Storia Information Retrieval

Anita Alicante

SICSI VIII CILCO

INFORMATION RETRIEVAL
Agenda

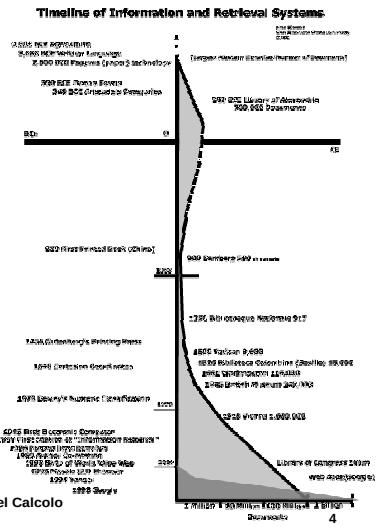
- Introduzione
- Definizione
- Genitori e Ideatori
- Evoluzione Storica dei Modelli per la costruzione di sistemi di IR
- Un'applicazione: Motori di Ricerca

INFORMATION RETRIEVAL
Introduzione

- **Information Retrieval**: ricerca (nel WEB e negli archivi aziendali) dei documenti che soddisfano una determinata esigenza
  - Possibilità di consultare le pratiche relative a casi in archivio in studi legali, assicurazioni, ma anche banche e aziende di servizi.
  - Il termine **Information Retrieval** fu coniato nel 1952 da Calvin Mooers che tra l'altro formulò la "legge di Mooers" che disse:
    - "Un sistema di reperimento delle informazioni tenderà a non essere usato quando trovare le informazioni è 'more painful and troublesome" (dice proprio così 'più noioso e doloroso') che non trovarle

INFORMATION RETRIEVAL
Introduzione

- **Information Retrieval** è un campo interdisciplinare che nasce dall'incrocio di discipline diverse.
- **Infomation Retrieval** coinvolge la psicologia cognitiva, l'architettura informativa, la filosofia (vedi la voce ontologia), il design, il comportamento umano sull'informazione, la linguistica, la semiotica, la scienza dell'informazione e l'informatica.
- Molte università e biblioteche pubbliche utilizzano sistemi di **Information Retrieval** per fornire accesso a pubblicazioni, libri ed altri documenti.



INFORMATION RETRIEVAL
Agenda

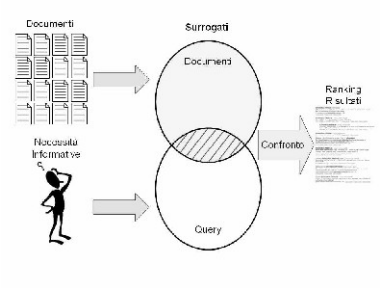
- Introduzione
- Definizione
- Genitori e Ideatori
- Evoluzione Storica dei Modelli per la costruzione di sistemi di IR
- Una applicazione: Motori di Ricerca

INFORMATION RETRIEVAL
Definizione

- Per recuperare l'informazione, i sistemi di IR usano i linguaggi di interrogazione basati su comandi testuali. Due concetti sono di fondamentale importanza, query ed oggetto:
  - Le query sono generalmente stringhe di parole-chiave rappresentanti l'informazione richiesta. Vengono digitate dall'utente in un sistema di IR
  - Un oggetto è un'entità che mantiene o racchiude informazioni in una banca dati. Un documento di testo, per esempio, è un oggetto di dati.
- A seguito di un'interrogazione, il sistema segnala il numero di documenti ritrovati e ordina i documenti per rilevanza.
- Un documento può essere rilevante o non rilevante per la query formulata dall'utente.

INFORMATION RETRIEVAL
Definizione

- L'esame dei documenti si avvale di due funzionalità:
  - RANKING: presentazione dei risultati in ordine decrescente di rilevanza, in funzione dei pesi assegnati ai termini. L'utente può farsi un'idea di quanto la ricerca sia efficace.
  - BROWSING: documenti raggruppati in classi di somiglianza, permettendo all'utente di sfogliarli secondo un ordine logico.



INFORMATION RETRIEVAL
Definizione

- **Precision(P)** valuta la capacità di trovare documenti rilevanti

$$\mathbf{P} = \frac{\#\,\text{documenti rilevanti recuperati}}{\#\,\text{documenti recuperati}}$$

- **Recall(R)** valuta la capacità di rigettare i documenti non rilevanti

$$\mathbf{R} = \frac{\#\,\text{documenti rilevanti recuperati}}{\#\,\text{documenti rilevanti}}$$

INFORMATION RETRIEVAL
Agenda

- Introduzione
- Definizione
- Genitori e Ideatori
- Evoluzione Storica dei Modelli per la costruzione di sistemi di IR
- Una applicazione: Motori di Ricerca

INFORMATION RETRIEVAL
Genitori e Ideatori (1°)

- Nel 1890 l'ingegnere statistico **H. Hollerith** aveva elaborato una tabulatrice riutilizzando l'idea delle schede perforate di **Babbage**, (questa volta però non per specificare il programma, ma i dati da elaborare o i risultati dell'elaborazione)
- Prima macchina usata classificare e contare automaticamente i dati del censimento americano

INFORMATION RETRIEVAL
Genitori e Ideatori (2°)

- Nell'articolo **As We May Think** del 1945, **Vannevar Bush** prospetta una apparecchiatura futuribile con la quale uno studioso possa raccogliere e organizzare i vari testi che possono servire per i suoi studi e la chiama **Memex**, abbreviando memory expansion. In questo articolo predice:
  - Compariranno nuovi tipi di enciclopedie confezionate con una rete di percorsi associativi che le collegano, pronte ad essere inserite in memex e qui ampliate.
- Il **Memex** era un calcolatore analogico dotato di un sistema di archiviazione, ideato dallo scienziato e tecnologo statunitense negli anni trenta e mai realizzato, da molti considerato il precursore del personal computer e degli ipertesti.
- Il suo progetto del Memex era partito da un'esigenza concreta, dalla ricerca della soluzione di un problema specifico:
  - rendere più efficiente l'archiviazione ed il reperimento del sapere, dato che l'informazione, nelle università, organizzata in biblioteche risultava spesso di difficile o impossibile accesso.

INFORMATION RETRIEVAL
Genitori e Ideatori (4°)

- Nell' articolo **The automatic creation of literature abstract**, pubblicato nell'IBM Journal nell'aprile del 1958, **Luhn** affermava che:
  - la **frequenza con cui alcune parole compaiono in un testo** forniscono un parametro importante del significato delle parole.
  - il **posizionamento di queste parole all'interno delle frasi** indica il significato e quindi l'importanza delle frasi.

  La frequenza con cui alcune parole compaiono in un testo, può essere usata per rappresentare un documento.
- Questi saranno i principi di base all'**indicizzazione automatica** dei testi

INFORMATION RETRIEVAL
Genitori e Ideatori (5°)

- Dopo di Mooers, **Gerard Salton** nel 1960 sviluppò un sistema di recupero d'informazione denominandolo **SMART**, grazie al quale vennero sviluppati molti importanti concetti sui quali, oggi, si basa l'attuale disciplina di Recupero d'Informazione .
- Due concetti fondamentali introdotti da Salton su cui si basa (ancora oggi) IR
  - Nel modello di spazio vettoriale (vector space model) per il recupero di informazioni i documenti sono modellati come vettori costruiti sulla base dei termini e dell'importanza dei termini nel documento.
  - Nel modello dei risultati pertinenti (relevance feedback), le informazioni ottenute dalla prima query sono utilizzate, in base alla loro pertinenza, per eseguire un'altra query.

INFORMATION RETRIEVAL
Agenda

- Introduzione
- Definizione
- Genitori e Ideatori
- Evoluzione Storica dei Modelli per la costruzione di sistemi di IR
- Una applicazione: Motori di Ricerca

INFOTMATION RETRIEVAL
Evoluzione Storica dei Modelli per la costruzione di sistemi di IR

- Negli anni 50 e 60 sono stati fatti la maggior parte degli esperimenti per costruire i sistemi di Information Retrieval.
- Molti degli attuali sistemi di librerie commerciali come **Dialog** e **BRS** sono stai ideati grazie agli esperimenti eseguiti in questi anni.
- In questi anni nascono anche le definizione di recall e precision per la valutazione delle prestazione degli IRS.
- La prima generazione di ricerca basata su Information Retrieval risale, infatti, agli anni 60, ed era dominata da costruzione di modelli, sperimentazione ed euristiche. I personaggi più illustri di questo processo furono **Gerry Salton** e **Karen Sparck Jones**.
- Questi Sistemi erano basati sul **Modello Booleano**.

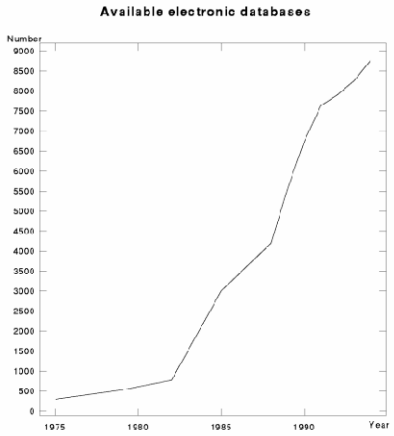
INFOTMATION RETRIEVAL
Evoluzione Storica dei Modelli per la costruzione di sistemi di IR

- Il secondo periodo, che inizia a metà degli anni 70, mostra uno spostamento verso la matematica e una crescita del modello di Information Retrieval basato sulla teoria probabilistica (**Modello probabilistico**). Il nome più famoso legato a questa teoria fu **Stephen Robertson**.
- In questo periodo i sistemi di IRS diventano maturi. Si cominciavano a scrivere **molti documenti in formato elettronico** (input del IRS) e nascevano **anche i sistemi time-sharing** (risposta immediata alle query)

Stephen Robertson, oggi, è ricercatore per la Microsoft Cambridge dall'Aprile del 1998. Negli 1970 era ricercatore nella University College London

INFOTMATION RETRIEVAL
Evoluzione Storica dei Modelli per la costruzione di sistemi di IR

- Negli anni 80 si è capito che non è solo un termine che caratterizza un documento ma diversi termini, formano il contesto semantico in cui il documento stesso è immerso. Per questo si affiancano al **modello probabilistico** una sua evoluzione basata sul **modello del linguaggio (linguista computazionale)**
- **Altro Incentivo per IR**: Crescita Esponenziale dei database online (input dei sistemi IRS)



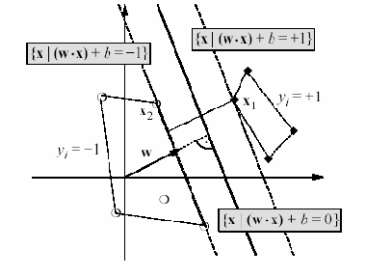
INFOTMATION RETRIEVAL
Evoluzione Storica dei Modelli per la costruzione di sistemi di IR

- Più recentemente **Keith van Rijsbergen** ha coordinato un gruppo che ha sviluppato i **modelli logici** di fondo dell'Information Retrieval.
- Negli ultimi anni è stato introdotto il **modello di spazio vettoriale** (vector space model) per il recupero di informazioni. In questo modello i documenti sono modellati come **vettori (features)** costruiti sulla base dei termini e dell'importanza dei termini nel documento.
  - Negli utlimi anni si rappresentano sempre più dati semi strutturati utilizzando il linguaggio di Markup XML (Extensible Markup Language) XML è usato per contenuti web, per i documenti scritti con programmi di video scrittura (MSOffice o Open Office), per l'importazione e l'esportazione di testo contenuto in generale, e molte altre applicazioni.

Nel 1980 docente di informatica all'University College Dublin; nel 1986 si è trasferito alla Glasgow University dove ancora oggi lavora.

INFOTMATION RETRIEVAL
Evoluzione Storica dei Modelli per la costruzione di sistemi di IR

- **Support Vector Machine (SVMs)**
  - Usate con successo in molti problemi di apprendimento automatico.
  - Basate su tecniche provenienti dalla teoria dell'ottimizzazione.
  - Ricercano un iperpiano che separa un insieme di dati in due distinte classi e massimizzando il margine tra esse.



INFOTMATION RETRIEVAL
Evoluzione Storica dei Modelli per la costruzione di sistemi di IR

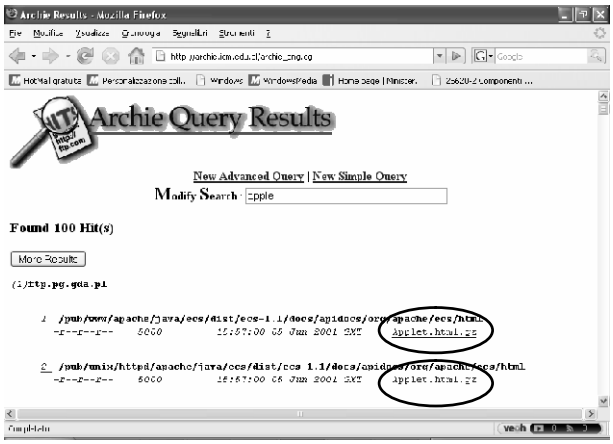
- Prime implementazione di sistemi di Information Retrieval System (IRS) sono stati introdotti negli anni 50 e 60.
- Negli anni 90 IRS iniziano a fornire risultati accetabili su piccoli corpora di testi (alcune migliaia di documenti)
- Negli **1990** nascita dei motori di Ricerca **(Google, Yahoo ecc…)**
  - Alcuni producono i risultati in relazione alla frequenza con cui determinate parole chiave che compaiono all'interno di ogni sito web
  - Altri motori di ricerca si basano su algoritmi di ranking
- Il primo tool per ricerche su internet fu chiamato **Archie**. Il nome viene dal "archive" senza la "v". Fu creato da **Alan Ematage** uno studente nel 1990 dell'**Università McGill** di **Montreal**.
  - Pur con caratteristiche molto diverse dagli attuali strumenti disponibili per la ricerca on line, questo antenato dei motori permetteva di cercare nella rete per nome di file (Archie non analizzava infatti il contenuto testuale di un documento).

INFORMATION RETRIEVAL
Agenda

- Introduzione
- Definizione
- Genitori e Ideatori
- Evoluzione Storica dei Modelli per la costruzione di sistemi di IR
- Una applicazione: Motori di Ricerca

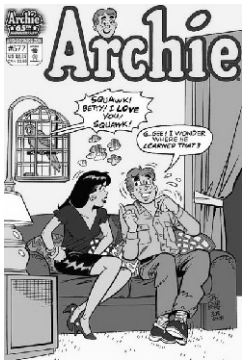
INFORMATION RETRIEVAL:
Un'applicazione: Motore di ricerca

INFORMATION RETRIEVAL:
Un'applicazione: Motore di ricerca



INFORMATION RETRIEVAL:
Un'applicazione: Motore di ricerca

- Nel 1991 vengono creati due nuovi motori di ricerca: **Veronica (Very Easy Rodentia-Oriented Net-wide Index to Computerized Archives)** e **Jughead (Jonzy's Universal Gopher Gerarchia Scavo E Display)** servizi del sistema Gopher per la gestione dei documenti.
  - **Veronica** ha fornito una ricerca per parola chiave
  - **Jughead** è stato uno strumento per ottenere informazioni dal menu Gopher server specifico
- Mentre il nome del motore di ricerca "Archie" non è stato un riferimento al Archie libro fumetti della serie, "Veronica" e "Jughead" sono personaggi della serie, quindi riferimento a loro predecessore.



INFORMATION RETRIEVAL:
Un'applicazione: Motore di ricerca

- Il primo vero motore di ricerca, che funzionava con un meccanismo di indicizzazione automatica dei documenti, è stato chiamato **Wandex** creato da un web developer **Matthew Gray** della MIT nel 1993.
- Nel 1993 **Martijn Koster** creò **ALIWEB** (Archie Like Indexing for the Web) il primo motore di ricerca per il Web.
- ALIWEB, nonostante permettesse l'inserimento da parte degli utenti della descrizione delle proprie pagine web e dell'inserimento delle parole chiave, non ebbe molta fortuna a causa della difficoltà di utilizzo da parte degli utenti.
- Anche **Lycos** nel 1994 (in collaborazione con l'Università di Carnegie Mellon) lanciò il suo motore di ricerca, il primo motore di ricerca creato a scopo commerciale.

INFORMATION RETRIEVAL:
Un'applicazione: Motore di ricerca

- In poco tempo si svilupparano molti motori di ricerca **Excite, Infoseek, Inktomi, Northern Light, and AltaVista.**
- **Yahoo!** È stato tra i più popolari modi per le persone a trovar le pagine web di interesse, ma la sua funzione di ricerca si basa directory web, piuttosto sull'analisi del testo integrale delle pagine web.
- Nell' 1998 nasce **Google. Larry Page** e **Sergey Brin** fondarono **Google** nel settembre del 1998. Oggi Google, a meno di dieci anni dalla sua apparizione, è il più grande Motore di Ricerca del mondo.
- **Google** usa una formula matematica denominata **PageRank** per giudicare la rilevanza delle pagine per una determinata ricerca. La formula del **PageRank** viene applicata attraverso un algoritmo che valuta tutti i siti collegati una pagina Web e assegna loro un valore, basato in parte sui siti ad essi collegati.

INFORMATION RETRIEVAL
I più popolari motori di ricerca Dec. 2007
http://www.comscore.com/press/release.asp?press=2018

Company	Millions of searches	Relative market share
Google	28,454	46.47%
Yahoo!	10,505	17.16%
Baidu (search engine in the People's Republic of China)	8,428	13.76%
Microsoft	7,880	12.87%
NHN	2,882	4.71%
eBay	2,428	3.9%
Time Warner (includes AOL)	1,062	1.6%
Ask.com and related	728	1.1%
Yandex	566	0.9%
Alibaba.com	531	0.8%
Total	**61,221**	**100.0%**

INFORMATION RETRIEVAL
Bibliografia

- Chris Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2007
- Ricardo Baeza-Yates e Berthier Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999
- Measuring in Digital World (http://www.comscore.com/press/release.asp?press=2018)
- H.P Luhn, The Automatic Creation of Literature Abstracts, IBM Journal of Research Development, 1958 (http://courses.ischool.berkeley.edu/i256/f06/papers/luhn58.pdf)
- Vannevar Bush**,** As We May Think, The Atlantic Monthly, 1945
- http://www.ifla.org/VI/5/op/udtop5/dbs.gif
- Hedvah L. Schuchman Information transfer in engineeringReport 461-46-27 The Futures Group (1981). ISBN 0-9605196-0-2.