UNIVERSITA' DEGLI STUDI DI FIRENZE

Facoltà di Ingegneria
Corso di Laurea in Ingegneria Informatica

Elaborato per l'esame di
"Telematica"
Prof. F.Pirri

Web Change

di Menchetti Sauro

A.A. 1999/2000

Introduzione

Ai nostri giorni, uno dei più potenti ed espressivi mezzi di comunicazione è certamente Internet. La sua crescita esplosiva ha fatto si che sempre più persone investissero risorse in questo campo: in questo momento le pagine presenti nel web sono stimate essere circa tre miliardi e mezzo. Una caratteristica peculiare di questo mezzo di comunicazione è la sua continua evoluzione: nascono, si sviluppano e scompaiono una grandissima quantità di pagine in breve tempo. Poiché ci sono grandi quantità di dati in gioco, è necessario realizzare degli strumenti automatici in grado di elaborarli senza richiedere l'intervento umano.

Un problema molto interessante riguarda le modifiche che una pagina subisce durante la sua permanenza nel web. Sarebbe interessante capire quando tali modifiche sono sufficientemente rilevanti tanto da poter dire che è cambiato il contenuto della pagina. Questo potrebbe spingere un utente a voler riaccedere alla pagina per leggere le nuove informazioni presenti. Si tratta quindi di stabilire quanto è cambiato il contenuto di una pagina, basandosi su modifiche strutturali. Si cerca quindi di passare da un'informazione di tipo sintattico-strutturale ad una di tipo semantico.

In sintesi, il problema può essere riformulato nel seguente modo: data una pagina web e data una sua evoluzione temporale, stabilire se la pagina originale e quella modificata sono abbastanza diverse tanto da poter affermare che è cambiato il contenuto delle due pagine: tutto questo in modo automatico, senza la supervisione di un umano.

Rappresentazione del documento

La rappresentazione del documento ha un forte impatto nella risoluzione del nostro problema. Una pagina web è tipicamente una stringa di caratteri. Per stabilire se due due pagine hanno lo stesso contenuto, tali pagine devono essere rappresentate in modo adatto allo scopo che ci si prefigge. Una tra le più sfruttate modellazioni utilizza le parole come unità di rappresentazione e trascura il loro ordinamento nel documento. Si ipotizza quindi che le parole siano tra di loro indipendenti e che siano anche indipendenti dalla posizione che occupano. Questo porta a rappresentare un documento come un "bag of words", cioè come un insieme non ordinato di parole indipendenti tra loro. La rappresentazione come bag of words è equivalente ad una modellazione attributo-valore: ad ogni distinta parola che rappresenta l'attributo, corrisponde una "feature" con associato il numero di volte che la parola compare nel documento od una valore ad esso proporzionale (valore

dell'attributo). Per evitare di creare degli insiemi di features troppo popolati, si esegue una procedura di feature selection che seleziona come features solo le parole che hanno maggiore contenuto informativo.

Selezione delle features rilevanti

Non tutte le parole presenti in un insieme di documenti hanno la stessa rilevanza e lo stesso valore semantico: ad esempio, una congiunzione od un articolo hanno un peso minore rispetto a certe parole chiave presenti nel documento. È necessario quindi procedere ad una selezione delle parole con più alto contenuto informativo.

Una prima selezione può essere fatta usando una lista di parole semanticamente meno significative dette "stop words" e mappando le parole presenti nel documento su tale lista: se una parola del documento compare anche nella lista, allora viene cancellata dal documento. Tale lista viene chiamata "stop list" e può essere realizzata mediante un file di testo: conterrà articoli, congiunzioni, etc.

Una volta eliminate le stop words dal documento, si può applicare un algoritmo di "stemming" alle parole rimaste. Tale algoritmo prende una parola ed estrae da essa la sua radice: ad esempio, una parola singolare ed una plurale vengono così ad assumere lo stesso valore, così come un verbo coniugato o meno.

Le due feature selection descritte si applicano ad ogni tipo di documento. Poiché il nostro problema coinvolge delle pagine html, possiamo introdurre anche una funzione che pesi le parole in modo diverso a seconda di quali tag html coinvolgono. In questo modo le parole presenti nel titolo della pagina, nei meta, negli h1, h2, h3, nei link assumono un peso maggiore delle altre. Gli altri tag html di minore importanza vengono cancellati dalla pagina.

Altre elaborazioni di minore importanza riguardano la cancellazione dei caratteri non alfabetici (ad esempio la punteggiatura) e nel rendere case insensitive la pagina.

A questo punto il documento è pronto per essere suddiviso in token: tali token sono le feature più informative di quella pagina.

Costruzione del vocabolario

Per stabilire se due pagine che trattano di un certo argomento hanno lo stesso contenuto, sarebbe utile disporre di un insieme di pagine che parlino di quell'argomento da cui poter estrarre un vocabolario di parole. Avendo a disposizione un certo insieme di pagine, si processano una ad una per estrarre le feature rilevanti: le feature estratte da ogni pagina contribuiscono così alla

costruzione di un dizionario comune. Sia $|V|$ il numero di elementi di tale vocabolario.

Criterio di confronto

L'algebra lineare permette di stabilire un criterio di confronto tra due documenti. Ecco ciò che è utile al nostro scopo.

Richiami di algebra lineare

Si definisce **prodotto scalare** fra gli elementi $X, Y \in \mathbb{R}^n$ il numero reale:

$$< X, Y >= x_1 y_1 + \cdots + x_n y_n = \sum_{i=1}^{n} x_i y_i.$$

Si verifica che valgono le seguenti proprietà:

$\forall\ X, Y, Z \in \mathbb{R}^n, \forall\ \lambda, \mu \in \mathbb{R}$
$< X, Y >=< Y, X >$
$< \lambda X + \mu Y, Z >= \lambda < X, Z > + \mu < Y, Z >$
$< X, X > \geq 0$ e $< X, X >= 0$ se e solo se $X = O$.

$X, Y \in \mathbb{R}^n$ si dicono **ortogonali** se $< X, Y >= 0$.
Si dice **norma** di $X \in \mathbb{R}^n$ il numero reale non negativo $\|X\| = \sqrt{< X, X >}$.
Vale la seguente **disuguaglianza di Schwarz**:

$$< X, Y > \leq \|X\|^2 \|Y\|^2 \quad \forall\ X, Y \in \mathbb{R}^n.$$

Dalla disuguaglianza di Schwarz segue immediatamente:

$$-1 \leq \frac{< X, Y >}{\|X\| \|Y\|} \leq 1.$$

Dati $X, Y \in \mathbb{R}^n \backslash \{O\}$, si dice **angolo** fra X e Y l'angolo $\theta \in [0, \pi]$ tale che

$$\cos\theta = \frac{< X, Y >}{\|X\| \|Y\|}.$$

Algoritmo di confronto

L'algoritmo di confronto si basa sulla seguente rappresentazione del documento. Ogni documento d è rappresentato come un vettore $d = (d_1, \cdots, d_{|V|})$, dove $|V|$ è il numero di elementi del vocabolario costruito in precedenza.

In questo modo i documenti con un contenuto simile avranno un vettore di rappresentazione simile in accordo ad una fissata metrica di similarità. Ogni elemento d_i rappresenta una distinta parola w_i. Il termine d_i viene calcolato come combinazione dei due termini $TF(w_i, d)$ e $DF(w_i)$. Il termine $TF(w_i, d)$ (term frequency) è il numero di volte che la parola w_i compare nel documento d, mentre il termine $DF(w_i)$ (document frequency) è il numero di documenti nei quali la parola w_i compare almeno una volta. Possiamo così definire il termine $IDF(w_i)$ (inverse document frequency) nel seguente modo:

$$IDF(w_i) = \log\left(\frac{|D|}{DF(w_i)}\right)$$

dove $|D|$ è il numero totale dei documenti. Intuitivamente il termine $IDF(w_i)$ è piccolo se la parola w_i compare in molti documenti ed è grande se la parola w_i compare in un solo documento. Il peso d_i di una parola w_i nel documento d è dato da

$$d_i = TF(w_i, d)\ IDF(w_i).$$

L'euristica per pesare una parola afferma che una parola w_i è termine importante per il documento d se compare molte volte in esso. D'altra parte, le parole che compaiono in molti documenti sono meno importanti come conferma la loro frequenza inversa.

Poiché i nostri documenti sono delle pagine html, non tutte le parole presenti nella pagina hanno la stessa importanza. Alcuni tags html attribuiscono un particolare significato alle parole a cui sono legati e sembra quindi opportuno pesare diversamente quelle parole. In conclusione, il peso d_i di una parola w_i nel documento d è dato da

$$d_i = TF(w_i, d)\ IDF(w_i)\ WEIGHT(w_i, tag)$$

dove $WEIGHT(w_i, tag)$ rappresenta il peso della parola associata ad un particolare tag html.

A questo punto interviene l'algebra lineare. Siano H_1 e H_2 le due pagine html da confrontare e siano d_1 e d_2 le rispettive rappresentazioni. Come criterio di confronto viene scelto il coseno dell'angolo compreso tra i due vettori d_1 e d_2:

$$\cos\theta = \frac{< d_1, d_2 >}{\|d_1\|\|d_2\|} \in [0, 1].$$

Se le due pagine sono molto simili, i due vettori di rappresentazione saranno molto vicini e l'angolo compreso sarà quindi piccolo, risultando in un valore del coseno prossimo ad 1. D'altra parte, se le due pagine sono completamente diverse, i due vettori di rappresentazione saranno ortogonali e il coseno dell'angolo compreso varrà 0.

Adesso non resta che definire una soglia al di sopra della quale si attribuisce lo stesso contenuto ai due documenti. Un valore appropriato sembra essere compreso tra $[0.8, 0.9]$.

Sketch dell'algoritmo di confronto

L'algorimo di confronto è composto di due passi:

- costruzione del vocabolario;
- confronto delle due pagine html.

Sia D un insieme di pagine html che trattano lo stesso argomento di cui parlano le due pagine che vogliamo confrontare. Per ogni pagina dell'insieme D, si selezionano le features mediante la procedura di feature selection descritta sopra. A questo punto, ogni pagina è rappresentata come un insieme di feature caratteristiche. A partire da questo insieme D di pagine rappresentate ognuna tramite le sue parole più significative, si costruisce un vocabolario di termini che riguardano l'argomento di interesse. Sia $|V|$ il numero di parole del vocabolario. Durante la costruzione del vocabolario, si valutano anche i termini $DF(w_i)$ per ogni parola del vocabolario.

Avendo a disposizione un vocabolario di termini per l'argomento di interesse, si rappresentano le due pagine che si vogliono confrontare ognuna come un vettore di $|V|$ elementi, cioè si mappano le due pagine nel vocabolario. Ad ogni parola w_i viene associato un peso d_i come descritto in precedenza.