

Informazione Elettronica, Ricerca ed Internet

Corso di *Informatica Generale*
(Roberto BASILI)

Teramo, 21 Dicembre, 1999

Informazione Elettronica, Applicazioni ed Internet

- Informazioni Automatizzate
- Strumenti di Automazione
- Applicazioni
- Il reperimento della informazione
- IR & Internet

Informazioni Automatizzate

Interno Esterno Pubblico	Sw libraries Lettere	Grafici	Campioni	Chioschi Elettronici
	Basi di Dati Basi di Conoscenza	Immagini	Video	Pubblicazioni Multimediali
	Thesaura, Ontologie	Videotex	Video Clips	Web pages Internet Portals
	DATI	TESTI	IMMAGINI	SUONI

Strumenti di Automazione

Accesso Elaborazione Comunicazione	Query	Iptestesi	Movie Players	Players Browser Iptestesi	WWW
	DBMS	Editor	Computer Graphics	Campionatori	
	Fogli Elettronici	Word Processor	Sequencer	Groupware	
	Servizi di rete	E-mail	Fax	Teleconferenze	
	DATI	TESTI	IMMAGINI	SUONI	MULTIMEDIALI

Applicazioni

Midware Professionali Gestionali Tecniche	Sistemi di Pagamento Elettronico	Biblioteche Digitali	Commercio Elettronico	
	Posta Elettronica			
	Sistemi di Supporto Individuale			
	Decisionali Controllo Operative	Gestione Documentazione		
	Robotica Commutazione Misure	Sound Editing Hard Disk Recording Telerilevamento	Controllo Ambientale	
	DATI	TESTI	IMMAGINI	SUONI

Reperimento della Informazione

- Se la memorizzazione (mediante dispositivi di memoria di massa) e' estensiva (testi, immagini, suoni, ...) si pone il problema di **ricercare** tale informazione
- Il livello di astrazione consentito dal *File System* e' solo un primo livello di supporto:
 - e' insufficiente in molti casi (ad es. anagrafica)
 - non e' ottimale (riguardo alla velocita' della ricerca)

Reperimento della Informazione

- Ricercare in generale significa
 - definire i propri bisogni informativi
 - memorizzare i risultati
 - raffinare la propria selezione
 - ridefinire i requisiti informativi
 - “navigare” attraverso i dati trovati
 - elaborare, cioè’ combinare i dati di diverse ricerche

Reperimento della Informazione

- Strumenti di supporto alla ricerca di informazione
 - determinano (a priori) una strutturazione dell’informazione da ricercare che
 - rifletta il linguaggio di interrogazione
 - rifletta la natura (cioè’ il contenuto) dei dati da ricercare
 - faciliti le operazioni interne di ricerca
 - forniscono un linguaggio per la definizione dei bisogni informativi, detto linguaggio di interrogazione (*Query language*)

Strutturazione dell’Informazione

- Documenti e loro surrogati
 - Ad ogni documento d da archiviare e ricercare è associata una immagine detta surrogato che
 - esprima sistematicamente il contenuto di d
 - sintetizzi l’informazione
 - faciliti le operazioni di ricerca
 - Ogni surrogato fa riferimento (univocamente) a d
 - Ricercare d equivale a ricercare il suo surrogato

Strutturazione dell’Informazione

- Cataloghi Bibliotecari
- Sistemi di IR
 - Archiviazione per Autore e Titolo
 - Sistemi di Classificazione Gerarchici
 - Sistemi di Classificazione Tematica
 - Sistemi Analitico-Sintetici
 - Sistemi a Chiave

Reperimento della Informazione

- Es. Elenco Telefonico
 - La Struttura è l’ordinamento Alfabetico per Cognome, poi Nome
 - è coerente con la specifica degli elementi da ricercare
 - Riflette la natura del problema (Telefono => Persona)
 - Rende più’ veloce la ricerca (binaria e non enumerativa)
 - Il Query Language è in generale la tripla
 <Cognome, Nome, Indirizzo>

Reperimento della Informazione

- La fase di specifica dei bisogni informativi è detta Fase di Interrogazione al sistema
- La fase di ricerca deve essere preceduta da una fase di strutturazione della informazione obbiettivo (O) detta tradizionalmente Fase di Indicizzazione
- Viene cioè stabilita’ una funzione
$$ind : O \rightarrow I$$
con $ind(o_i) = \{ i_{i1}, \dots, i_{iN} \}$

Indici

- Gli indici stabiliscono una corrispondenza (!) tra gli elementi di O e il loro contenuto
- Gli *indici* hanno un ruolo
 - logico, esprimendo il contenuto dei dati (per es. i cognomi)
 - fisico poiché sono collegati al livello fisico della memorizzazione (per es. determinano la posizione nell'ordinamento alfabetico)
- Gli indici poi entrano a far parte del linguaggio di interrogazione

Ricerca

- La ricerca quindi procede per
 - la traduzione della query in una struttura di indici
 - un confronto tra gli indici della query e la strutturazione di O
 - un ordinamento di tutti gli elementi di O che hanno superato positivamente l'esito del confronto
- Nei linguaggi booleani la query è una composizione logica di indici
 - Es. "Basilì" AND "Roberto"
 - "Basilì" AND ("Roberto" OR "Roma")

Informazione Bibliografica

- A differenza della informazione in un database, l'informazione testuale è caratterizzata da una natura molto complessa
 - il contenuto è *sperso* attraverso il testo
 - nessuna parola è "*da sola*" capace di esprimere l'intero contenuto
 - le *relazioni logiche* espresse da un testo possono essere molto complesse

Informazione Bibliografica

- Es.
 - "Il Dipartimento di Letteratura si è occupato di Scienze della Comunicazione"
 - La letteratura dice che il Dipartimento di Scienze non ne fece mai comunicazione a ..."
 - Non tutte le parole sono significative
 - Alcune parole sono più importanti !
 - L'insieme delle parole non basta

Metodi per l'indicizzazione

- Metodi statistici
 - selezione di indici
 - pesatura degli indici
- Metodi basati su conoscenza o linguistici
- Metodi Ibridi
 - selezione linguistica
 - pesatura statistica

Metodi per l'indicizzazione

- Es. Selezione (statistica)
 - D1: La letteratura dice che il Dipartimento di Scienze non ne fece mai comunicazione a ..."
 - D2: Il Dipartimento di Letteratura si è occupato di Scienze della Comunicazione
- D1: { letteratura, Dipartimento, Scienze, comunicazione, fare, dire }
- D2: { Dipartimento, Letteratura, occupare, Scienze, Comunicazione }

Metodi per l'indicizzazione

- Es. Pesatura statistica
 - D1: La *letteratura* dice che il *Dipartimento di Scienze* non ne fece mai *comunicazione* a ..."
 - D2: Il *Dipartimento di Letteratura* si e' occupato di *Scienze della Comunicazione*

D1: { (*letteratura*,0.5),(*Dipartimento*,0.5), (*Scienze*,0.5)
(*comunicazione*,0.5), (*fare*,1.0), (*dire*,1.0) }

D2: {(*Dipartimento*,0.5), (*Letteratura*,0.5), (*occupare*,1.0),
(*Scienze*,0.5), (*Comunicazione*,0.5)}

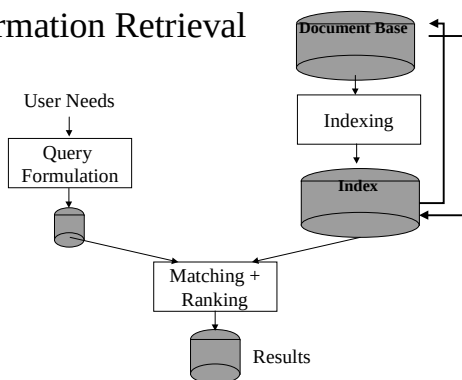
Metodi per l'indicizzazione

- Es. Selezione e pesatura basata su conoscenza
 - D1: La *letteratura* dice che il *Dipartimento di Scienze* non ne fece mai *comunicazione* a ..."
 - D2: Il *Dipartimento di Letteratura* si e' occupato di *Scienze della Comunicazione*

D1: { (*letteratura*,1.0), (*Dipartimento_di_Scienze*,1.0),
(*comunicazione*,1.0), (*fare*,0.5), (*dire*,0.5) }

D2: {(*Dipartimento_di_Letteratura*,1.0), (*occuparsi*,0.5),
(*Scienze_della_Comunicazione*,1.0)}

Information Retrieval



Information Retrieval

- I metodi di IR si diversificano per
 - metodo di selezione degli indici
 - metodo di pesatura degli indici in un documento
 - metodo di selezione dei documenti (matching)
 - metodo di presentazione

Information Retrieval

- Classificazione dei sistemi IR
 - Indicizzazione
 - manuale
 - statistica
 - basata su conoscenza

IR

- Classificazione dei sistemi IR
 - Pesatura numerica
 - frequenza
 - probabilistica
 - metodi empirici
 - Pesatura basata su conoscenza
 - indici come proprietà
 - regole di inferenza (es. alberi di decisione)

IR

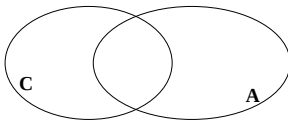
- Classificazione dei sistemi di IR
 - Metodi di selezione simbolici
 - risultato logico (*boolean*)
 - regole per la decisione relativa al matching
 - Metodi quantitativi
 - metodi bayesiani
 - vector-space model (Salton,79)

IR

- Valutazione delle prestazioni dei sistemi IR
 - Copertura dei fenomeni (Recall)
 - Quanti documenti corretti (cioè che erano da trovare) sono stati trovati?
 - Accuratezza (Precision)
 - Quanti documenti corretti tra quelli ritrovati?

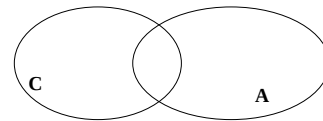
IR

- Precision and Recall di un metodo M
- Data una query Q , l'insieme dei documenti C corretti e l'insieme A dei documenti rilevati dal metodo M



Precision and Recall di un metodo M

- Data una query Q , l'insieme dei documenti C corretti e l'insieme A dei documenti rilevati dal metodo M

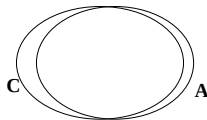


$$\text{recall} = |A \cap C| / |C|$$

$$\text{precision} = |A \cap C| / |A|$$

Precision and Recall di un metodo M

- Data una query Q , l'insieme dei documenti C corretti e l'insieme A dei documenti rilevati dal metodo M



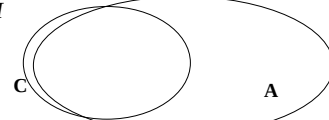
*Alta Recall e
Alta Precision*

$$\text{recall} = |A \cap C| / |C|$$

$$\text{precision} = |A \cap C| / |A|$$

Precision and Recall di un metodo M

- Data una query Q , l'insieme dei documenti C corretti e l'insieme A dei documenti rilevati dal metodo M



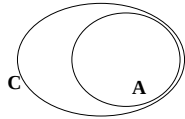
*Alta Recall ...
Bassa Precision*

$$\text{recall} = |A \cap C| / |C|$$

$$\text{precision} = |A \cap C| / |A|$$

Precision and Recall di un metodo M

- Data una query Q , l'insieme dei documenti C corretti e l'insieme A dei documenti rilevati dal metodo M



Massima Precision:
 $A \subset C$

$$\text{recall} = |A \cap C| / |C|$$

$$\text{precision} = |A \cap C| / |A|$$

Ricerca in Internet

- Internet e' una *meta rete* di calcolatori in grado di comunicare secondo protocolli indipendenti dallo specifico tipo di macchina
- Essa rappresenta una infrastruttura in grado di erogare diversi servizi
 - comunicazione dati (es. Ftp)
 - Posta Elettronica
 - accesso all'informazione multimediale (es. WWW)

Internet

- In quanto rende disponibile informazione distribuita geograficamente Internet rappresenta un enorme patrimonio informativo
 - Multimediale, anche se ...
 - La maggior parte della informazione e' disponibile in forma testuale

Internet

- Internet consente quindi l'accesso
 - ad archivi molto strutturati (data base e/o ipertesti)
 - ad estensive sorgenti testuali
 - poco strutturate
 - eterogenee (diversi autori, diversi linguaggi e sottolinguaggi)
 - dinamiche (cioe' continuamente in divenire)

Internet

- L'accesso ad informazione su Internet e' garantito da
 - Tecnologia
 - Architetture Client-Server
 - Browser Multimediali
 - Sistemi di IR
 - Servizi
 - Motori di Ricerca dedicati (Es. Altavista, Yahoo)
 - Mailing-list
 - News Group
 - Portali

Tecnologie Avanzate per l'accesso all'informazione su Internet

- I limiti degli strumenti attuali ...
- ... e l'esplosione delle potenziali applicazioni
 - Knowledge Management
 - Finanza On-line
 - Commercio Elettronico (E-Commerce)

spingono alla progettazione di nuovi strumenti

Tecnologie Avanzate per IR su Web

Obbiettivi

- Migliorare le prestazioni dei sistemi attuali
- Estendere il bacino di utenza
 - elevare la autonomia dell'utente
 - supportare la multilingualita'
 - diminuire il "peso" sull'utente
 - navigazione
 - *lost in the hyperspace*

Tecnologie Avanzate per IR su Web

Strumenti

- Cross-Linguistic Information Retrieval (CLIR)
- Apprendimento Automatico
- Data and Text Mining
- Nuovi Standard per la rappresentazione
 - Extended HTML
 - XML

Tecnologie Avanzate per IR su Web

Strumenti

Natural Language Processing

- Information Extraction
- Machine Learning of Natural Language
- Terminology
- Content-driven Retrieval