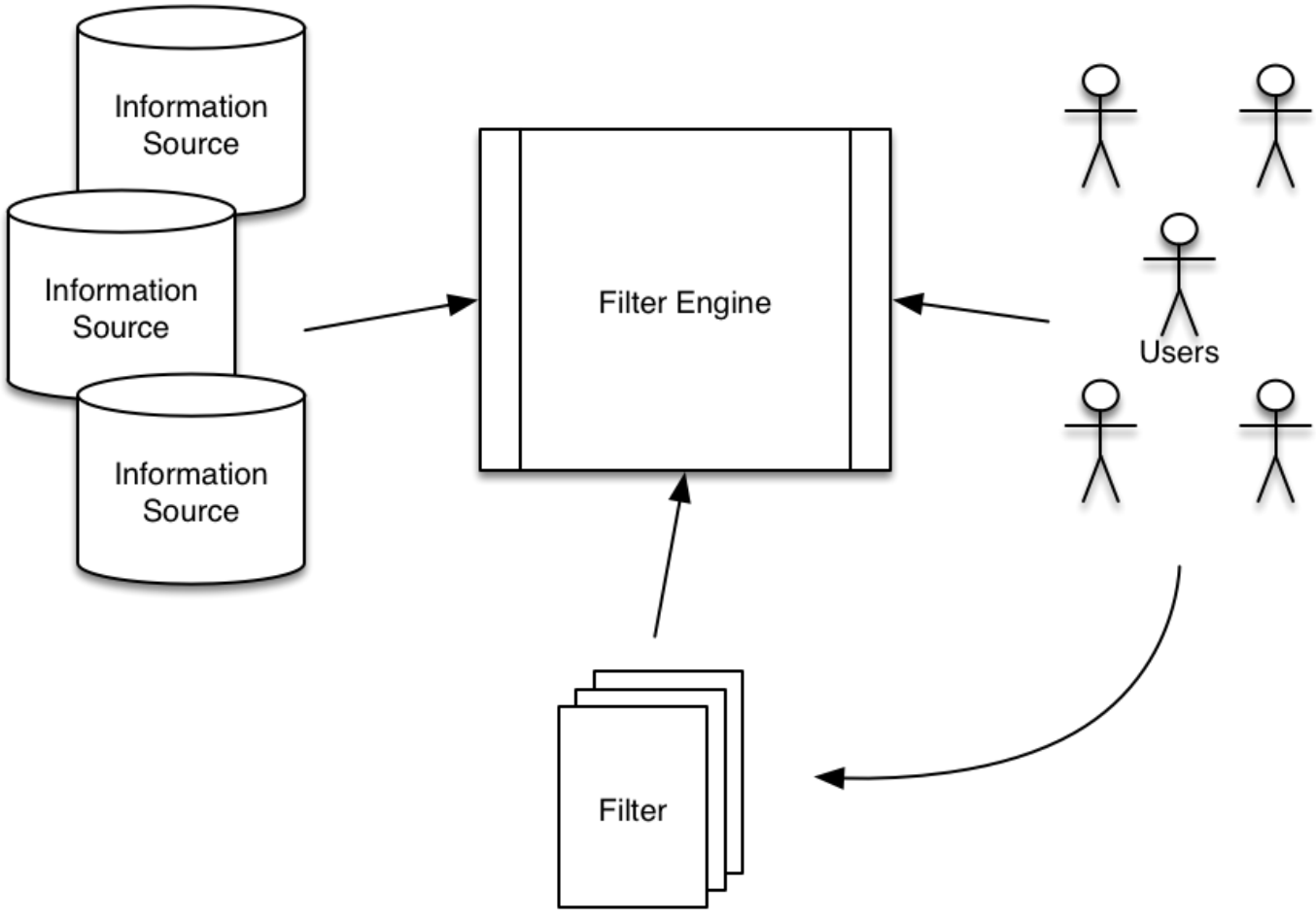
Information Filtering

L'*Information Filtering* è il processo che consente il monitoraggio di grandi quantità di informazioni generate dinamicamente al fine di presentare all'utente un sottoinsieme di queste. Questo sottoinsieme conterrà le informazioni identificate dal sistema come rilevanti per gli interessi e bisogni informativi dell'utente.

Il componente fondamentale di un sistema di Information Filtering è rappresentato dall'Information Filter. L'*Information Filter* è un componente che cerca di rappresentare gli interessi di un dato utente e ha come obiettivo quello di identificare solamente le informazioni interessanti questo.

Come può però un Information Filter acquisire informazioni riguardo ai bisogni di un utente e, possibilmente, adattarsi dinamicamente ai cambiamenti che questi possono subire?



In seguito vengono presentate alcune definizioni di Information Filtering.

- Un processo di selezione e di presentazione di elementi a partire da un grande insieme di possibilità in cui gli elementi vengono presentati in maniera ordinata per priorità. (Malone et al., 1987)
    - es. commercial recommendations

- Un campo di studio progettato per la creazione di un approccio sistematico all'estrazione di informazioni, che una particolare persona trova importanti, a partire da un più grande flusso di informazioni. (Cavanese, 1994)
    - es. news stream filtering

- Strumenti che cercano di filtrare materiale irrilevante. (Khan & Card, 1997)
    - es. spam filtering

Tipologie di Information Filtering

Possono essere identificate tre categorie di sistemi di Information Filtering. Queste categorie differiscono per le modalità con cui l'Information Filter viene definito e acquisito:

- Content-based Filtering, gli oggetti da filtrare sono generalmente rappresentati da testi e il filter engine è basato sull'analisi dei contenuti;
- Collaborative Filtering, gli oggetti da filtrare sono prodotti o beni e il filter engine è basato sull'analisi dell'utilizzo ;
- Hybrid Filtering, una combinazione degli approcci precedenti.

Content-based Filtering

Nel Content-based Filtering, gli oggetti che devono essere filtrati sono documenti, generalmente di tipo testuale. Il filter engine di un sistema content-based applica un'analisi del contenuto dei documenti. Il filtro a partire dal quale viene svolta questa analisi, noto anche come profilo utente, rappresenta gli interessi di un utente o di gruppo di utenti. In particolare, un filtro cerca di riflette gli interessi a lungo termine degli utenti.

Content-based Filtering e IR

IF e IR: sono due facce della stessa medaglia. Mentre l'IR si occupa della selezione di testi da un repository di documenti, l'IF si occupa della selezione di testi da un flusso di dati dinamico. Mentre l'IR si occupa di fornire alle risposte ad una query proposta attivamente dall'utente, soddisfando quindi i bisogni immediati, il Content-based IF si occupa di soddisfare i bisogni informativi a lungo termine, espressi attraverso una serie di ricerche di informazioni.

Inoltre, come affermato durante le prime lezioni, un'altra differenza fondamentale tra IR e IF è rappresentata dal fatto che i sistemi del primo tipo agiscono in modalità pull, i secondi agiscono in modalità push, le informazioni non devono essere richieste dall'utente ma gli vengono fornite in maniera automatica.

IF as a classification problem

Il Content-based Filtering richiede la definizione di un modello formale in grado di rappresentare sia i documenti che i profili degli utenti, in maniera tale da consentire il calcolo della similarità tra un documento e ciascun profilo utente. Generalmente i sistemi di Information Filtering fanno uso di una soglia di rilevanza da applicare al loro filtro: dato un profilo p e una soglia di rilevanza q, un documento D è rilevante per P se

$sim(D, P) > q$. Uno dei problemi più importanti nell'Information Filtering è la definizione di questa soglia di rilevanza, la quale rappresenta la risoluzione di problemi di classificazione.

Profili e modelli dell'utente

I profili sono di grande importanza per le performance dei sistemi di information Filtering. La costruzione di un "buon" profilo rappresenta ancora l'ostacolo più significativo per ottenere un sistema di Information Filtering le cui performance siano ragionevoli: "la costruzione di profili accurati è una sfida chiave - il successo del sistema dipende in gran parte dall'abilità del profilo appreso di rappresentare i reali interessi dell'utente." (Balabanovic & Shonan, 1997)

La costruzione di un modello dell'utente può essere:

- esplicita, in questo caso il modello viene costruito dal sistema a partire da informazioni che vengono esplicitamente fornite dall'utente;
- implicita, in questo caso il modello viene costruito attraverso un processo di apprendimento basato sui feedback che vengono forniti dall'utente, inferiti dalle risposte ricevute, e sul comportamento dell'utente, inferito dalle azioni compiute.

In questo contesto, le problema fondamentali da tenere in considerazione sono rappresentate dalla sforzo richiesto all'utente e dal grado di controllo esercitato da questo.

Vector space filtering model

Collaborative Filtering

Nell'ambito del Collaborative Filtering, o Social Filtering, gli oggetti che devono essere filtrati sono documenti, prodotti o servizi e il filtraggio si basa sull'analisi del comportamento degli utenti/consumatori. Un sistema di Collaborative Filtering utilizza un database delle preferenze degli utenti per trovare utenti con interessi simili e predire se un certo elemento possa risultare interessante o meno per un particolare particolare utente a partire da come gli altri utenti hanno valutato quello stesso elemento, prima che l'utente abbia mai visto quel particolare elemento.

Formalmente, dato l'insieme degli utenti C, l'insieme degli elementi C e una funzione di utilità che misura l'utilità di un elemento s per un utente c, definita come $u : C \times S \rightarrow R$, dove R è un insieme totalmente ordinato, l'utilità $u(c, s)$ di un elemento s sconosciuto all'utente c viene stimata a partire dall'utilità $u(c_j, s)$ assegnata all'elemento s dagli utenti $c_j \in C$ considerati "simili" all'utente c.

I metodi per la creazione di collaborative recommendations possono essere classificate all'interno di due classi generali: i metodi memory-based e quelli model-based.

I metodi memory-based determinano le loro previsioni di valutazione a partire dall'intera collezione di elementi valutati già valutati dagli utenti. La valutazione sconosciuta $r_{c,s}$ rispetto ad un elemento s di un utente c viene calcolata come un aggregato delle valutazioni fornite da altri utenti (tipicamente gli N più

simili) per lo stesso elemento s:

$$r_{c,s} = aggr_{c' \in C^*}(r_{c',s})$$

dove C^* denota l'insieme degli N utenti che sono considerati più simili all'utente c che hanno valutato l'elemento s.

Esistono diversi approcci per il calcolo della similarità tra due utenti utilizzati nell'ambito del Collaborative Filtering. La maggior parte di questi utilizza le valutazioni espresse dai due utenti relativamente agli stessi elementi come base per il calcolo della similarità. I due approcci più diffusi sono quello basato sulla correlazione e quello basato sul coseno.

Altri metodi

I metodi model-based invece utilizzando la collezione delle valutazioni espresse dagli utenti per apprendere un modello, il quale viene poi utilizzato per prevedere le valutazioni per i nuovi elementi.

I metodi ibridi possono a loro volta essere divisi in quattro tipologie distinte:

1. metodi che implementano Collaborative Filtering e Content-based Filtering separatamente e combinano i risultati ottenuti da questi;
2. metodi che incorporano caratteristiche del Content-based Filtering all'interno di approcci di tipo Collaborative;
3. metodi che incorporano caratteristiche del Collaborative Filtering all'interno di approcci di tipo Content-based;
4. metodi che costruiscono un modello generale che incorpora sia caratteristiche del Content-based Filtering sia caratteristiche del Collaborative Filtering.