Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2015/2016

Corso di Laurea Magistrale in
Editoria e Giornalismo

Sistemi per il recupero delle informazioni sul Web

Introduzione

- Il Web
  - Negli ultimi anni ruota attorno ai motori di ricerca
  - Presente ovunque in modo distribuito e perciò enorme, pubblico e non strutturato
  - Necessità di strumenti appositi per la gestione e il recupero delle informazioni
  - Altamente dinamico, nei contenuti e nella struttura

Principali difficoltà (I)

- Le difficoltà che i SRI sul Web devono affrontare si possono classificare in due classi
  - Data-centric
  - Interaction-centric
- Data-Centric: legati ai dati e ai contenuti
  - Dati distribuiti
  - Dati altamente dinamici e volatili
  - Grande quantità di dati
  - Dati ridondanti e non strutturati
  - Varia qualità dei dati
  - Dati eterogenei

Principali difficoltà (II)

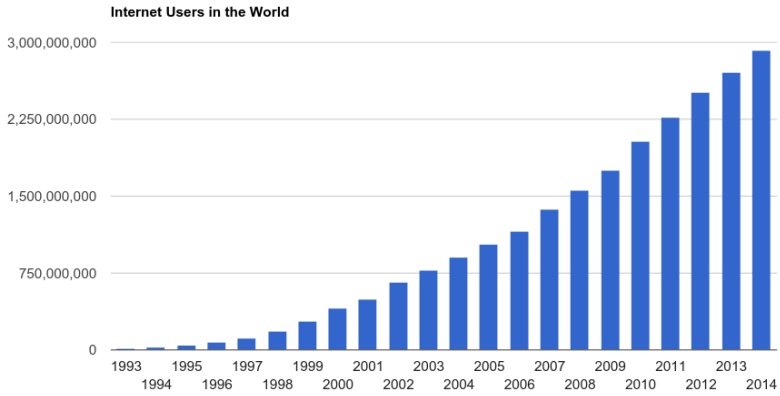
- Interaction-centric: legati agli utenti e alle loro interazioni
  - Come esprimere le query
  - Come rendere facile scrivere buone query
  - Come interpretare i risultati
  - Eseguire le ricerche velocemente e ritornare dei buoni risultati
    - Anche nel caso di query espresse male

Il Web

- Diverse ricerche hanno studiato proprietà e caratteristiche di sottoinsiemi del Web
  - Manca una comprensione completa del Web e delle sue dinamiche
  - Il web è una rete
    - Ha alcune proprietà e caratteristiche comuni a tutte le reti (umane e non), come ad es. i social network

Il Web: alcuni numeri (I)

- ~3G internet users
  - 2,6G email users
    - 4,3G email accounts
    - 900M GMail accounts
  - 3,6G social network account
    - 1,4G FB active acc.s
    - 300M G+ active acc.s
    - 316M Twitter active acc.s
  - 2,2G smartphones



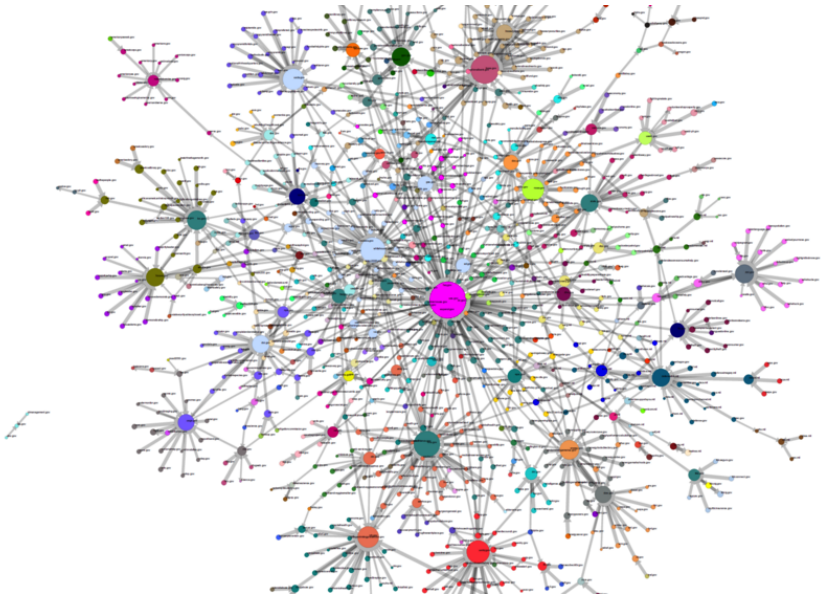
M: milioni; G: miliardi

Il Web: alcuni numeri (II)

- 1,1G siti web
- 1,8G Foto condivise/giorno
  - 760M on snapchat
  - 300M on FB
  - 700M on WhatsApp
  - 245M on Instagram
- Ricerche
  - 1608G ricerche su Google all'anno (~51000 r/sec)
  - Google: 90% del mercato globale, 92% del mercato mobile, 68% del mercato desktop (settembre 2015)

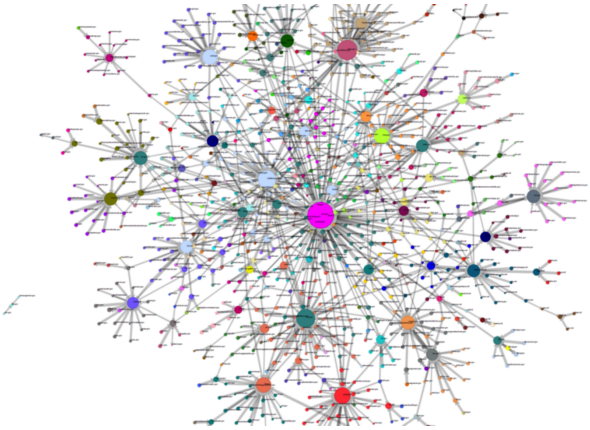
Struttura del web (I)

- Il web può essere rappresentato come un grafo
  - I nodi rappresentano singole pagine
  - Gli archi rappresentano i link tra le pagine



Struttura del web (II)

- La distanza media tra 2 pagine qualsiasi è 19 (nel 1999 ☹)
  - Teoria del piccolo mondo
  - http://en.wikipedia.org/wiki/Small_world_experiment
  - M. Buchanan. *Nexus - Perché la natura, la società, l'economia, la comunicazione funzionano allo stesso modo*. Mondadori.



Struttura del web (III)

- Secondo uno studio del 2000 la struttura del web ad alto livello ha la forma di un fiocco

SCC: insieme delle pagine che si possono raggiungere tutte l'un l'altra

- Una minoranza delle pagine

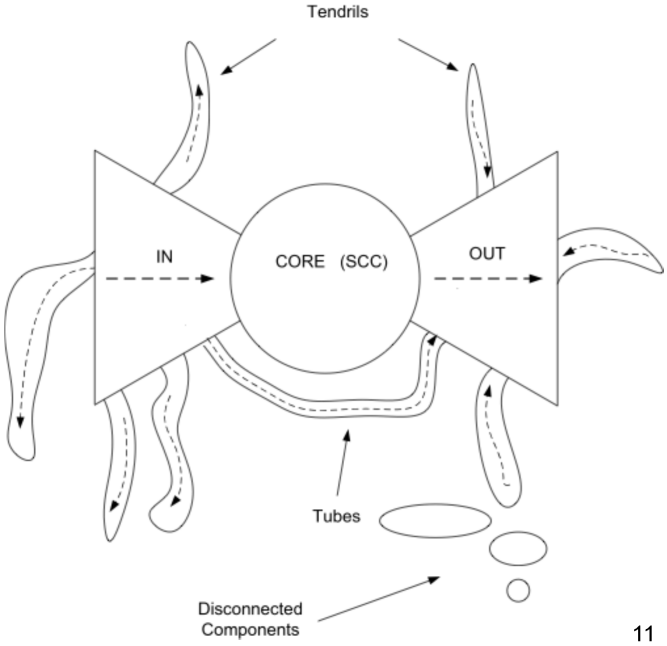
IN: insieme delle pagine che possono raggiungere le pagine in SCC ma che non possono essere raggiunte

- Nuovi siti che non sono ancora stati "scoperti" e linkati

OUT: insieme delle pagine che possono essere raggiunte dall'SCC ma che non lo possono raggiungere

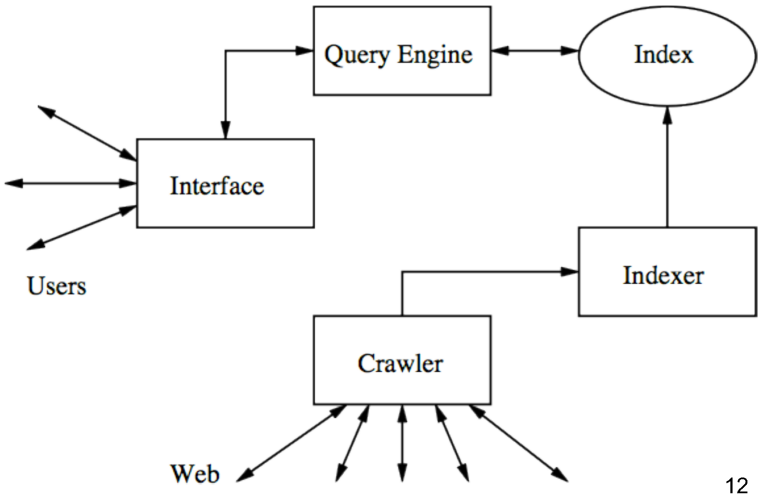
- Siti aziendali, con soli link interni

Tendrils: insieme delle pagine che non possono raggiungere né essere raggiunte dall'SCC



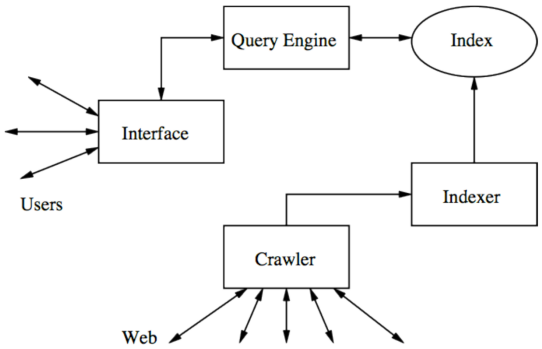
Architettura dei motori di ricerca (I)

- **Architettura centralizzata crawler-indexer**
  - Usata dalla maggior parte dei motori di ricerca
- I crawler sono programmi che navigano in automatico il Web seguendo i link copiando le pagine che trovano
- L'indexer legge le pagine recuperate dai crawler e le indicizza (nel modo già visto)



Architettura dei motori di ricerca (II)

- L'indice è usato centralmente (sul server) per rispondere alle query
  - Senza bisogno di accedere alle pagine originali
  - Per generare eventuali surrogati si usano copie locali delle pagine
  - Indici invertiti posizionali con frasi frequenti indicizzate come termini unici
- L'indice può essere affiancato ad alcuni metadati
  - Data di creazione della pagina, titolo, dimensione, …



Architettura dei motori di ricerca (III)

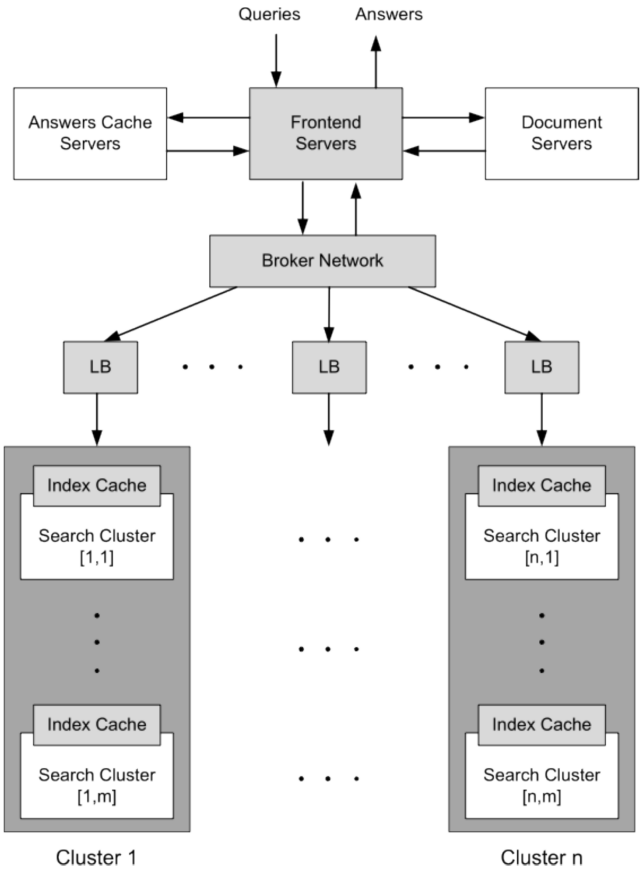
- I risultati di una query possono essere migliaia o milioni
  - Il loro calcolo completo può
    - richiedere troppo tempo
    - essere inutile in quanto solitamente gli utenti leggono solo una pagina (o poco più)
- I motori di ricerca non calcolano tutti i risultati
  - Calcolare solo i primi risultati
  - I risultati successivi sono calcolati su richiesta

Architettura dei motori di ricerca (IV)

- L'architettura è leggermente modificata al fine di migliorare l'efficienza: **architettura cluster-based**
  - Indici e crawler sono replicati su diversi server
    - Dislocati in diversi luoghi del pianeta
      - Riduce i tempi di comunicazione e attesa
    - Le query sono inviate al server meno occupato
    - La ridondanza riduce i problemi dovuti ad eventuali problemi software o hardware
  - Google si stima abbia 19 centri negli USA, 12 in Europa, 1 in Russia, 1 in Sud-America e 3 in Asia

http://www.datacenterknowledge.com

Architettura dei motori di ricerca (V)



Caching (I)

- Un'altra tecnica per aumentare la velocità e l'efficienza è l'uso della cache
  - Si salvano nella memoria principale i risultati di alcune query
    - Riduce il lavoro richiesto ai server
    - Riduce il tempo di risposta
    - Riduce la quantità di dati trasmessi in rete

Caching (II)

- Quali query salvare nella cache?
- Le query seguono una distribuzione esponenziale
  - Poche query eseguite tante volte
  - Tante query eseguite poche volte
- Salviamo le query più richieste
  - Con una “piccola” cache si può rispondere ad un gran numero di richieste
  - Se la cache è usata il 30% delle volte l'efficienza dei motori aumenta almeno del 43%
- Si possono anche salvare le ultime query eseguite
- Diverse altre strategie sono state proposte

Ranking

- È la più importante e difficile funzionalità di un motore di ricerca
  - Valutare adeguatamente la rilevanza dei risultati rispetto all'utente
  - Valutare la qualità dei contenuti nel web
  - Evitare, prevenire e gestire il web spam
  - Definire una funzione di ranking

Parametri per il ranking (I)

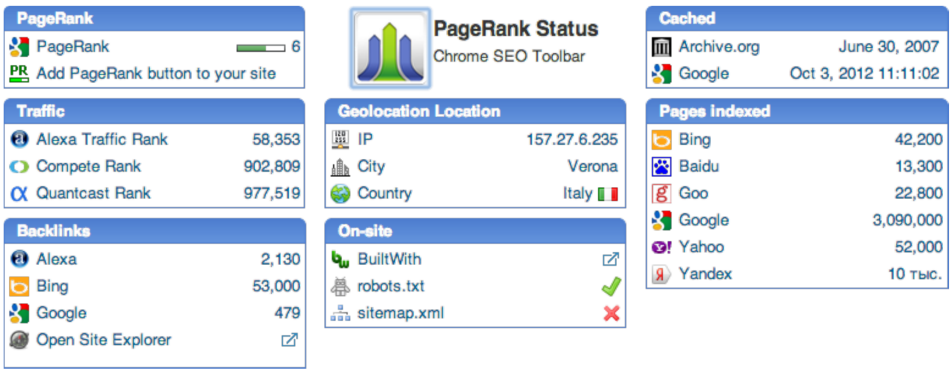
- Diversi parametri possono essere usati per valutare la rilevanza dei risultati, basati su
  - Contenuto
  - Struttura
  - Utilizzo
- Parametri basati sul contenuto: correlati al testo contenuto nella pagina
  - Numero di termini della query che vi compaiono
  - Posizione dei termini nella pagina (titolo,header, …)
  - Prossimità dei termini della query nella pagina

Parametri per il ranking (II)

- Parametri basati sulla struttura: derivano dalla natura “hyperlinked” del web
  - Testo dei link
  - Numero e qualità dei link entranti (pagine che puntano alla pagina in esame) ed uscenti (pagine a cui la pagina in esame punta)
- Parametri basati sull'utilizzo: tengono conto di come l'utente utilizza il e interagisce nel web
  - Numero di click sulle diverse pagine
  - Storia delle pagine visitate dall'utente
  - Posizione geografica e lingua dell'utente

PageRank (I)

- Il primo algoritmo di ranking più utilizzato è stato PageRank (®Google): basato sui link
  - Il numero di link che puntano ad una pagina “misurano” la sua popolarità e qualità
  - “pagine di qualità puntano a pagine di qualità”
- Alcune estensioni e servizi online permettono di calcolare il PageRank di un sito



PageRank (II)

$$PR(a) = \frac{q}{T} + (1-q)\sum_{i=1}^{n} \frac{PR(p_i)}{L(p_i)}$$

- Dove:
  - $L(p)$ numero di link nella pagina p
  - $p_1,...,p_n$ pagine che puntano alla pagina a
  - q: probabilità che l'utente non segua un link nella pagina (inizialmente 0.15)
    - $1\text{-}q$: probabilità che l'utente segua un link nella pagina
  - $PR(a)$: PageRank della pagina a
  - T: numero totale di pagine nella collezione (nel web)
- all'aumentare del numero di link complessivi dei siti che puntano ad a il PageRank di a aumenta
- Più il PageRank dei siti che puntano ad a è alto, più il PageRank di a è alto

NB: la formula non è richiesta all'esame, il suo "significato" sì invece

PageRank (III)

- Altri utilizzi:
  - Per misurare l'importanza di una rivista scientifica
  - Usato per "predire" il numero di veicoli/persone che andranno in una via/luogo
  - Usato per determinare quali specie animali in un ecosistema sono essenziali al benessere ambientale
- Problemi di PageRank
  - "facilmente manipolabile", ad esempio, era (è?) possibile comprare da alcune aziende link su siti con alto PageRank in modo da ottenere un PageRank migliore
  - Basato sul **random user model**
    - Un utente clicca in modo casuale su un link (con probabilità q)
    - Gli utenti non si comportano in modo casuale!!

Oltre PageRank (I)

- Per risolvere i problemi di PageRank esso è stato modificato e usato insieme ad altri metodi di ranking
- **Intentional user model**:
  - Un utente non naviga casualmente nel web
  - Si calcola l'importanza di una pagina osservando quanti utenti la visitano
    - Google Toolbar e Chrome, MS Internet Explorer, ecc... registrano e inviano il comportamento degli utenti a tal fine

Oltre PageRank (II)

- Febbraio 2011: “Panda” → valutare in modo più completo la qualità di una pagina anche basandosi sul suo contenuto e su ciò che pensano gli utenti
- Gennaio 2012: “page layout algorithm” → valutare la qualità di una pagina in base al suo layout e a quanta informazione è direttamente visibile appena la pagina appare
- Aprile 2012: “Pinguin” → ridurre il ranking dei siti che cercano di “frodare” i motori di ricerca per ottenere un ranking migliore
- Settembre 2013: “Hummingbird” → tenere maggiormente conto dell'intera query anziché delle singole parole, per andare incontro anche al sempre maggior uso della ricerca vocale
- Il ranking di una pagina è quindi calcolato sulla base di un mix di molte diverse sue caratteristiche
  - Bing ne utilizza più di 1000
  - Google ne utilizza più di 200

Valutazione delle performance (I)

- Usiamo Precisione e Richiamo?
  - Il richiamo è normalmente difficilmente calcolabile, nel web è impossibile
  - I grafici precisione-richiamo non possono essere costruiti
  - Gli utenti di solito leggono solo la 1ª (2ª) pagina dei risultati → 10-20 risultati
  - Le query tendono ad essere vaghe
    - Difficile valutare la rilevanza dei risultati

Valutazione delle performance (II)

- Quindi?
  - Misuriamo la precisione solo considerando i primi 5, 10, 20 risultati: P@5, P@10, P@20
  - Ogni coppia query-risultati deve essere valutata indipendentemente da diversi utenti

Valutazione delle performance (III)

- Altra possibilità, valutiamo i risultati in base al comportamento degli utenti
  - Rimane tanto tempo sulla pagina su cui ha cliccato?
    - Magari ha trovato qualcosa di interessante
  - Salta da un risultato ad un altro?
    - I risultati non sono soddisfacenti
  - È tornato ai risultati pochissimo dopo aver cliccato su uno di questi?
    - Web spam?
- Metodo complesso e tenuto "segreto"

Web spam

- Detto anche spamdexing, search spam, search engine spam
- Azioni il cui fine è l'acquisizione di visibilità nei motori di ricerca utilizzando metodologie e/o tecniche ritenute illecite o comunque apertamente in contrasto con i termini d'uso dei motori di ricerca
- Perché?
  - Perché più visibilità significa più visitatori, più visibilità/click della pubblicità sulla propria pagina, quindi più introiti

Web spam: tecniche

- Esistono diverse tecniche
- Qualche idea?

Web spam: tecniche (II)

- Due principali tipi di tecniche
  - Content spam
    - Puntano a modificare la percezione della pagina web da parte dei motori di ricerca
  - Link spam
    - Uso di link per aumentare il proprio ranking secondo gli algoritmi basati su essi (PageRank)

Content spam (I)

- Keyword stuffing:
  - Eccesivo ricorso alle parole chiave nel testo di una pagina
  - Caduto in disuso
  - I motori di ricerca analizzano la distribuzione delle parole chiave (keyword pattern) e valutano se è o no “fraudolenta”

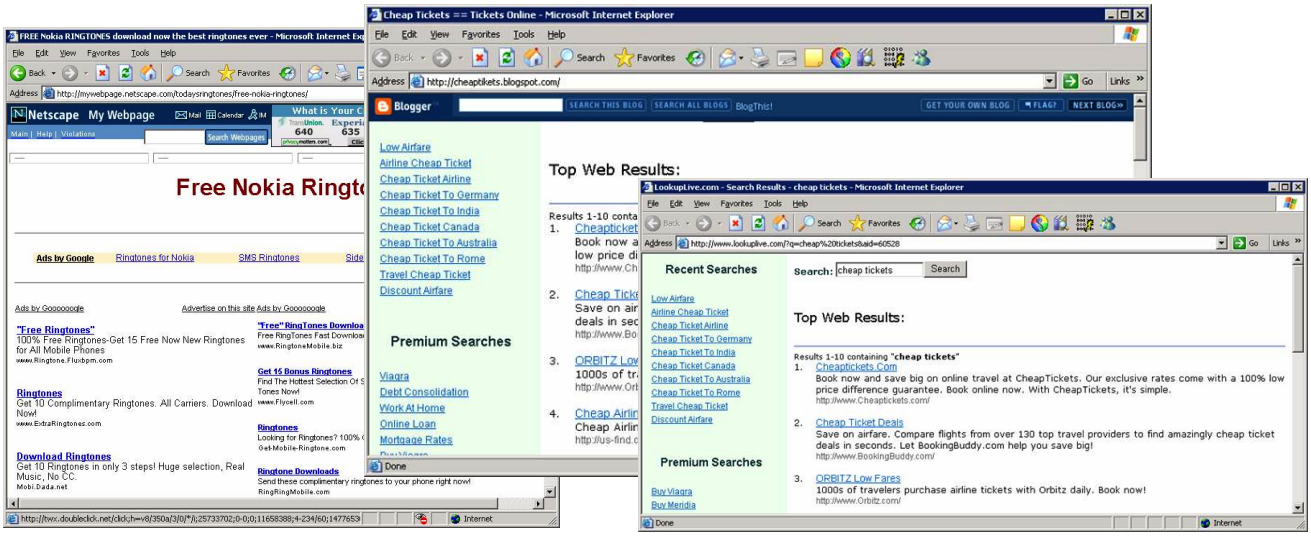
L’azienda joy produce **camicie** di alta qualità. Le nostre **camicie** rispettano i più alti livelli qualitativi. **Camicie** uniche. **Camicie** di qualsiasi colore. Produzioni di **camicie**. Vendiamo **camicie** per cerimonia. **Camicie** in lino. **Camicie** in

Content spam (II)

- Testo nascosto o invisibile
  - Testo dello stesso colore dello sfondo
  - Testo troppo piccolo
  - Testo nascosto nel sorgente della pagina ma reso invisibile
  - Invisibile agli utenti, ma visibile ai crawler/indexer

Content spam (III)

- Doorway pages
  - Pagine costruite appositamente per attrarre i motori di ricerca
  - Redirigono automaticamente l'utente su un'altra pagina o caricano il contenuto da un'altra pagina



Content spam (IV)

- Cloaking:
  - Evoluzione delle doorway pages per superare i controlli dei motori di ricerca
  - Mostrano un certo contenuto ai motori di ricerca
    - In modo da essere “ben indicizzati”
  - Mostrano un contenuto diverso (pura pubblicità) agli utenti

Link spam

- Link farm
  - Siti di una “comunità” che puntano l'una all'altra
- Blog spam
  - Simile al link farm
  - Creazione di blog “vuoti” con soli link verso una pagina target (che si vuole posizionare meglio)
- Using world-writable pages
  - Siti pubblicamente modificabili possono essere usati per inserirvi link ad una pagina target
    - Nei commenti/post di un blog/forum
    - In pagine delle wiki

Filtri contro la spamdexing

- Al fine di evitare lo spam e fornire ai propri “clienti” risultati attendibili e di qualità i motori di ricerca hanno introdotto nei propri crawler e algoritmi di indicizzazione dei filtri per individuare eventuale spam
  - Introdotti un po' alla volta nel tempo, all'apparire di un nuovo tipo di spam o di ricerca

Penalizzazione

- Quando il crawler identifica un possibile caso di spam, il motore di ricerca avverte il proprietario del sito
- Eventuali comportamenti non leciti vengono “puniti” dai motori di ricerca con diversi tipi di penalizzazione
  - Esclusione dagli indici
    - Permanente per pagine completamente fraudolente
    - Temporanea per siti “veri” ma non conformi alle regole del motore di ricerca
  - Penalità nella posizione
    - Limite minimo (e.g., almeno 30°) alla posizione di un sito nei risultati delle ricerche