Information Retrieval

Informazione e IR

L'*informazione* può essere definita come la l'unione tra un insieme di dati e della loro interpretazione, dove i dati rappresentano fatti elementari che devono appunto essere interpretati per poter arricchire la conoscenza. Si dice che l'informazione produce una variazione della conoscenza. L'informazione può assumere diverse forme, dai testi espressi in linguaggio naturale alle immagini, dai suoni alle mappe.

L'*Information Retrieval* è una disciplina informatica. Il suo obiettivo è la costruzione di software che permetta la memorizzazione di ingenti quantità di documenti all'interno di un archivio. Questo archivio deve consentire un efficiente reperimento e ordinamento dei documenti, in maniera tale da poter soddisfare le necessità informative degli utenti.

Information Retrieval System

Un *Information Retrieval System*, o IRS, è quindi definito come un sistema che opera da intermediario, interpretando le necessità informative dell'utente e stimando la rilevanza dei documenti rispetto ad esse.

Un IRS può effettuare il reperimento delle informazioni richieste in due modalità differenti:

1. *pull*, l'utente richiede esplicitamente informazioni in maniera interattiva;
2. *push*, l'utente viene automaticamente aggiornato con informazioni che il sistema ritiene che siano di interesse per lui.

Rilevanza, efficienza ed efficacia

Le due misure chiave dell'information retrieval sono efficienza e efficacia:

1. l'*efficienza* è un problema di tipo tecnico, si studiano problemi di efficienza quando si ricercano modalità per rappresentare e manipolare l'informazione utilizzando degli elaboratori;
2. l'*efficacia* è un problema di tipo semantico, si studiano problemi di efficacia quando si ricercano modalità per sintetizzare e memorizzare l'informazione conservandone il significato originario.

La *rilevanza*, detta anche Retrieval Status Value, è definita come la pertinenza, utilità di un documento in accordo ad una query espressa dall'utente. Alla luce di questa definizione, l'obiettivo di un IRS può essere ridefinito come il reperimento di tutti i documenti rilevanti per l'utente, trascurando al tempo stesso quelli non rilevanti.

Differenza tra DBMS e IRS

Nonostante i DBMS e gli IRS siano entrambi sistemi per l'accesso ad informazioni è importante fare

attenzione alle differenze. Mentre nel contesto delle Basi di Dati le interrogazioni sono rigide e la semantica dei dati è ben definita, nel contesto del Information Retrieval la semantica delle interrogazioni e dei documenti è spesso vaga e i risultati ottenuti presentano spesso piccoli errori in quanto prodotti a partire da una stima di rilevanza. Queste differenze sono causate soprattutto delle differenti necessità che hanno portato alla nascita di queste due diverse tipologie di dati, mentre i primi nascono dalla necessità di gestire crescenti quantità di dati relativi ad applicazioni aziendali tradizionali, i secondi nascono dalla necessità di gestire, classificare, reperire libri e articoli in biblioteche e librerie

Struttura di un IRS

Un IRS è basato su un *modello matematico* che fornisce una descrizione formale dei documenti, delle query e del modo in cui confrontare le rappresentazioni di questi due in maniera tale da produrre delle liste di documenti stimati rilevanti.

Ad alto livelli un IRS è composto da 4 elementi:

- archivio di documenti;
- rappresentazione formale dei documenti, che sintetizza il contenuto informativo dei documenti e viene ottenuta mediante il processo di indicizzazione;
- linguaggio di query, che consente all'utente di esprimere le condizioni per la selezione di documenti di interesse;
- meccanismo di matching, che confronta la rappresentazione dei documenti e le condizioni di selezione espresse attraverso una query al fine di produrre la lista dei documenti rilevanti rispetto ad essa.

Rispetto ad una particolare query, un documento può essere classificato come rilevante/non rilevante e come reperito/non reperito. A partire da queste due classificazioni sono definite le due metriche fondamentali dell'Information Retrieval:

$$precision = \frac{|rilevante \cap reperito|}{|reperito|}$$

$$recall = \frac{|rilevante \cap reperito|}{|rilevante|}$$

Le principali difficoltà dell'Information Retrieval sono rappresentate da:

- incompletezza della rappresentazione dei documenti;
- soggettività del concetto di rilevanza;
- ambiguità del significato dei termini;
- vaghezza delle richieste utente;
- incertezza rispetto alla correttezza del risultato;
- approssimatività del meccanismo di confronto.

Un altro problema particolarmente significativo nel campo dell'IR è quello dell'*ottimizzazione* rispetto a spazio e tempo di esecuzione. La *compressione* nasce dalla necessità di ottimizzare l'occupazione di

memoria e i tempi di trasmissione nell'IR distribuito a discapito dei tempi necessari per compressione e decompressione. Alcuni tipi di compressione consentono il cosiddetto *matching compresso*.

Documenti e linguaggi di markup

Un *documento* è definito come l'unità di informazione reperibile.

Un *archivio* è definito come un insieme di documenti, un insieme che può essere sia centralizzato che distribuito.

A livello pratico, un documento è composto da sezioni o campi distinti e non sovrapposti, di contenuto testuale o multimediale e di lunghezza variabile, delimitati da tag. È possibile distinguere tra documenti strutturati e documenti semi-strutturati:

- strutturati, documenti con una struttura rigidamente fissata;
- semi-strutturati, documenti caratterizzati da irregolarità nella struttura.

I *documenti semi-strutturati* devono comunque essere conformi ad un modello di dati semi-strutturato, definito utilizzando un meta-linguaggio. Sono rappresentabili da un grafo diretto in cui: i nodi rappresentano i campi del documento e gli archi rappresentano le relazioni tra campi.

Linguaggi di markup

I *linguaggi di markup* sono stati definiti per permettere l'utilizzo di istruzioni di strutturazione e formattazione tramite comandi testuali all'interno dei documenti. Le istruzioni di marcatura utilizzate da questi linguaggi sono chiamate *tag* e delimitano la parte di testo a cui sono applicate. *SGML* è il meta-linguaggio standard per la definizione di linguaggi di markup. Un esempio di linguaggio prodotto a partire da SMGL è HTML.

XML è un sottoinsieme di SGML adatto a rappresentare documenti strutturati concepiti come aggregazioni di unità indipendenti, chiamate entità o oggetti. Queste unità rappresentano il contenuto informativo del documento. I vincoli strutturali per i documenti appartenenti ad una collezione di documenti XML possono essere espressi utilizzando un *DTD*, il quale però è del tutto opzionale.

Un documento XML può essere rappresentato in maniera semplificata attraverso una *struttura ad albero* in cui:

- a ogni campo corrisponde un nodo interno;
- al testo libero corrispondono le foglie;
- i figli di un nodo sono i campi o il testo in esso contenuti.

Questo modello ad albero può anche essere esteso in maniera tale da includere attributi, commenti e istruzioni e altri elementi che possono comporre un documento XML.

Il *Document Object Model*, o DOM, rappresenta un'astrazione tra un documento (XML o HTML) e l'applicazione che lo deve utilizzare o elaborare. Questo livello di astrazione aggiuntivo consente la

definizione di operazioni standard utili per lavorare con i documenti.

Metadati

I metadati sono definiti come un insieme di dati associati ad un documento e relativi ad esso. Possono essere classificati in:

- descrittivi, se relativi alla creazione del documento;
- semantici, se relativi all'argomento trattato dal documento.