Valutazione

Valutare le prestazioni di un sistema di Information Retrieval è un compito piuttosto complesso. Le valutazioni che vengono presentate in questo capitolo sono dette *laboratory based evaluations*. Generalmente il procedimento applicato è il seguente:

1. si installa una collezione standard;
2. si osserva il comportamento del sistema rispetto ad insieme di query definite ad hoc da chi ha prodotto la collezione;
3. vengono misurate i valori medi di richiamo e precisione ottenuti dal sistema per le query proposte.

Per svolgere questo procedimento è necessario avere a disposizione:

- una collezione di documenti standard e un insieme di query relative al contenuto di questa collezione, l'utilizzo di collezioni standard e di un insieme specifico di query è infatti l'aspetto che rende possibile il confronto tra sistemi;
- un insieme di utenti, anch'esso relativo alla collezione utilizzata, nel caso in cui questo sia necessario;
- dei criteri di valutazione e delle misure che consentano di esprimere le performance del sistema;
- il progetto dell'esperimento, il quale specifica come si intende unire i risultati delle misurazioni effettuate per ottenere la valutazione del sistema.

Esistono diverse collezioni standard, definite nel corso degli anni da enti e organizzazioni note nel campo dell'Information Retrieval. La più nota tra queste organizzazioni è la TREC: Text REtrieval Conference/Competition.

Criteri di valutazione

Esistono diversi criteri di valutazione. Questi criteri di valutazione si distinguono per l'aspetto di un sistema di IR che tengono in considerazione. La valutazione a cui siamo maggiormente interessati è quella relativa all'*efficacia del retrieval*: la capacità di un sistema di IR di identificare l'insieme preciso e completo dei documenti utili all'utente.

Altri criteri di valutazione misurano aspetti come:

- funzionalità, la facilità con cui l'utente interagisce con il sistema;
- correttezza, l'assenza di errori semantici o sintattici nei risultati;
- performance, l'efficienza del sistema in termini di tempo e spazio.

Durante la valutazione di un sistema di IR è necessario tralasciare gli aspetti legati alla soggettività del retrieval. Un documento potrebbe essere più o meno utile ad un utente e in alcuni casi un utente potrebbe scoprire un nuovo bisogno informativo a partire dai documenti ritornati dal sistema*. Per poter svolgere la valutazione in modo più agevole, durante le laboratory based evaluations si tralasciano gli aspetti legati alla

soggettività e si suppone che un documento possa o soddisfare o non soddisfare l'utente. In sostanza, la rilevanza di un documento viene modellata in modo binario.

* quest'ultimo fenomeno viene detto *apprendimento accidentale*

Richiamo e precisione

Lo scopo di un buon sistema di retrieval è quello di reperire quanti più documenti rilevanti possibile, minimizzando il numero di documenti non rilevanti reperiti. In termini insiemistici, l'obiettivo di un sistema di IR è quello di massimizzare l'intersezione tra i l'insieme dei documenti reperiti e l'insieme dei documenti rilevanti.

L'efficacia del retrieval viene quindi studiata a partire da due misure:

- *richiamo*, la proporzione di materiale rilevante che viene reperito;

$$recall = \frac{| \, rilevanti \; e \; reperiti \, |}{| \, rilevanti \, |}$$

- *precisione*, la proporzione di materiale reperito che è rilevante.

$$precisione = \frac{| \, rilevanti \; e \; reperiti \, |}{| \, reperiti \, |}$$

Per poter studiare queste due misure, si compie un'ulteriore assunzione: si assume che esista un insieme ben definito di documenti che soddisfa il bisogno informativo dell'utente e che questo insieme sia noto a priori per ogni query. Questo non è mai vero durante il funzionamento reale di un sistema, soprattutto nel campo del web, in cui non si è neppure a conoscenza dei confini della collezione su cui si basa il sistema. È un'assunzione molto forte ma fondamentale per poter calcolare le misure richiamo e precisione di un sistema.

A partire da queste due misure è possibile individuare due casi limite particolarmente interessanti:

- $precisione = 1$, se la precisione è massima il richiamo sarà probabilmente molto basso; questo perché nella maggioranza dei casi solamente un insieme ridotto di documenti della collezione è rilevante rispetto ad una query e, in caso di precisione massima, solamente quell'insieme piccolo insieme verrà reperito;
- $richiamo = 1$, in questo caso l'insieme dei documenti ritornati rappresenta l'intera collezione; sicuramente questo insieme conterrà anche tutti i documenti quelli rilevanti rilevanti, l'utente sarà però obbligato a scorrere un lungo elenco di documenti non rilevanti per poterli identificare, rendendo inutile il lavoro del sistema.

Da un punto di vista matematico, il rapporto tra richiamo e precisione è inversamente proporzionale e può quindi essere descritto da una funzione monotona decrescente. A valori alti di precisione corrispondono valori bassi di richiamo e viceversa. Tipicamente, un sistema viene giudicato positivamente se offre un compensazione tra queste due misure e se raggiunge una media tra precisione e richiamo del 60% circa.

Un sistema di retrieval può essere modificato in maniera tale da favorire una misura piuttosto che l'altra. Tuttavia, queste due misure dipendono in parte anche da aspetti sui quali non è possibile agire: dalla collezione utilizzata e da ciò che l'utente ritiene rilevante.

Metodologie per la valutazione

Essendo che i sistemi di IR restituiscono una lista di documenti ordinata per rilevanza, solitamente l'efficacia viene valutata analizzando in primi n risultati. Esistono due metodologie per effettuare questa analisi: considerando sottoinsiemi di cardinalità n o considerando i *livelli di recall*.

Il primo metodo prevede che:

1. venga definita una serie di dimensioni precise, ad esempio 10, 20, 30, 40 e 50;
2. venga calcolata la precisione per insiemi costruiti a partire dalle dimensioni specificate al passo precedente. In pratica questo significa che viene calcolata la precisione nel caso di un insieme contenente i primi 10 documenti ritornati del sistema, poi viene calcolata la precisione sull'insieme contenente i primi 20 documenti ritornati e così via;
3. la valutazione finale viene espressa calcolando la media per i valori di precisione ottenuti sui diversi sottoinsiemi.

Il secondo metodo prevede di utilizzare i livelli di richiamo (percentuali di richiamo) e calcolare la precisione per ognuno dei tagli prodotti da questi. In questo caso si considera una lista contenente tutti i documenti della collezione ordinati per rilevanza e la precisione viene calcolata per un primo taglio contenente il 10% della collezione (richiamo del 10%), per un secondo taglio contenente il 20% della collezione (richiamo del 20%) e così via. Al termine dei calcoli la valutazione viene espressa come la media del rapporto tra precisione e richiamo ai diversi livelli.

Le valutazioni prodotte utilizzando queste due metodologie possono essere usate in modo molto semplice per confrontare due sistemi. È infatti sufficiente osservare quale sistema ha una valutazione maggiore per conoscere quale dei due sistemi è generalmente* migliore.

* “generalmente” perché, essendo che il valore finale è ottenuto facendo una media, il sistema con la valutazione più bassa potrebbe essere migliore del sistema con una valutazione più alta su un certo numero di tagli.

Problemi

Alcuni problemi delle laboratory based evaluations sono i seguenti:

- è necessario avere a disposizione un gran numero di query per avere un’idea dei risultati reali del sistema;
- è necessario avere a disposizione una collezione standard, che tipicamente viene offerta a pagamento; questo perché se utilizzassi una collezione proprietaria non potrei confrontare i risultati del mio sistema con quelli di altri sistemi;

- valutazioni di questo tipo sono solamente approssimazioni del reale perché non possiamo conoscere i veri valori richiamo e precisione, soprattutto nell'ambito web;
- si assume in maniera ingiustificata una modalità batch di retrieval, ossia che l'utente non possa cambiare la propria interpretazione di rilevanza durante l'interazione con il sistema;
- si assume in maniera ingiustificata che l'ordinamento stretto non sia importante ma sia sufficiente analizzare la rilevanza in modo binario.

Altre misure interessanti

L'*accuratezza* è un'altra grandezza molto importante per la valutazione di un sistema di IR, in particolare per i sistemi di Information Filtering. L'accuratezza rappresenta il grado di accordo tra utente e sistema, ed è definita matematicamente come:

$$accuratezza = \frac{|\ rilevanti\ e\ reperiti\ | + |\ non\ rilevanti\ e\ non\ reperiti\ |}{|\ intera\ collezione\ |}$$

ossia *quanto ci becca™*.

La *f measure* del documento j-esimo è definita come la media armonica tra la precisione P_j e il richiamo R_j corrispondenti al j-esimo documento nella lista ordinata dei documenti reperiti dal sistema.

$$F_j = \frac{2}{\frac{1}{R_j} + \frac{1}{P_j}}$$

In particolare, se $F_j = 0$ significa che nessun documento rilevante è stato reperito, mentre $F_j = 1$ quando tutti i documenti rilevanti sono stati reperiti.

La *e measure* è molto simile alla f measure ma consente di calibrare il rapporto tra precisione e richiamo utilizzando il parametro α:

$$E_j = \frac{2}{\alpha(\frac{1}{R_j}) + (1-\alpha)\frac{1}{P_j}}$$

User studies

I sistemi di IR personalizzato non possono essere valutati utilizzando le laboratory based evalutations, questo perché nell'IR personalizzato solamente l'utente, attraverso le sue interazioni con il sistema, sarà in grado di decidere quali documenti sono rilevanti e quali no.

I sistemi di IR personalizzato vengono valutati utilizzando i cosiddetti user studies. Uno *user study* è una procedura mediante la quale viene considerato un certo numero di utenti, generalmente compreso tra 15 e 50 utenti, vengono definiti i profili per ogni utente e si realizza una valutazione sulla base dell'interazione tra gli utenti reali e il sistema.

Misure dipendenti dall'utente

Esiste poi un altro insieme di misure utili alla valutazione di un sistema di IR. Queste misure hanno in comune l'essere tutte dipendenti dall'utente.

La *novità* è una misura che viene privilegiata da molti motori di ricerca, i quali implementano delle tecniche apposite per fare in modo che i risultati mai visti dall'utente vengano privilegiati all'interno della lista dei documenti presentata a fronte di una query. La novità è definita come:

$$novità = \frac{|R_u|}{|R_u| + |R_k|}$$

dove R_k è l'insieme dei documenti rilevanti e reperiti ed R_u è l'insieme dei documenti rilevanti, reperiti e sconosciuti all'utente fino ad ora.

La *copertura* rappresenta la percentuale di documenti reperiti e considerati rilevanti dal sistema rispetto a quelli effettivamente effettivamente considerati rilevanti dall'utente. Si tratta di una sorta di richiamo adattato all'utente ed è definito come:

$$copertura = \frac{|R_k|}{|U|}$$

dove U rappresenta l'insieme dei documenti rilevanti secondo l'utente.

Il *recall relativo* misura la percezione che l'utente ha dell'efficacia del sistema ed è definita come il rapporto il numero di documenti reperiti e rilevanti e numero di documenti rilevanti che l'utente si aspetta in risposta a una query:

Lo *sforzo di recall* è formalmente definito come il rapporto tra il numero di documenti rilevanti che l'utente si aspetta in risposta a una query e il numero di documenti esaminati tra quelli reperiti per trovare quelli rilevanti.