il testo - quindi le parole -

attraverso una struttura a matrice, le cui righe rappresentino tutte le singole parole che compongono i documenti e le colonne possono rappresentare ad esempio le parole stesse, le frasi, i paragrafi, o comunque un qualsiasi altro sottoinsieme di tutti i documenti da analizzare. Una tale rappresentazione e´ detta “Metodo dello spazio vettoriale” [13].

Il metodo dello spazio vettoriale si pone come obiettivo quello di poter rappresentare i termini dei documenti tramite vettori, affinche´ sia possibile effettuare misure di distanza tra il vettore della query e quelli dei documenti ed ottenere dei “punteggi” per identificare il documento piu´ vicino alla richiesta dell’utente.

La scelta di cio´ che caratterizzera´ le colonne della matrice varia di volta in volta, secondo le necessita´ di chi dovra´ eseguire l’algoritmo di codifica ed in funzione di cio´ che dovra´ essere la risposta del sistema alle richieste dell’utente, sia essa una frase o un documento.

2.4.1 Rappresentazione termine-documento

Una rappresentazione termine-documento e´ utile quando si vogliono produrre come risposte del sistema, interi documenti o frasi (oppure ancora qualsiasi sottoinsieme del materiale disponibile), associando a ciascuno di essi un “punteggio” che ne identifica una misura di relazione con la query immessa dall’utente.

Informazioni utili sulla rappresentazione termine-documento possono essere trovati in [14].

Data la matrice **A** termine-documento:

$$A = [a_{ij}]$$

Ciascun elemento della matrice mettera´ quindi in relazione le righe con le colonne nel seguente modo: il valore assunto dall'elemento $\mathbf{a_{ij}}$ della matrice stessa potrebbe rappresentare il valore di occorrenza del termine **i** nel documento **j**. In questo modo, su ciascuna riga della matrice - identificante il termine i-esimo - si troveranno i valori di occorrenza della parola **i** in ciascuno dei documenti - j-esima colonna - considerati.

Il valore dell'elemento $\mathbf{a_{ij}}$ potrebbe invece rappresentare qualcosa di leggermente piu´ complesso e di conseguenza piu´ preciso ed efficace. Un metodo per attribuire un valore piu´ preciso all'elemento $\mathbf{a_{ij}}$ prevede l'utilizzo di tre componenti per la pesatura del valore che esso dovra´ assumere. Tali componenti rappresentano il peso locale, quello globale ed un fattore di normalizzazione [12].

L'elemento $\mathbf{a_{ij}}$ sara´ quindi rappresentato come segue:

$$a_{ij} = l_{ij} \cdot g_i \cdot n_j$$

Dove:

$\mathbf{l_{ij}}$ → rappresenta il peso locale del termine **i** nel documento **j** ed e´ una misura di quanto il termine **i** sia importante per il documento **j**, considerando che se un termine si trovi molte volte all'interno di un documento, probabilmente esso e´ relativamente importante per il documento preso in esame.

$\mathbf{g_i}$ → rappresenta il numero di occorrenza del termine **i** in tutti i documenti: e´ il peso globale.

n_j → rappresenta un fattore di normalizzazione che tiene conto ad esempio della lunghezza del documento per poter fare in modo che termini che hanno lo stesso numero di occorrenza in documenti di lunghezza differente abbiano un peso differente: il termine relativo ad un documento piu´ corto avra´ un peso maggiore, poiche´ ha una percentuale di occorrenza maggiore in quel documento rispetto al termine relativo ad un documento piu´ lungo, sempre a parita´ di numero di occorrenza.

Per scegliere il documento maggiormente correlato alla query e´ necessario quindi calcolare il suo "punteggio" tramite semplici operazioni di prodotto tra vettori e matrici.

Ciascuna query immessa dall'utente puo´ essere considerata semplicemente come un documento, quindi puo´ essere anch'essa rappresentata da un vettore di lunghezza pari al numero di termini totali, le cui componenti sono rappresentate dal numero di occorrenza che il termine **i** ha nella stessa query. Si ottiene cosi´ il vettore colonna **q** di query, rappresentabile come segue:

$$q = [q_i]$$

Effettuando il prodotto tra **q´**, il trasposto del vettore della query e la matrice termine-documento gia´ creata, otteniamo, come illustrato di seguito, il vettore riga **s** dei punteggi relativi ai vari documenti:

$$s = q' \cdot A$$

Tale vettore avra´ dimensione pari al numero dei documenti - le colonne della matrice **A** - e la sua componente a valore maggiore in

corrispondenza del documento piu´ attinente rispetto alla query immessa dall'utente.

2.4.2 Rappresentazione termine-termine

Una rappresentazione termine-termine si potrebbe rendere necessaria quando si vuole attribuire a ciascuna parola una propria codifica vettoriale per poter realizzare ad esempio codifiche dei documenti ottenute come somma - o qualsiasi altra operazione - tra vettori relativi alle parole contenute nel documento stesso.

In una matrice di questo tipo, il valore assunto dall'elemento **i**,**j** potrebbe invece rappresentare quante volte il termine **i** ed il termine **j** - la coppia di termini **i**,**j** - occorrano contemporaneamente nelle frasi dei documenti.

Ottenuta quindi la matrice delle occorrenze delle coppie dei termini nei documenti, si sfrutta la tecnica di scomposizione ai valori singolari (cfr. §2.5 e segg.) per ottenere le codifiche vettoriali delle singole parole, presenti nella matrice $\mathbf{U_k}$. Le codifiche vettoriali dei documenti si possono ottenere, come precedentemente accennato, semplicemente come somma vettoriale delle codifiche dei singoli termini componenti il documento, oppure direttamente dalle righe della matrice $\mathbf{V_k}$. Lo stesso si puo´ fare per la query immessa dall'utente, anche se si possono effettuare metodi di codifica differente per la query rispetto ai documenti.

A questo punto si dispone dei vettori rappresentativi sia della query che di tutti i documenti. Per poter scegliere il documento da mostrare all'utente in risposta alle proprie richieste, si devono effettuare misure di distanza nello spazio vettoriale tra query e documenti, come ad esempio misure dell'angolo tra i due vettori.

Una misura interessante potrebbe essere il rapporto tra la parte ortogonale - seno - e la parte parallela - coseno - tra il vettore **q** della query e tutti i vettori **d** dei documenti, come illustrato di seguito [Fig. 2].

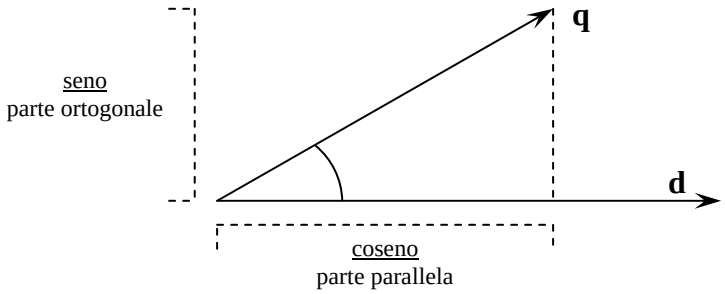


Fig. 2 - Rappresentazione di parte ortogonale e parallela

$$dist = \frac{parte \quad ortogonale}{parte \quad parallela}$$

dove il coseno - parte parallela - e´ pari al rapporto tra il prodotto scalare dei vettori ed il prodotto delle loro norme (la norma corrisponde al modulo, o lunghezza, del vettore):

$$\cos(q,d) = \frac{q \cdot d}{\|q\|_2 \cdot \|d\|_2}$$

mentre il seno - parte ortogonale - si puo´ ricavare facilmente come sottrazione vettoriale tra il vettore della query ed il coseno precedentemente ottenuto.

Come si puo´ gia´ intuire dalla misura di distanza, questo valore e´ tanto piu´ piccolo quanto il vettore della query va sempre piu´ andandosi a sovrapporre ai vettori dei documenti. La distanza risulta, al limite, pari

a zero nel caso in cui il vettore della query e quello del documento coincidano.

2.5 Decomposizione ai valori singolari

Dopo aver realizzato la matrice in una delle differenti modalita´ possibili, si rende necessario utilizzare un metodo per poter diminuire la dimensione delle matrici su cui lavorare, ma soprattutto per ottenere altre matrici che conterranno codifiche importanti da utilizzare successivamente. Si puo´ ben capire che, in funzione del numero di termini e delle frasi o comunque dei documenti utilizzati come base per la misura delle occorrenze dei termini stessi, si possono ottenere matrici di considerevoli dimensioni, le quali non permettono agevolmente di effettuare calcoli abbastanza velocemente, ma soprattutto potrebbero portare a risultati non corretti sulle rappresentazioni vettoriali dei termini. Inoltre, come gia´ accennato, questo metodo ci consentira´ di ottenere la rappresentazione vettoriale dei termini, necessaria per poter codificare successivamente i documenti.

2.5.1 Matrici ottenute dalla scomposizione

Una tecnica molto adatta allo scopo e´ la scomposizione ai valori singolari (Singular Value Decomposition - SVD) [16], [17] della matrice precedentemente ottenuta. La matrice verra´ cosi´ scomposta in tre sottomatrici, il cui prodotto riportera´ alla matrice orginaria [Fig. 3].

La figura 3 mostra le relazioni, in termini di dimensione, tra la matrice **A** e le matrici risultanti dalla scomposizione.

$$A = U \cdot \Sigma \cdot V'$$

Data la matrice **A**, di dimensioni **m*n** e rango **r**, le matrici risultanti dalla scomposizione sono rispettivamente:

U di dimensioni **m*r**

Σ diagonale di dimensioni **r*r**

V di dimensioni **n*r**

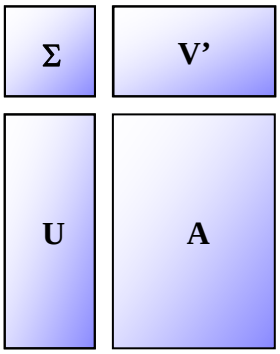


Fig. 3 - Scomposizione SVD

Una caratteristica particolare delle matrici **U** e **V** e´ che, se moltiplicate per la propria trasposta, producono la matrice Identita´ a dimensione **n*n**:

$$UU' = V'V = I_n$$

Mentre per quanto riguarda la matrice **Σ**, come precedentemente indicato, e´ una matrice quadrata diagonale, i cui elementi (non negativi) - detti valori singolari - sono ordinati in maniera decrescente lungo la diagonale, cosi´ come esposto di seguito:

$$\Sigma = diag(\sigma_1,...,\sigma_n)$$

$$\text{Con } \sigma_i > 0 \text{ per } 1 \le i \le r \text{ e } \sigma_i = 0 \text{ per } 1 \ge r+1$$

Questa e´ una caratteristica importante per procedere ad una successiva approssimazione delle matrici, illustrata in §2.5.2.

Le colonne della matrice **U** rappresentano gli autovettori normalizzati relativi agli autovalori non nulli della matrice **AA'** e le

colonne della matrice **V** rappresentano gli autovettori relativi agli autovalori non nulli della matrice **A'A**.

Gli elementi della matrice **Σ** invece rappresentano le radici quadrate non negative, disposte in ordine decrescente lungo la diagonale, degli autovalori di **AA'** e costituiscono i valori singolari della stessa matrice **A**.

Ottenute le tre matrici dalla scomposizione, possono essere utilizzate per effettuare confronti, tra termini, tra documenti e anche tra termini e documenti.

Il confronto tra i termini puo´ avvenire tramite prodotto scalare delle righe della matrice. La matrice dei prodotti scalari e´ la matrice quadrata e simmetrica **AA'**, che puo´ anche essere espressa come segue:

$$AA' = U \cdot \Sigma^2 \cdot U'$$

Per il confronto tra i documenti, invece, devono essere messe in relazione le colonne della matrice di partenza. I prodotti scalari delle colonne possono essere riassunti nella matrice **A'A**, anch'essa quadrata e simmetrica, che puo´ essere espressa come:

$$A'A = V \cdot \Sigma^2 \cdot V'$$

Il confronto tra termini e documenti tiene conto del fatto che la matrice A puo´ essere ottenuta, come gia´ detto come il prodotto delle tre matrici:

$$A = U \cdot \Sigma \cdot V'$$

Quindi, il generico elemento $\mathbf{a_{ij}}$ della matrice **A** puo´ essere ottenuto come il prodotto della riga **i** della matrice $U\sqrt{\Sigma}$ con la riga **j** della matrice $V\sqrt{\Sigma}$.

Come si vedra´ nel successivo paragrafo, le stesse operazioni viste finora possono essere eseguite con le matrici approssimate.

2.5.2 Approssimazione delle matrici

Una attenzione molto particolare deve essere rivolta al parametro della scomposizione - **k** -, ovvero il valore della dimensione minore (il "lato piu´ corto" delle matrici) che caratterizzera´ le matrici risultanti dalla scomposizione stessa (la matrice **Σ** sara´ una matrice quadrata il cui lato avra´ appunto dimensione **k**); se tale parametro e´ assente, le matrici non verranno approssimate e la dimensione minore delle stesse sara´ il rango della matrice **A** di partenza.

Questo parametro rappresenta appunto il grado di approssimazione delle matrici: tanto maggiore sara´ la dimensione scelta - ovvero tanto piu´ **k** si avvicinera´ ad **r** -, quanto piu´ accurata sara´ la ricostruzione della matrice originaria.

Ovviamente questo comporta sia vantaggi che svantaggi. Uno dei vantaggi e´ sicuramente la mancanza di perdita di informazione sulla matrice risultante rispetto a quella originaria, mentre uno svantaggio importante per il nostro caso e´ che si conserverebbe senz'altro rumore aggiunto che porterebbe ad una cattiva e quindi non corretta interpretazione del significato dei termini da codificare. Il troncamento della dimensione puo´ essere effettuato anche perche´, come gia´ detto, i termini della matrice **Σ** - che e´ una matrice diagonale - sono disposti in

ordine decrescente ed il troncamento ha effetto a partire dai termini meno significativi [Fig. 4].

La figura successiva mostra le realzioni che intercorrono tra la matrice $\mathbf{A_k}$, - approssimazione della matrice **A** - e le tre matrici approssimate, risultanti dalla scomposizione.

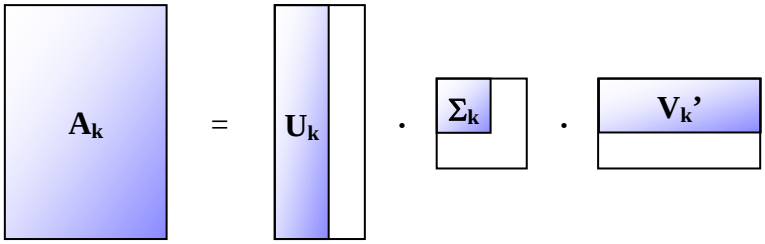
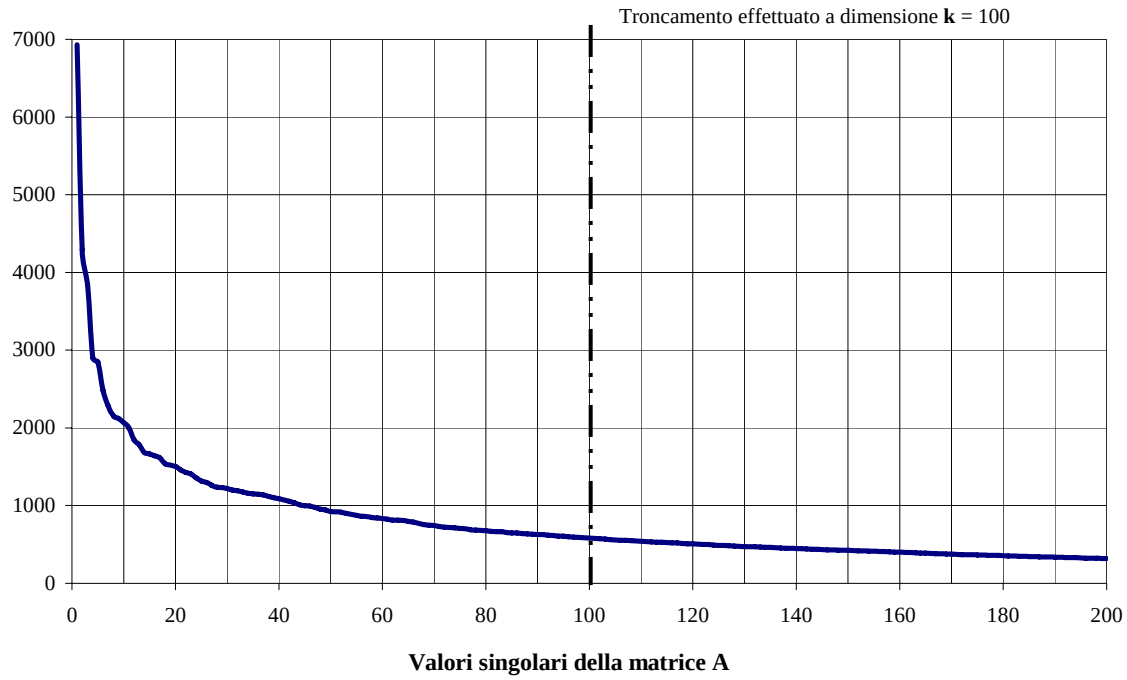


Fig. 4 - Approssimazione a dimensione k delle matrici

E´ dimostrato che un troncamento sulle dimensioni delle matrici porta a sensibili miglioramenti nella rappresentazione delle parole [2], inoltre la matrice $\mathbf{A_k}$ rappresenta la miglior approssimazione della matrice **A** nel senso dei minimi quadrati [11]. Effettuando una approssimazione sui valori delle matrici, e´ possibile eliminare il rumore residuo ed ottenere risultati migliori nella codifica dei termini.

La scelta della dimensione **k** del troncamento effettuato sulle matrici prodotte dalla scomposizione SVD non e´ fissata a priori e il valore di **k** e´ un valore del tutto empirico, ovvero esso puo´ variare da un particolare caso ad un altro, in funzione del tipo e del contenuto dei documenti analizzati. Su certi documenti potrebbe essere sufficiente una dimensione relativamente piccola, mentre su altri la stessa dimensione porterebbe a risultati palesemente errati. Valori tipici per **k** variano, in genere, tra cento e trecento [19].

Di seguito e´ riportato il grafico dei primi duecento valori singolari della matrice **A**, di cui sono stati utilizzati i primi cento piu´ significativi:



Valori singolari della matrice A

2.6 Utilizzo delle matrici ottenute dalla scomposizione

Tornando alle tre matrici ottenute dalla scomposizione ai valori singolari, possiamo notare che, utilizzando la matrice originaria **A** di dimensioni **m*n** ed effettuando un troncamento SVD a dimensione **k**, si ottengono le seguenti tre matrici: la matrice $\mathbf{U_k}$ di dimensioni **m*k**, la matrice $\mathbf{\Sigma_k}$, matrice quadrata diagonale di dimensioni **k*k**, ed infine la matrice $\mathbf{V_k}$ di dimensioni **n*k**.

La matrice $\mathbf{A_k}$ [Fig. 4] (approssimazione a dimensione **k** della matrice **A**) puo´ essere ottenuta come semplice prodotto delle tre matrici come di seguito indicato:

$$A_k = U_k \cdot \Sigma_k \cdot V_k'$$

Come precedentemente esposto, la matrice $\mathbf{U_k}$ ha un numero di righe pari a quelle della matrice di partenza, a loro volta uguali in numero ai termini dei documenti. La matrice $\mathbf{U_k}$ conterra´ quindi le codifiche vettoriali a dimensione **k** dei termini stessi. Tali codifiche verranno utilizzate successivamente per caratterizzare le frasi dei documenti o i documenti stessi. A questo punto, ottenute le codifiche vettoriali delle parole, e´ possibile ricavare quelle delle singole frasi o comunque dei documenti.

La matrice $\mathbf{V_k}$ invece, avra´ un numero di righe pari al numero dei documenti analizzati. Da questo si puo´ intuire che le righe della matrice $\mathbf{V_k}$ potranno essere utilizzate per recuperare una codifica vettoriale per ciascuno dei documenti analizzati.

Poiche´ la matrice $\mathbf{V_k}$ e´ il risultato dell'applicazione della tecnica SVD alla matrice originaria di partenza; le codifiche dei documenti in essa contenute non corrispondono alle codifiche vettoriali dei documenti ottenute invece come somma dei vettori rappresentativi dei termini che costituiscono i documenti stessi.

Per quanto riguarda l'elaborazione della query immessa dall'utente, un altro modo potrebbe essere quello di trasformarla sotto forma di pseudo-documento grazie alla relazione successiva:

$$\hat{q} = q' \cdot U_k \cdot \Sigma_k^{-1}$$

In questo modo si rende possibile il confronto tra la query immessa dall'utente e tutti i micro-documenti dell'insieme, le cui codifiche

vettoriali si trovano, come gia´ detto precedentemente, nelle righe della matrice $\mathbf{V_k}$.

2.6.1 Codifica dei documenti e della richiesta dell'utente

In funzione di cio´ che dovra´ essere infine presentato all'utente - come ad esempio singole frasi o interi documenti - in risposta alle sue richieste, esistono diversi metodi per procedere con la codifica dei documenti al fine di reperire quello maggiormente legato alla query.

Dopo aver ottenuto le codifiche vettoriali di tutti i termini - presenti, come detto, nella matrice $\mathbf{U_k}$ - la codifica dei documenti si puo´ ottenere realizzando una combinazione lineare delle rappresentazioni vettoriali dei termini presenti nel documento stesso, sia esso una frase o altro. Quindi, un possibile metodo, probabilmente il piu´ semplice, per ottenere un vettore rappresentativo di un documento e´ sostituire a ciascuna parola dello stesso, il suo vettore rappresentativo ed ottenere il vettore risultante come normalizzazione della somma delle singole rappresentazioni vettoriali dei termini.

E´ inoltre possibile ottenere altre codifiche vettoriali dei documenti sfruttando le righe della matrice $\mathbf{V_k}$. Con questo metodo, pero´, i confronti tra i vettori dei documenti e quello della query, dovranno essere effettuati dopo aver ottenuto il vettore di query $\hat{q}$ come esposto precedentemente, sfruttando le matrici $\mathbf{U}$ e $\mathbf{\Sigma^{-1}}$ ottenute dalla scomposizione ai valori singolari.

L'utente fara´ fronte alle proprie esigenze di conoscenza ponendo una richiesta - la query - che potra´ essere codificata o meno allo stesso modo dei documenti.

Sono disponibili adesso le codifiche vettoriali sia della query che di tutti i documenti.

2.7 Scelta del documento piu´ attinente

Per procedere al recupero del corretto documento, ovvero del documento semanticamente piu´ vicino alla richiesta dell'utente, si possono calcolare le distanze tra il vettore della query ed i singoli vettori di tutti i documenti, al fine di poter identificare quello geometricamente piu´ vicino al vettore della query.

Alcune metodologie tengono conto dell'angolo compreso tra il vettore della query e quello del documento, altre considerano la quantita´ in comune tra gli angoli n-dimensionali costituiti dai vettori rappresentativi di tutte le parole presi singolarmente rispettivamente della query e del documento, altre ancora si basano su misure piu´ o meno complesse e/o precise.

La metodologia utilizzata in questa tesi per la misura della distanza e´ illustrata nel capitolo terzo, paragrafo 5 (cfr. §3.5) e utilizza misure delle parti ortogonali e parallele dei vettori di cui si deve ottenere una misura di distanza.

Capitolo Terzo

Soluzione proposta

3.1 Raccolta dei documenti

Come gia´ velocemente trattato nell'introduzione, questa tesi si occupa di fornire una conoscenza sui documenti del repository del parlamento europeo, e´ stato quindi necessario effettuare la raccolta preliminare dei documenti in lingua inglese presenti nel sito del parlamento europeo [1], per poter dotare l'applicazione sia di uno spazio vettoriale con cui rappresentare le parole, sia di una conoscenza specifica sull'argomento.

Di seguito e´ riportato uno schema illustrativo dei passaggi principali seguiti nello sviluppo di questa tesi che verranno successivamente trattate con maggior dettaglio [Fig. 5].

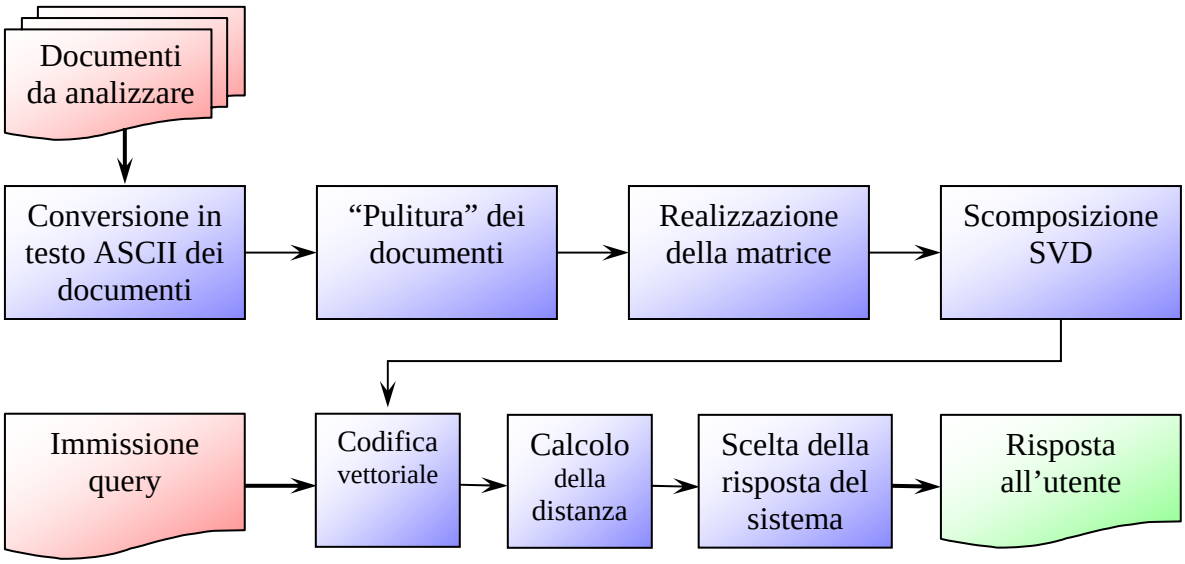


Fig. 5 - Passaggi principali dello sviluppo della tesi

3.2 Pre-elaborazione e conversione dei documenti

3.2.1 Pulitura del testo

Affinche´ i documenti reperiti possano essere utilizzati, essi devono passare per una fase preliminare di elaborazione. Dopo aver trasformato tutti i documenti - originariamente in formato pdf - in files costituiti da puro testo ascii, e´ stato necessario effettuare alcune successive elaborazioni per poter rappresentare i documenti nello spazio semantico.

Sono stati eliminati segni di punteggiatura, caratteri speciali e parole che non portano alcun significato aggiuntivo alle frasi, tra queste parole si possono ricordare articoli, congiunzioni, proposizioni, avverbi e aggettivi frequenti, etc. Dette parole vengono chiamate stop-words.

Inoltre e´ stata considerata soltanto la rappresentazione in soli caratteri minuscoli dei termini, quindi, per ciascuna parola - che essa sia presente nel testo in caratteri minuscoli e/o maiuscoli - le verra´ associata sempre la medesima codifica vettoriale.

La fase preliminare di processamento dei documenti e´ stata effettuata in ambiente Linux (come del resto tutto lo sviluppo dell'applicazione) tramite classi Java e, soprattutto in questa fase, grazie a script di shell (bash) per la gestione di flussi di caratteri, molto veloci e semplici da gestire.

A questo punto, dall'insieme di tutti i documenti in lingua inglese presenti nel sito del parlamento europeo, oltre a mantenere ciascun file di testo relativo ai documenti analizzati, e´ stato ottenuto un unico file comprensivo delle frasi di tutti i documenti, frasi ordinate riga dopo riga, che costituiranno i micro-documenti su cui effettivamente agire. Per realizzare lo spazio di rappresentazione delle parole, verranno utilizzati

tali micro-documenti, costituiti appunto dalle singole frasi prelevate dall'insieme di tutti i documenti.

3.2.2 Elenco dei termini presenti nei documenti

Si prepara adesso l'elenco di tutti i termini presenti nei micro-documenti con il relativo numero di occorrenza, per poter determinare quelli piu´ frequenti, e quindi, piu´ importanti rispetto ad altri termini. Il file risultante sara´ costituito da due colonne, separate da un carattere di tabulazione, in cui la riga i della prima colonna conterra´ il numero di occorrenza della i-esima parola presente nella seconda colonna, occorrenza relativa all'insieme di tutti i documenti. L'insieme dei termini presenti in questo file sara´ costituito da tutti i termini - ovviamente presi una sola volta, giacche´ si tratta appunto di un insieme - presenti nei documenti.

A ciascuno di questi termini verra´ successivamente associata la propria codifica vettoriale grazie all'elaborazione della matrice termine-termine che conterra´ dati sulle proprieta´ statistiche delle singole parole nei documenti analizzati.

3.3 Realizzazione matrice e scomposizione SVD

Come detto nel paragrafo 2.3, bisogna ricondurre i documenti ad una rappresentazione matriciale su cui poter effettuare calcoli statistico-matematici atti ad ottenere la codifica vettoriale dei termini. Nella matrice termine-termine realizzata, l'elemento **i**,**j** e´ dato dal valore di occorrenza della coppia "termine **i**, termine **j**" nei micro-documenti. Si puo´ verificare quindi che alcune coppie possano essere piu´ frequenti di

altre o molto piu´ frequentemente che altre coppie non si trovino in alcun micro-documento. Questo a causa della presenza dei piu´ disparati termini presenti in elenco, i quali non hanno probabilita´ di trovarsi assieme nello stesso micro-documento.

Una volta realizzata la matrice termine-termine, e´ possibile scomporla in tre matrici sfruttando l'approssimazione a dimensione **k** (cfr. §2.5 - Decomposizione ai Valori Singolari - SVD) cosi´ da poter lavorare su matrici di dimensioni molto ridotte rispetto a quelle della matrice termine-termine originaria, ma soprattutto in modo da ottenere un'approssimazione a dimensione **k** dei valori originari che consente di eliminare il rumore di fondo per la codifica delle parole. Con rumore di fondo si intende tutto cio´ che potrebbe ad esempio riguardare imperfezioni delle frasi, casi particolari di occorrenza delle parole, etc. che porterebbe ad una errata codifica dei termini.

3.4 Codifica vettoriale

3.4.1 Codifica delle singole parole

I vettori riga della matrice $\mathbf{U_k}$ ottenuta dalla scomposizione ai valori singolari sono linearmente indipendenti e quindi potranno rappresentare univocamente le codifiche vettoriali a dimensione **k** dei termini (tutte le parole presenti nei documenti) ad essi associati. Le codifiche ottenute verranno successivamente utilizzate affinche´ si possa trovare una relazione tra le parole stesse, in modo da identificare parole che sono reciprocamente vicine nello spazio vettoriale, potendone quindi misurare una distanza in tale spazio di codifica.

Le codifiche delle parole possono essere impiegate per ottenere una rappresentazione vettoriale delle frasi - i micro-documenti - come normalizzazione della somma delle singole rappresentazioni vettoriali delle parole componenti le frasi stesse (cfr. §2.6.1).

3.4.2 Codifica dei micro-documenti

Ottenute le codifiche vettoriali di tutte le parole, si puo´ adesso procedere alla rappresentazione vettoriale dei micro-documenti. Poiche´ essi sono costituiti dai termini gia´ codificati, si possono utilizzare le codifiche delle singole parole per ottenere una codifica univoca di ciascun micro-documento.

I documenti da codificare adesso sono i micro-documenti completi di stop-words e punteggiatura (come detto precedentemente, le stop-words sono termini di uso comune, quali aggettivi, avverbi o altro, che non apportano un contenuto informativo aggiuntivo alla frase o al documento in cui si trovano), poiche´ essi verranno utilizzati come risposta alle richieste dell'utente e devono essergli presentate in una forma grammaticale corretta. Il fatto che si utilizzino micro-documenti "completi" non rappresenta un problema, poiche´ i caratteri di punteggiatura - e comunque i caratteri che non siano lettere - verranno ignorati durante il processo di codifica: a ciascuna stop-word non corrispondera´ infatti alcuna rappresentazione vettoriale poiche´ non e´ presente nell'elenco generale dei termini. Esse verranno semplicemente ignorate e cio´ corrisponde al recupero di una rappresentazione vettoriale nulla del termine.

Per tutte le parole incluse nel micro-documento, si considera la relativa codifica vettoriale e si sommano, componente per componente, i vettori di ciascuna parola che compone il micro-documento. Di tali vettori somma verra´ utilizzata successivamente la rispettiva

rappresentazione normalizzata per poter effettuare misure di similarita´ tra il vettore della query e quelli dei micro-documenti.

Un altro modo per poter ottenere la codifica dei documenti, e´ quello di sfruttare i vettori riga gia´ presenti nella matrice $\mathbf{V_k}$, come risultato della precedente scomposizione SVD. Per poter utilizzare tali codifiche e´ necessario ottenere il vettore di query come esposto in §2.6.

A questo punto si e´ ottenuto un file contenente su ciascuna riga la codifica vettoriale di tutti i micro-documenti. A ciascuna riga del file dei micro-documenti corrisponde la relativa riga del file delle codifiche vettoriali dei micro-documenti stessi, in modo da mantenere una relazione tra gli indici di riga del file dei micro-documenti e quello delle corrispondenti codifiche vettoriali, tale accorgimento si rende necessario in quanto, effettuando la misura delle distanze tra i vettori (cfr. §2.7), viene ritornato l'indice della riga in cui e´ presente il vettore del micro-documento a minor distanza rispetto al vettore della query ed e´ possibile quindi ricavare il micro-documento originale dal proprio file.

3.4.3 Codifica della richiesta dell'utente

Una volta ottenute con uno dei metodi precedentemente discussi le codifiche vettoriali dei micro-documenti e dopo averle salvate su file, si puo´ procedere all'analisi della query immessa dall'utente.

Essa puo´ essere considerata a sua volta un micro-documento per cui su di essa si possono - ma non necessariamente - applicare le stesse tecniche utilizzate per codificare i micro-documenti; quindi per ciascuna parola appartenente alla stringa della query, se ne potrebbe considerare il rispettivo vettore rappresentativo; la codifica finale per l'intera query e´ data quindi dalla normalizzazione della somma di tali vettori.

Se si vuole invece utilizzare una codifica che ricorra alla matrice $\mathbf{V_k}$ per le codifiche vettoriali dei documenti, la richiesta dell'utente deve essere codificata a sua volta con il metodo descritto in §2.6.

Adesso si dispone anche del file contenente la codifica vettoriale della query.

3.5 Calcolo della distanza tra la richiesta dell'utente e i micro-documenti

Avendo a disposizione il file contenente la codifica vettoriale della query ed il file contenente le codifiche vettoriali dei micro-documenti, si puo´ effettuare una misura della distanza tra il vettore della query e i vettori di ciascun micro-documento per poter trovare quello semanticamente piu´ correlato - quindi vicino - alla query immessa, di volta in volta, dall'utente.

Come criterio di misura della distanza tra i vettori e´ stato adottato un metodo che tiene conto del rapporto tra la parte ortogonale e quella parallela di un vettore (quello relativo alla query) rispetto all'altro (il vettore che codifica il micro-documento).

Poiche´ la parte ortogonale e´ una misura di quanto differiscono i due vettori, mentre la parte parallela rappresenta invece quanto tali vettori siano simili: tanto piu´ il rapporto tra parte ortogonale e parte parallela e´ prossimo allo zero, quanto piu´ simili sono i due vettori.

Il metodo utilizzato puo´ essere sintetizzato dalla figura, per completezza, di seguito riportata (cfr. §2.4.2) [Fig. 6].

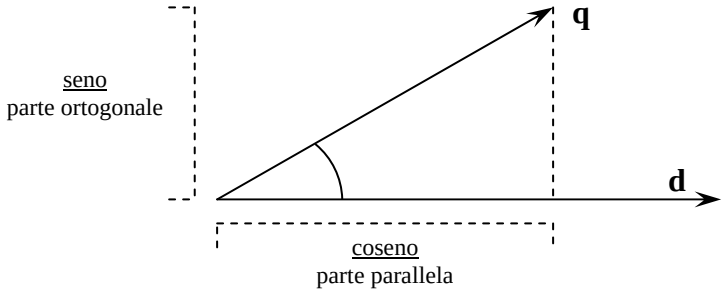


Fig. 6 - Rappresentazione di parte ortogonale e parallela

3.6 Scelta della risposta

La procedura di calcolo delle distanze tra il vettore di query e quelli dei micro-documenti, produce un file in cui sono memorizzate appunto tali distanze. Il file delle distanze verra´ quindi analizzato per trovare la posizione in cui si trova il valore di distanza minore: la posizione nel file - indice - identifichera´ il micro-documento da mostrare all'utente in risposta alle sue richieste.

Adesso che e´ stato possibile ottenere il micro-documento che ha maggior attinenza semantica con la query, questo viene visualizzato all'utente tramite l'interfaccia grafica dell'applicazione.

L'utente puo´ quindi effettuare un'altra richiesta al sistema che provvedera´ ad eseguire i passi precedentemente esposti al fine di poter trovare nell'insieme dei documenti, quello semanticamente piu´ legato alla nuova query immessa.

Capitolo Quarto

Caratteristiche del sistema e sviluppi futuri

4.1 Documenti necessari per ottenere uno spazio semantico

I sistemi che utilizzano la tecnica LSA al fine di ottenere uno spazio semantico dei termini, necessitano di avere a disposizione una ragguardevole mole di documenti; come gia´ detto, questo lavoro di tesi si basa sui documenti in lingua inglese presenti nel repository del parlamento europeo [1]; i documenti di testo utilizzati sono oltre un migliaio.

Su tali documenti dovranno essere eseguiti algoritmi statistico-matematici per poter individuare le relazioni semantiche presenti tra i termini che li compongono: maggiore e´ il numero di documenti su cui operare, maggiore sara´ la precisione delle relazioni semantiche tra i termini, dove con *precisione* si potrebbe intendere la capacita´ del sistema di identificare sinonimie, polisemie, diverse forme di genere e numero dei termini, etc. producendo risultati utili anche nei casi in cui l'utente non fornisca una query particolarmente pertinente dal punto di vista terminologico a cio´ che in realta´ si vorrebbe ottenere dalla ricerca.

4.2 Analisi della richiesta dell'utente

Un aspetto molto importante per i sistemi di recupero informazioni e sanz'altro quello relativo all'analisi della query immessa dall'utente.

Nel caso in esame non e´ stato necessario effettuare alcuna fase preliminare di elaborazione (cfr §2.6.1 e segg.) della query, poiche´ la classe java che si occupa di associare a ciascuna query (considerata a sua volta come un micro-documento) una codifica vettoriale, provvede a recuperare le codifiche della sola forma minuscola dei termini in essa presenti, escludendo cosi´ automaticamente qualsiasi altro carattere diverso dalle lettere minuscole. In realta´, i caratteri che non sono nell'insieme **[a-z]**, vengono utilizzati come separatori per individuare i singoli termini della query: la query viene prima trasformata in minuscolo e successivamente viene analizzata la nuova stringa in cerca di caratteri diversi dalle lettere minuscole, che costituiranno, come gia´ detto, i separatori per identificare i termini, a cui successivamente verra´ associato il rispettivo vettore per poter calcolare quello risultante che identifichera´ la query.

Nel caso in cui si voglia ottenere una codifica vettoriale della query e´ possibile sfruttare le codifiche dei singoli termini che la compongono. La codifica risultante sara´ data dalla somma delle codifiche vettoriali dei termini della query.

4.2.1 Rappresentazione della richiesta dell'utente per il caso termine-documento

Nel caso in cui invece si voglia utilizzare una rappresentazione termine-documento (cfr. §2.4.1) si possono utilizzare le codifiche vettoriali dei documenti gia´ presenti nelle righe della matrice $\mathbf{V_k}$.

Per utilizzare tali codifiche, pero´ e´ necessario codificare la query in modo differente. Si realizza un vettore con un numero di elementi pari al numero di termini totali dei documenti, a ciascun termine corrispondera´ quindi una posizione all'interno di tale vettore. Il valore che assumera´ l'elemento **i** del vettore sara´ in funzione di quante volte il termine i-esimo e´ presente all'interno della stringa di query. Ottenuto questo primo vettore di occorrenze, si rende necessario proiettarlo nello stesso spazio vettoriale delle codifiche dei documenti, questo e´ possibile tramite il prodotto con la matrice $\mathbf{U_k \cdot \Sigma^{-1}}$ che produrra´ un vettore della stessa dimensione **k** del troncamento SVD.

Tale vettore potra´ adesso essere confrontato con i vettori contenuti nella matrice $\mathbf{V_k}$ utilizzando le misure di distanza precedentemente trattate.

4.3 Funzionalita´ del sistema

Il sistema in esame prevede tre principali modalita´ di funzionamento:

1. Interazione semplice con l'utente su conoscenze generali.

2. Incremento della base di conoscenza con nuove frasi immesse dall'utente.

3. Recupero di documenti del Parlamento Europeo.

Di seguito verranno descritte le varie funzionalita´ con maggior dettaglio.

4.3.1 Interazione semplice

Per la funzionalita´ di interazione semplice tra utente e software, si e´ deciso di utilizzare alcune risposte del chat-bot ALICE [21] per poter simulare un semplice dialogo naturale.

Le singole frasi di risposta sono state codificate come somma dei vettori rappresentativi delle parole componenti le risposte stesse; tali vettori derivano da una scomposizione ai valori singolari della matrice delle occorrenze dei termini in tutte le possibili frasi di risposta.

La domanda dell'utente viene codificata allo stesso modo. Ottenuto il vettore rappresentativo della domanda, lo si puo´ confrontare con i vettori delle possibili risposte gia´ calcolati precedentemente.

La risposta che possiede un valore di somiglianza piu´ elevato viene infine mostrata all'utente.

4.3.2 Incremento della conoscenza

Ad una conoscenza generale potrebbe essere preferibile una conoscenza particolareggiata su un certo argomento.

Questa funzionalita´ permette all'utente di incrementare la base di conoscenza dell'applicazione semplicemente inserendo la possibile nuova

risposta nel campo di immissione del sistema, che verra´ quindi codificata come somma dei vettori rappresentativi dei suoi termini.

La frase e la sua codifica vettoriale verranno rispettivamente inseriti in coda ai file delle possibili risposte e delle codifiche vettoriali delle risposte.

4.3.3 Recupero documenti

Per il recupero di documenti invece il processo di codifica della richiesta dell'utente e´ diverso.

Dalla query si ottiene un vettore con un numero di elementi pari al numero dei possibili termini gia´ codificati (righe della matrice $\mathbf{U_k}$ ottenuta tramite scomposizione ai valori singolari della matrice delle occorrenze); il valore assunto dagli elementi del vettore e´ rappresentato dal numero di occorrenza del termine i-esimo nella query stessa.

Adesso si deve poter ottenere un vettore paragonabile ai vettori rappresentativi dei documenti (presenti nelle righe della matrice $\mathbf{V_k}$). Per fare cio´ e´ necessario effettuare opportuni prodotti tra le matrici disponibili per trasformare il vettore di query in un vettore con un numero di componenti pari al valore di approssimazione della scomposizione ai valori singolari, in questo caso, cento.

L'operazione che permette tale trasformazione e´ la seguente:

$$\hat{q} = q' \cdot U_k \cdot \Sigma_k^{-1}$$

Ottenuto il nuovo vettore della query, e´ possibile confrontarlo con i vettori rappresentativi dei documenti. Il confronto dei vettori risulta in un indice, indicante la posizione del vettore di risposta all'interno della

matrice $\mathbf{V_k}$. Il sistema provvedera´ adesso a visualizzare all'utente il relativo documento.

4.4 Confronto con i Chat-bot tradizionali

4.4.1 Introduzione ai Chat-bot

Un chat-bot e´, in generale, un agente software in grado di elaborare una richiesta dell'utente e di interagire con lui fornendogli una risposta. Da questa definizione, seppur molto generale, si puo´ comunque capire che un chat-bot e´ un interfaccia tra l'uomo e la macchina.

I chat-bot piu´ diffusi sono dotati di interfacce grafiche e/o testuali, con cui l'utente puo´ *dialogare* inserendo la sua richiesta (spesso una richiesta in forma di testo, anche se esistono chat-bot con supporto audio e/o video) ed ottenendo infine una risposta. Si possono trovare chat-bot specializzati nel dialogo naturale, chat-bot con una conoscenza specifica su un particolare argomento o, piu´ semplicemente, chat-bot che fungono da motori di ricerca sul web che, essendo in continua espansione, necessita di conseguenza, di strumenti per poter agevolare l'utente nelle sue ricerche.

Altre e piu´ numerose ed interessanti informazioni sui chat-bot si possono trovare in [21], [22], [23].

4.4.2 Funzionamento dei Chat-bot

Il funzionamento dei chat-bot tradizionali si basa sull'analisi dei singoli termini - parole chiave - immessi dall'utente, individuandone strutture sintattiche piu´ o meno semplici. Tutto questo e´ possibile

grazie all'utilizzo di un linguaggio specifico, l'AIML (Artificial Intelligence Mark-up Language) [20], basato sul probabilmente piu´ noto XML (eXtensible Mark-up Language) [24].

4.4.3 Organizzazione della conoscenza nei Chat-bot

L'unita´ fondamentale di conoscenza per un chat-bot e´ la *categoria*, costituita da due parti: la prima - pattern - relativa alla domanda e la seconda - template -relativa alla risposta.

La risposta viene data all'utente in base al riconoscimento di tali pattern, procedura detta pattern-matching. Affinche´ sia possibile rispondere all'utente, e´ necessario che tutte le possibili domande siano contemplate nella sezione "pattern" della categoria. Se invece la richiesta dell'utente non corrisponde a nessuna categoria, viene prodotta una risposta generica che avrebbe il compito di stimolare l'utente ad immettere richieste piu´ precise oppure semplicemente a suggerirgli un nuovo argomento da discutere, proprio come in un dialogo umano.

Tornando al linguaggio AIML, e´ possibile trovare in esso tre tipi principali di categorie:

1. Categorie atomiche:
   Sono le categorie piu´ semplici, in cui a ciascun pattern corrisponde un template costituito spesso semplicemente da una frase del linguaggio naturale.

2. Categorie predefinite:
   Consentono al chat-bot di poter rispondere anche a richieste in cui il pattern e´ soltanto in parte verificato.

3. Categorie ricorsive: Consentono al chat-bot di risolvere i principali problemi verificabili nell'analisi del testo, permettendo di implementare forme di *sinonimia*, *riduzione simbolica*, *suddivisione di pattern*, *correzione grammaticale*, etc.

   Relativamente alla sinonimia, viene fatto corrispondere lo stesso template di risposta a diversi pattern per la domanda, in modo che a differenti domande, contenenti pero´ sinonimi, venga associata la stessa risposta.

   Per la riduzione simbolica, vengono ridotte forme grammaticali complesse in forme piu´ semplici, tutte riconducibili alla medesima coppia pattern-template della categoria.

   Per la suddivisione di pattern, viene suddivisa la richiesta originaria dell'utente in piu´ sottorichieste corrispondenti a categorie distinte.

   Infine la correzione grammaticale permette al chat-bot di rispondere anche a richieste grammaticalmente errate dell'utente potendo individuare il pattern adatto. Per far cio´ si rende necessario pero´ prevedere tutti i casi in cui l'utente potrebbe incorrere in errori di digitazioni o simili.

4.5 Differenze tra il sistema in esame ed i Chat-bot tradizionali

Come probabilmente il paragrafo §4.3 ha lasciato intuire, ci sono profonde limitazioni all'utilizzo dei chat-bot tradizionali come strumento principale per la simulazione del dialogo umano; essi si basano infatti su strutture abbastanza rigide e definite a priori, senza che possano essere modificate.

A causa di cio´ e´ impossibile pensare che con i sistemi attuali, basati su chat-bot pattern-template, si possano ottenere ottimi risultati in tale direzione.

Le applicazioni che sfruttano l'analisi della semantica letente hanno maggiori possibilita´ nel poter simulare un dialogo umano in quanto, a differenza di altre applicazioni che utilizzano strutture predefinite e, nella maggior parte dei casi, statiche (ovvero non modificabili), e´ possibile interpretare dal punto di vista semantico il testo della richiesta immessa dall'utente e fornire di conseguenza la risposta piu´ adeguata alle circostanze. Come detto in precedenza, l'analisi semantica del testo permette di effettuare associazioni tra le parole dal punto di vista del loro significato, grazie a questo e´ quindi possibile non soltanto rispondere a richieste dell'utente utilizzando la ricerca per corrispondenza dei termini immessi, ma anche - e soprattutto - tramite una ricerca basata sul significato dei termini.

4.6 Possibili sviluppi futuri

Un sistema che utilizzi l'analisi della semantica latente quale motore del suo funzionamento potrebbe essere impiegato in differenti

campi di applicazione, uno fra tutti, la ricerca automatica ed intelligente dell'informazione.

E´ infatti possibile utilizzare tali sistemi in differenti modalita´,tra cui:

Intrattenimento ed interazione con l'utente (si vedano tali sistemi come complemento o addirittura sostituzione agli attuali chat-bot).

Recupero di informazioni tramite un'interazione con l'utente sotto forma di dialogo naturale.

Oppure recupero di interi documenti, o parti di essi, come evoluti sistemi di ricerca e recupero informazioni.

4.6.1 Miglioramenti successivi

Per rendere il sistema piu´ efficace ed efficiente, e´ possibile agire su alcuni limiti strutturali presenti nell'attuale versione del programma.

1. Innanzi tutto, sarebbe possibile realizzare uno spazio semantico in maniera tale da renderlo piu´ preciso e particolareggiato, utilizzando un numero maggiore di documenti e verificando rigorosamente la loro formattazione ed i loro contenuti. Con tale affermazione si intende dire che il testo dei documenti, utilizzati come base per la realizzazione dello spazio semantico, potra´ essere sottoposto ad una elaborazione preliminare ancora piu´ accurata, soffermandosi soprattutto sulla fase di *stemming*, ovvero di reperimento delle radici delle parole.

2. La fase di stemming porterebbe ad una considerevole riduzione del numero di termini necessari per creare la matrice dello spazio vettoriale, in quanto molti attuali termini verranno riassunti da un'unica parola che ne e´ la radice, eliminando variazioni di singolare/plurale, maschile/femminile per i sostantivi, gli aggettivi, etc... mentre per i verbi verranno trascurate molte forme di modo e tempo.

Uno spazio semantico cosi´ corretto e´ senza dubbio molto piu´ funzionale poiche´, pur ottenendo una matrice di dimensioni ridotte rispetto a quella di partenza - senza aver effettuato lo stemming - riesce a far ottenere risultati migliori in quanto non e´ piu´ presente il rumore di fondo causato da tutti i termini completi di desinenza.

3. E´ possibile inoltre introdurre ulteriori misure di similarita´ per poter trovare la risposta piu´ attinente alle richieste dell'utente.

Si ricorda che la richiesta dell'utente viene trasformata in forma vettoriale e deve essere confrontata con i vettori rappresentativi delle possibili risposte. La presente applicazione fornisce un metodo di misura che tiene in considerazione le parti parallele ed ortogonali di tali vettori, producendo una misura di similarita´ data dal rapporto tra la parte ortogonale e la parte parallela dei vettori rispettivamente della query e delle risposte.

Tale misura tiene conto, quindi, non solo di quando due vettori possano essere simili - parte parallela - ma anche di quanto siano differenti - parte ortogonale - avendo cosi´ una misura piu´ accurata della distanza tra i vettori.

Si potrebbero introdurre altri tipi di misure, ad esempio la sola parte parallela o la sola parte ortogonale, oppure la distanza di tanimoto, o altre ancora.

4. Per il recupero di interi documenti, si possono introdurre nuove metodologie di costruzione del vettore della query, o addirittura dei vettori rappresentativi dei documenti, concentrandosi sui modi possibili per poter reperire informazioni utili anche da una richiesta non sufficientemente lunga e completa.

Di fatti, essendo i documenti formati da un numero relativamente grande di parole rispetto a quelle della query, potrebbe essere difficoltoso individuare correttamente un documento a partire da una query con un cosi´ minimo contenuto informativo quale puo´ essere quello di una singola frase immessa dall'utente. Utilizzando altri metodi di codifica potrebbe essere possibile individuare un documento gia´ con un ristretto numero di parole e concetti che costituiscono un sottoinsieme di quelli presenti nell'intero documento.

Bibliografia

- [1] - http://www.europarl.eu.int
- [2] - Landauer, T. K., Foltz, P. W., Laham D. - An Introduction to Latent Semantic Analysis - Discourse Processes, 25:259-284, 1998;
- [3] - Landauer, T. K., & Dumais, S. T. (1996). - How come you know so much? - From practical problem to theory. In D. Hermann, C. McEvoy, M. Johnson, & P. Hertel (Eds.), - Basic and applied memory: Memory in context. - Mahwah, NJ: Erlbaum, 105-126;
- [4] - Foltz, P. W. (1996) Latent Semantic Analysis for text-based research - Behavior Research Methods, Instruments and Computers - 28(2), 197-202
- [5] - Landauer, T. K., & Dumais, S. T. (1997). - A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. - Psychological Review. - 104, 211-240.
- [6] - http://lsa.colorado.edu
- [7] - B. Ribeiro-Neto R. Baeza-Yates - Modern Information Retrieval - Addison Wesley, 1999
- [8] - ISO 2382/1, 1984
- [9] - Salton, G. and McGill, M. J. (1983) - Introduction to Modern Information Retrieval - McGraw-Hill, New York, NY

- [10] - Introduction - The Need for Smarter Search Engines - Precision, Ranking, and Recall - the Holy Trinity - http://javelina.cet.middlebury.edu/lsa/out/lsa_intro.htm
- [11] - Berry, Michael, Theresa Do, Gavin O'Brien, Vijay Krishna, and Sowmini Varadhan, 1993 - SVDPACKC (Version 1.0) User's Guide, 1993
- [12] - A. R. Popescu - Implementation of term weighting in a simple IR system - Information Retrieval Project Department of Computer Science University of Helsinki, 2001
- [13] - T. G. Kolda and D. P. O'Leary - A semidiscrete matrix decomposition for latent semantic indexing information retrieval - ACM Transactions on Information Systems - 16(4):322-346, 1998
- [14] - The Term-Document Matrix - http://javelina.cet.middlebury.edu/lsa/out/tdm.htm
- [15] - S. Millozzi - Un approccio innovativo al content management basato su latent semantic indexing e reti di similarita´ - Tesi di Laurea, Universita´ La Sapienza, Roma, 2001
- [16] - M.W. Berry. Large scale singular value computation - International Journal of Supercomputer Application - 6(1):13-49, 1992;
  http://www.cs.utk.edu/~berry/pubs2.html
- [17] - Landauer, T. K., Laham, D., & Foltz, P. W., (1998) - In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), Advances in Neural Information Processing Systems 10,(pp. 45-51) - Cambridge: MIT Press.
- [18] - M. Porter. - An algorithm for suffix stripping. - Automated Library and Information Systems, 1980

- [19] - J. A. Wisniewski - On solving the large sparse generalized eigenvalue problem - Thesis, University of Illinois at Urbana-Champing - 1981
- [20] - http://alice.sunlitsurf.com/alice/aiml.html; http://www.alicebot.org/TR/2001/WD-aiml
- [21] - http://www.alicebot.org
- [22] - http://www.aimlbots.com
- [23] - http://www.pandorabots.com
- [24] - http://www.w3.org/XML; http://www.xml.it:23456/XML/REC-xml-19980210-it.html
- [25] - Latent Semantic Indexing - http://javelina.cet.middlebury.edu/lsa/out/lsa_definition.htm; How LSI Works - http://javelina.cet.middlebury.edu/lsa/out/lsa_explanation.htm; LSI Journal Articles and WebSites - http://javelina.cet.middlebury.edu/lsa/out/lsa_biblio.htm
- [26] - LSI Example - Indexing a Document - http://javelina.cet.middlebury.edu/lsa/out/tutorial.htm
- [27] - Indexing by Latent Semantic Analysis - by S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Journal of the Society for Information Science, 41(6), 391-407, (1990); http://lsi.research.telcordia.com/lsi/papers/JASIS90.pdf

GNU Free Documentation License

Version 1.2, November 2002

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you

copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain

ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of
transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which
states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough

number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
D. Preserve all the copyright notices of the Document.
E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
H. Include an unaltered copy of this License.
I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all

the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties--for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this
License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the

Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents
released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions

of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright (c) YEAR YOUR NAME.
> Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation;
> with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
> A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Note dell'autore:

Questo documento e´ liberamente distribuibile secondo la GNU Free Documentation License (GFDL) in quanto la conoscenza e´, in tutti i suoi aspetti, un bene prezioso al quale non si puo´ in alcun modo vietare la libera diffusione.
Per ogni dubbio e/o perplessita´, oppure semplicemente per ulteriori informazioni, e´ possibile contattare l'autore.

Un sincero grazie a tutti coloro che hanno deciso di spendere un po´ del loro prezioso tempo nel leggere questo documento.

– Salvatore La Bua –
[http://www.shogoki.it]