Indicizzazione

Fasi preliminari

La prima attività necessaria per il funzionamento di un IRS è la generazione di un archivio di documenti. Si tratta di una fase che avviene tipicamente offline e attraverso i seguenti passaggi:

1. *localizzazione*, l'inserimento manuale, semi-automatico o automatico dei documenti alla collezione;
2. *decodifica del formato*, la trasformazione del documento in stringhe;
3. *indicizzazione*, la creazione di una rappresentazione utilizzabile in maniera più efficiente del contenuto di un documento;
4. *generazione struttura dati*, la memorizzazione della struttura dati prodotta durante la fase precedente e dei documenti.

Indicizzazione

L'indicizzazione di un documento testuale è il processo che ne esamina il contenuto informativo, producendo un'opportuna lista di termini indice, molto spesso pesati. I *termini indice* sono definiti come gli elementi base della rappresentazione formale di un documento o di una query, dal punto di vista pratico possono essere: parole, radici di parole, intere frasi estratte da documento oppure parole estratte da un vocabolario controllato. La lista dei termini indice prodotta dalla fase di indicizzazione verrà quindi utilizzata dall'IRS come una versione più compatta del documento di partenza. Il processo di indicizzazione si compone di 5 fasi, anche se non è necessariamente detto che tutti i sistemi di IR le implementino tutte:

1. analisi lessicale e selezione delle parole;
2. rimozione delle stopwords;
3. riduzione delle parole alle rispettivi radici (stemming);
4. pesature degli elementi indice;
5. compressione.

L'*analisi lessicale* e la selezione delle parole rappresentano un processo di trasformazione del testo originario del documento, visto a livello macchina come un flusso di caratteri, in un flusso di tokens, ossia sequenze di caratteri, molto spesso parole, con un significato specifico attraverso l'identificazione di caratteri ritenuti di separazione. (quello che viene fatto dai tokenizer in Lucene)

Il processo di *rimozione delle stopwords* elimina le parole molto frequenti in tutti i documenti, e quindi di scarso contenuto informativo, consentendo una riduzione del 30-50% del numero di tokens utilizzati per la rappresentazione di un dato documento. Questo processo può essere svolto sia utilizzando un dizionario e andando a ricercare i termini funzionali per una data lingua (articoli, preposizioni, ...) oppure a partire dall'analisi statistica della frequenza dei termini nella collezione.

Lo *stemming* è un processo che riduce tutte le parole con la stessa radice (in inglese, stem) ad un unico termine. L'assunzione alla base di questo processo è che le parole con la stessa radice possano avere la stessa origine etimologica e quindi un contenuto informativo molto simile, assunzione valida per molte lingue.

Una volta terminato il processo di indicizzazione di una collezione di documenti, il risultato è rappresentato da una matrice sparsa, in cui i pesi associati ad un termine indice possono essere valori binari, reali o interi positivi.

Termini indice

Come affermato in precedenza, i termini indici fungono da rappresentazione del contenuto informativo di un documento. Per questa ragione, le caratteristiche che si ricercano nei termini indice sono:

- *esaustività*, in quanto devono esprimere quanto più possibile del contenuto informativo del documento,
- *specificità*, in quanto devono consentire il più possibile di distinguere il documento dagli altri.

I termini la cui frequenza è alta in tutti i documenti di una collezione prendono il nome di *termini funzionali*. I termini che identificano il contenuto del documento e hanno una frequenza variabile da un documento all'altro vengono invece detti *indicatori del contenuto* di un documento.

Peso, rank e frequenza

Essendo che non tutte le parole di un documento lo descrivono con la stessa precisione, è possibile associare un *peso* ai termini indice, il quale verrà utilizzato per tenere conto della significatività del termine all'interno del documento. La più semplice funzione di pesatura che si possa definire associa un peso uguale a 1 nel caso in cui il termine sia presente all'interno del documento, 0 in caso contrario. Si tratta come detto di una funzione molto semplice, che non tiene conto però, ad esempio, della frequenza del termine all'interno del documento.

Per ogni termine associato ad un documento è possibile definire due funzioni: frequenza e rank. La frequenza indica la *frequenza* con cui il termine compare all'interno del documento; il *rank* indica la posizione del termine all'interno della lista dei termini presenti nel documento ordinata per frequenza. È interessante sottolineare che il valore del prodotto tra rank e frequenza è costante per tutti i termini di una stessa collezione.

Analisi di Luhn

La *curva di Zipf* rappresenta una funzione che descrive il potere discriminante dei termini all'interno di una collezione andando a presentare il rapporto tra frequenza e rank dei termini. Si osserva che questa curva presenta due forti pendenze in corrispondenza dei termini più frequenti (upper cut-off) e dei termini meno frequenti (lower cut-off). La capacità dei termini di discriminare il contenuto dei documenti è massima nella posizione intermedia tra i due livelli di cut-off. Quest'ultima osservazione è nota come *Analisi di Luhn*.

Molti criteri di indicizzazione sono basati sull'analisi di Luhn, in particolare:

- pesatura dei termini indice, associa un peso minore alle parole più frequenti (upper cut-off);
- stop lists, elimina dagli indici i termini più frequenti (upper cut-off);
- parole significative, elimina dagli indici i termini sia i termini più frequenti che quelli meno frequenti (upper cut-off + lower cut-off).

Inverse document frequency

La *significatività* w di un documento è una funzione composta da due fattori:

$$w_{td} = f_d * discr_t$$

dove f_{td} rappresenta la frequenza del termine t nel documento d, ed è in relazione alla esaustività (fattore di recall), mentre $discr_t$, è in relazione alla specificità (fattore di precision).

L'*Inverse Document Frequency*, o IDF, di un termine t è definita come

$$discr_t = idf_t = log \frac{N}{df_t}$$

dove d_f è il numero di documenti in cui il termine t_j appare e N è il numero di documenti della collezione.

A partire da queste misure, il peso w_{ij} del termine t_i in un documento d_j è definito come

$$w_{ij} = tf_{ij} \times log \frac{N}{df_t}$$

Il peso viene associato ad ogni termine in base al documento in analisi e viene calcolato solamente dopo aver eliminato i termini funzionali. Essendo che la frequenza assoluta tf_{ij} di un termine cresce all'aumentare della lunghezza del documento d_j, il peso viene normalizzato utilizzando la seguente formula

$$w_{ij} = \frac{tf_{ij}}{max\ tf_j} \times log \frac{N}{df_i}$$

dove $max\ tf_j$ è la frequenza massima dei termini nel documento d_j e il primo fattore rappresenta la frequenza relativa del termine t_i nel documento t_j.

Tesauri

Un tesauro è una sorta di mappa per un linguaggio o, nel caso di tesauri tematici, per una sotto-parte di un linguaggio. (es. lessico specifico di una particolare disciplina)

Si possono distinguere 3 diverse tipologie di tesauro: gerarchici, clustered e associativi.

Tesauri gerarchici

I *tesauri gerarchici* rappresentano le relazioni tra i termini da loro contenuti, con particolare attenzione alle relazioni che esprimono l'ordinamento gerarchico tra i vari termini. I tesauri di questo tipo vengono utilizzati soprattutto in fase di espansione delle query e degli indici. Le grandi difficoltà legate a questa tipologia di tesauri sono rappresentate dalle fasi di generazione, in quanto questa necessita di esperti del settore e può essere svolta esclusivamente in maniera manuale, e di manutenzione, la quale si rivela necessaria al fine di mantenere il tesauro in linea con la continua evoluzione del linguaggio ed è anch'essa caratterizzata dagli stessi problemi della fase di generazione.

Tesauri clustered

Un *tesauro clustered* rappresenta un grafo contenente gruppi di parole. All'interno di questo grafo, due gruppi sono connessi se tra di essi è definita una relazione semantica. Un esempio di tesauro di questo tipo, per la lingua inglese, è Wordnet. A differenza della tipologia precedente questi tesauri possono essere generati in maniera automatica.

L'attività di *clustering* consiste in un'attività di raggruppamento dei documenti appartenenti alla collezione in classi all'interno delle quali si trovano documenti simili tra di loro. Si parla di clustering globale in caso di raggruppamenti basati sulla co-occorrenza degli indice nell'intera collezione, di clustering locale se effettuato sulla base del contesto della query.

Tesauri associativi

Infine, i *tesauri associativi*, o pseudo-tesauri, rappresentano dei sottogruppi di termini utilizzando relazioni gerarchiche di similarità e associando peso e verso ad ognuna di queste relazioni. Uno dei vantaggi di questa tipologia di tesauri è rappresentato dal fatto che la loro costruzione può essere automatizzata sfruttando una matrice di similarità e una funzione soglia, entrambe definite a partire dalla co-occorrenza e co-assenza dei termini all'interno dei documenti appartenenti alla collezione.