

La Tecnologie per le Ricerche su Internet



Moreno Marzolla
INFN Sezione di Padova
moreno.marzolla@pd.infn.it
<http://www.dsi.unive.it/~marzolla>

Ringraziamenti

- prof. Francesco Dalla Libera
 - Corso di Commercio Elettronico, Dipartimento di Informatica, Università Ca' Foscari di Venezia.

Scegliere?

- Si può comperare un cellulare Siemens:
 - direttamente dalla casa madre per €250
 - da un negozio elettronico (buy.com \$224)
 - in un'asta (ebay, ~ centinaia di offerte)
 - da un aggregator (costa \$230 adesso, ma il prezzo scenderà a \$218 se ne ordini almeno 10)

Come effettuare queste scelte?

- *Directories*
 - Organizzate da operatori umani
 - Yahoo, Open Directory Project, About.com, LookSmart
 - Molto laboriose!
- *Ricerche su internet*
 - Purché i dati non siano nascosti dietro a form
 - Purché le informazioni non richiedano autorizzazioni
 - Purché le pagine WEB non chiedano espressamente di non essere indicizzate dai robot
 - Purché le informazioni siano testuali (niente immagini, suoni ecc.)

Le tecnologie

- Strumenti per individuare informazioni su Web
 - Problemi: database “nascosti”, ad es. La Repubblica
- Indice Sistemático (directory)
 - Un elenco, organizzato a mano, di gerarchie di concetti (Yahoo)
 - simile alle Pagine Gialle
- Motore di ricerca (search engine)
 - Un indice, costruito automaticamente (di solito per parole-chiave)

Indici sistemáticos

- Organizzazione di tutta la conoscenza in una qualche struttura e classificazione delle singole pagine web secondo questa struttura
- Yahoo è l'esempio principale di un indice sistemático
- Problemi:
 - La classificazione è un attività molto “faticosa”: ci sono molte più persone che pubblicano su Web che persone che classificano
 - Che fare se l'informazione cercata non è presente nella classificazione?

Indici sistemáticos

- Un indice, come Yahoo, dipende da persone fisiche:
 - Si può inviare una breve descrizione per un proprio sito all'indice
 - Un editore scrive tali rassegne per i siti che visita
- Con una ricerca si esaminano le “concordanze” solo con le descrizioni inviate
- Se le pagine web cambiano di contenuto, questo non ha riflesso sull'indice

Dimensioni

Service	Type	Editors	Cats	Links...	As Of
Open Directory	D	36,000	361,000	2.6 million	4/01
LookSmart	D	200	200,000	2 million	8/00
Yahoo	D	100+	n/a	1.5 to 1.8 million	8/00
NBCi (Snap)	D	30	80,000	1.5 million	12/00
Askjeeves	AS	150	n/a	128 million	3/01
AltaVista	SE	See LookSmart			
Excite	SE	See LookSmart			
HotBot	SE	See Open Directory			
Lycos	D	See Open Directory			
MSN Search	SE	See LookSmart			
Netscape	SE	See Open Directory			

Legenda

- **Type:**
 - Mostra se il servizio è una directory (D) o un motore di ricerca (SE).
- **Editors:**
 - Mostra quante persone sono coinvolte nella produzione dell'indice. Tante persone non significa necessariamente un miglior indice. Però tanti editori sono un segno di qualità per una directory
- **Cats:**
 - Quante categorie ci sono in ogni directory
- **Links:**
 - Quanti link esistono nella directory.

Motori di ricerca

- Un motore di ricerca crea i propri elenchi (indici) automaticamente, senza intervento umano.
- Un motore attraversa periodicamente la rete, recupera e indicizza le pagine, gli utenti quindi ricercano usando gli indici così costruiti
 - Se una pagina cambia, il motore (prima o poi) troverà questi cambiamenti e modificherà i suoi elenchi
- **Esempi:**
 - AltaVista, Google, Excite, Infoseek, Lycos, HotBot

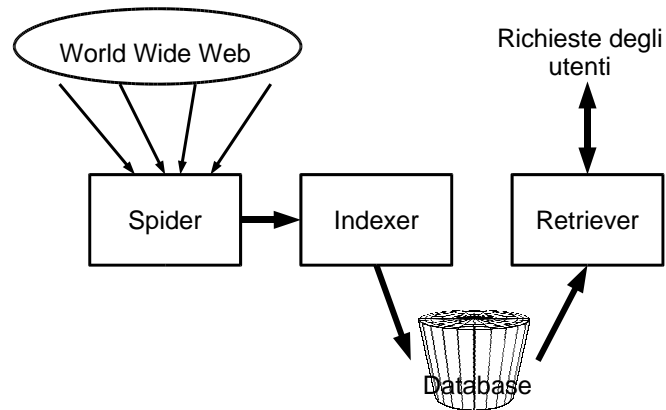
Componenti di un motore di ricerca

- Un "robot" (o "spider") che esplora la rete web e recupera le pagine
- Un database di informazioni aggiornato a partire dal contenuto di queste pagine
 - Interrogare questo database con le query utente e ordinare i risultati trovati
- Una interfaccia utente per creare le query e per presentare i risultati
- **Problema: l'informazione testuale è mal strutturata**

Architettura del motore di ricerca

- **Spider**
 - attraversa la rete per recuperare pagine. Segue i link presenti. Non si ferma mai
- **Indexer**
 - Produce le strutture dati per una ricerca veloce delle parole contenute nelle pagine
- **Retriever**
 - interfaccia di query
 - ricerca nel database informativo
 - ordinamento delle risposte (ranking)

Architettura del motore di ricerca



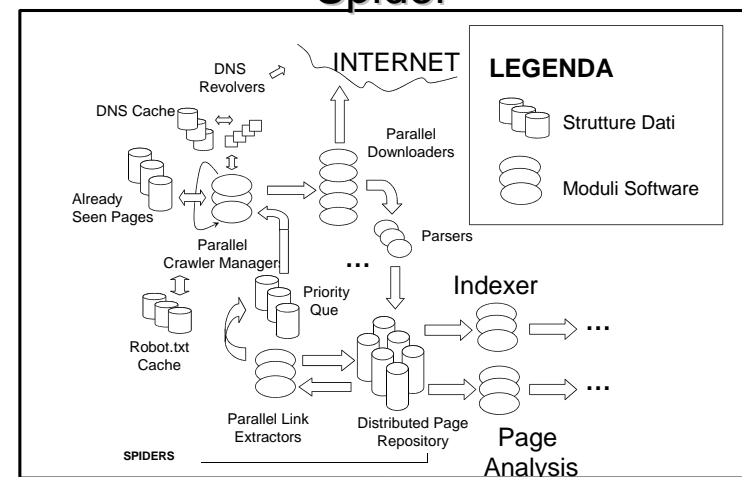
Spider / 1

- Lo spider visita una pagina web page (con una certa regolarità), la legge e segue i link ad altre pagine del sito
 - Recupera le pagine perchè siano indicizzate
 - Inizia con una pagina iniziale P0. Identifica le URL qui contenute e le accoda alle pagine da visitare
 - Finito con P0, la passa al programma di indicizzazione, recupera la pagina P1 dalla coda e ripete le operazioni
 - Può essere specializzato (e.g. recupera solo gli indirizzi email)

Spider / 2

- Problemi
 - Quale pagina esaminare in seguito? (argomenti speciali, la più recente)
 - Non sovraccaricare un sito
 - Con che profondità visitare un sito?
 - Con che frequenza?

Spider



Architettura spider

- Principali Moduli Software
 - Crawler Managers (si coordinano per decidere i prossimi task da assegnare ai Downloaders, Bilanciamento, controllo del carico...)
 - DNS Resolvers (risolvono gli indirizzi IP in modo async)
 - Parallel Downloaders (effettuano il download vero e proprio delle pagine Web usando HTTP)
 - Parsers (convertono i formati, parserizzano le informazioni, memorizzano le pagine nel repository)
 - Link Extractors (estraggono i link dalle pagine presenti nel repository e li memorizzano, senza sosta, nella struttura dati che tiene traccia delle prossime pagine da visitare)

Alcuni esempi

- Google (www.google.com)
 - 3,083,324,652 pagine raccolte
 - Centinaia di milioni di Web server distinti
 - Stimando 4Kb per pagina compressa → >10Tera
 - ~1.000 server linux (10% totale)
 - <http://www.google.com/bot.html>
- Fast (www.alltheweb.com)
 - 2,112,188,990 pagine raccolte
- Liste di spider noti
 - <http://joseluis.pellicer.org/ua/>
 - <http://www.robotstxt.org/wc/active/html/index.html>

Indexer

- L'indice contiene una *copia* di ogni pagina che lo spider trova
- Organizzato come un elenco di parole/termini significative e dei relativi riferimenti alle pagine
 - Esclude le parole da non considerare nella costruzione dell'indice (articoli, avverbi, ...)

L'informazione testuale è mal strutturata

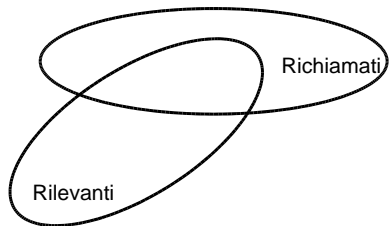
- *Richiamo (recall)*
 - Quanto materiale rilevante è effettivamente ritrovato
- *Precisione*
 - Quanto del materiale ritrovato è effettivamente rilevante

Precisione vs Richiamo

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$

Tutti i documenti



Moreno Marzolla

Tecnologie Web

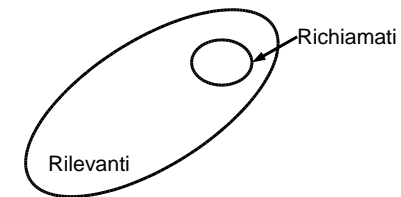
21

Alta Precisione, basso Richiamo

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$

Tutti i documenti



Moreno Marzolla

Tecnologie Web

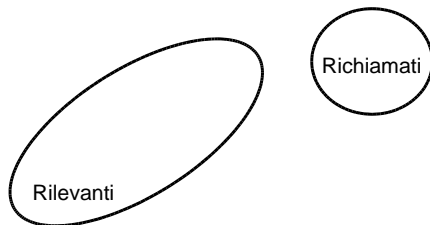
22

Bassa Precisione basso Richiamo

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$

Tutti i documenti



Moreno Marzolla

Tecnologie Web

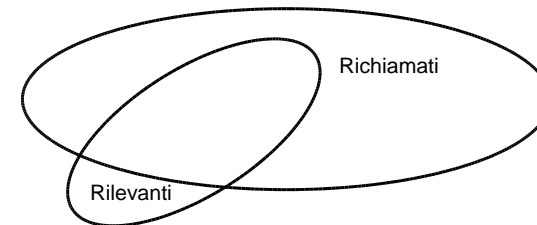
23

Alto Richiamo bassa Precisione

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$

Tutti i documenti



Moreno Marzolla

Tecnologie Web

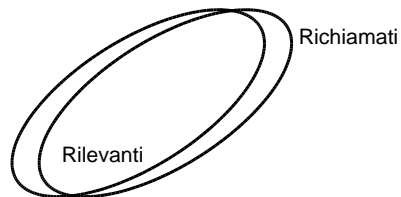
24

Alta Precisione alto Richiamo

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$

Tutti i documenti



Motori di ricerca

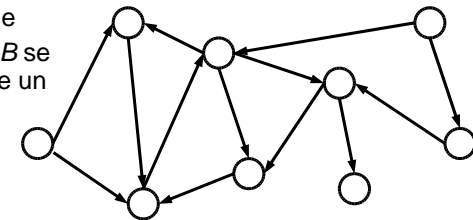
- Programma che attraversa l'indice per trovare concordanze (*matches*) con quanto richiesto dall'utente e ordina (*rank*) quanto ritrovato secondo un proprio criterio di rilevanza

Ricerche per giorno

Service	Searches Per Day	As Of/Notes
Google	150 million	5/02 (as reported to me by Google, for queries at both Google sites and its partners)
Inktomi	80 million	8/01 (as per various public statements)
AltaVista	50 million	3/00 (as cited by AltaVista in press release)
Direct Hit	20 million	4/01 (covers searches on DirectHit.com or through distribution partners such as Salon.com)
FAST	12 million	10/00 (probably for the FAST site itself and doesn't include partners)
Overture (GoTo)	6.5 million clicks	4/02 (based on Overture press release about activity for the first quarter of 2002)

Caso di studio: Google

- Idea di fondo: pagerank
 - Definire un "ordinamento" delle pagine web in modo tale da individuare quelle più rilevanti per la ricerca
- Si calcola pagerank considerando la struttura a grafo del web
 - I nodi sono le pagine
 - C'è un arco tra *A* e *B* se la pagina *A* contiene un link alla pagina *B*



Pagerank / 1

- d è una costante, $0 \leq d \leq 1$ (di solito $d=0.85$)
- T_1, \dots, T_n sono pagine che contengono un link alla pagina A
- $PR(A)$ è il pagerank di A
- $C(A)$ è il numero di link in uscita da A

$$PR(A) = (1-d) + d \times (PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))$$

Pagerank / 2

- Notare che $PR()$ è una distribuzione di probabilità
 - La somma di tutti i Pagerank di tutte le pagine WEB da 1
- $PR()$ può essere calcolato iterativamente
 - Rappresenta l'autovettore della matrice normalizzata che rappresenta la struttura del WEB

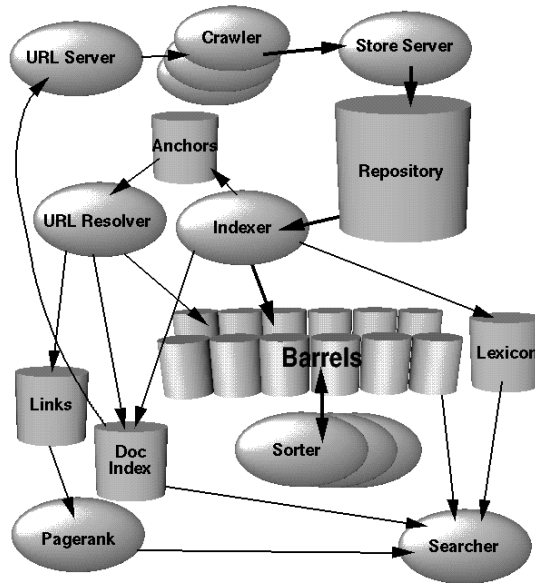
Giustificazione intuitiva

- Pagerank modella il comportamento di un utente "casuale"
 - L'utente inizia a navigare partendo da una pagina WEB scelta a caso
 - Segue link a caso sulla pagina, senza mai tornare indietro; continua a seguire link sulle pagine in cui arriva
 - Dopo un po' si stanca di seguire link, e riparte da una nuova pagina scelta a caso
 - d rappresenta la probabilità che l'utente, giunto ad una qualsiasi pagina, si stanchi e ricominci da capo

Osservazione

- Le pagine con elevato $PR()$ sono quelle "più citate"
 - Ossia quelle con più link che le puntano
 - Intuitivamente, le pagine che vengono maggiormente citate sono quelle che probabilmente contengono informazioni più autorevoli
 - In più, pagine direttamente raggiungibili da altre pagine con elevato $PR()$ (tipo yahoo.com) sono probabilmente meritevoli di attenzione

Anatomia di Google



Moreno Marzolla

Componenti di Google

- **Crawler**
 - Sistema distribuito per il recupero delle pagine WEB
- **URL Server**
 - Fornisce ai crawler i link da scaricare
- **Store Server**
 - Comprime e memorizza le pagine WEB in un repository
- **Indexer e Sorter**
 - Indicizzano il contenuto delle pagine WEB recuperate dal repository
 - Il Sorter costruisce la struttura dati che associa alle parole chiave le pagine WEB corrispondenti

Moreno Marzolla

Tecnologie Web

34

Rispondere alle query

1. Parse the query.
2. Convert words into wordIDs.
3. Seek to the start of the doclist in the short barrel for every word.
4. Scan through the doclists until there is a document that matches all the search terms.
5. Compute the rank of that document for the query.
6. If we are in the short barrels and at the end of any doclist, seek to the start of the doclist in the full barrel for every word and go to step 4.
7. If we are not at the end of any doclist go to step 4.
8. Sort the documents that have matched by rank and return the top k.

Moreno Marzolla

Tecnologie Web

35

Com'è fatto il WEB?

- **Web di "superficie" (*surface web*)**
 - Costituito dalle pagine statiche (HTML puro) pubblicamente disponibili
- **Web "profondo" (*deep web*)**
 - Costituito dai siti web dinamici e dai database accessibili attraverso una interfaccia web

Moreno Marzolla

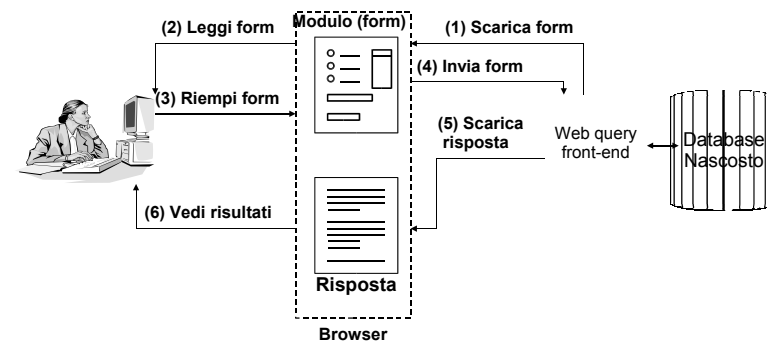
Tecnologie Web

36

Quanto è grande il WEB? Ad agosto 2000

- WEB (pagine statiche)
 - 2.5 miliardi di pagine
- WEB profondo: 550 volte più grande
 - database commerciali
 - siti che richiedono login
 - cataloghi, elenchi, orari
- I motori di ricerca generalizzati possono “vedere” solo il primo

Database “nascosti”



Quanto è grande il WEB?

- Il WEB di “superficie” è composto da circa 2.5 miliardi di documenti; tasso di crescita, 7.3 milioni di pagine al giorno
 - La dimensione media delle pagine è compresa tra 10 e 20 kbytes
- L'ammontare di informazioni del WEB di superficie varia tra 25 e 50 Terabyte

Grande?

- Il WEB “profondo” è composto da 550 miliardi di documenti, con una dimensione media di pagina di 14 kbytes, e il 95% delle informazioni è pubblicamente accessibile
- Lo spazio richiesto per tutto questo è 7500 terabytes
 - 150 volte lo spazio richiesto per memorizzare tutte le informazioni del WEB di “superficie”
 - Il 56% del WEB profondo è composto da pagine HTML, che da sole costituiscono 4200 terabytes di informazioni

Comportamento degli utenti

- Di fronte a un gran numero di risultati, la maggior parte degli utenti visita solo i primi siti elencati
- Gli algoritmi che ordinano i documenti ritrovati giocano un ruolo fondamentale
 - Vantaggio competitivo
 - Quale ordinamento (*ranking*) usare

Meccanismi di ranking

- In che ordine mostrare i documenti trovati?
 - Prima il più rilevante
 - Prima il più recente
 - Prima il più popolare
 - Prima il più affidabile
 - ...e altri ancora

Tre approcci base

- Ordinamento per rilevanza
 - Si considerano le parole chiave con più alta rilevanza statistica
 - altavista.com
- Ordinamento “a pagamento” (pay-per-click)
 - Più paghi più è alta la tua rilevanza
 - overture.com
- Ordinamento per popolarità
 - Una pagina WEB che viene maggiormente citata da altre pagine è assunta essere più importante
 - google.com

Link popularity

- I documenti web non sono solo testo, contengono anche link ad altri documenti
- Si usano i link ad un documento per classificare la sua rilevanza con i termini della ricerca
 - Qualcosa di simile al Science Citation Index, in cui il numero di articoli scientifici che citano un dato lavoro è uno strumento efficace per misurare la qualità del contenuto dell'articolo citato e la sua rilevanza per un determinato argomento
- I diversi link ad un sito sono le “citazioni”.
 - www.linkpopularity.com

Ipotesi

- Due aspetti per il ranking:
 - Siti puntati da un numero maggiore di link sono di miglior qualità
 - Se le pagine che puntano a quella in esame sono buone, allora anche quella è una pagina buona
- Le parole che descrivono i link che puntano alla pagina in esame sono degli indicatori utili sul contenuto della pagina stessa

Perché funziona?

- Il sito ufficiale della Ferrari sarà linkato da un sacco di altri siti ufficiali (o di alta qualità)
- Ma anche il miglior sito di fan-club Ferrari avrà (probabilmente) molti link che lo puntano
- Siti di minor qualità non avranno altrettanti siti di buona qualità che li puntino

Meta-motori di ricerca

- L'idea è quella di sfruttare le attività di diversi motori programmando “meta-motori” che effettuano le ricerche utilizzando un insieme di altri motori di ricerca
 - www.metacrawler.com
 - www.copernic.com

Meta-motori

- I motori di ricerca operano diversamente
 - Dimensioni
 - Diversi linguaggi per effettuare le interrogazioni
 - Diversi algoritmi di visita della rete
 - Diverse politiche per visualizzare i risultati
 - Diverse frequenze di aggiornamento
 - Diverse politiche di ordinamento dei risultati (ranking)
- *Si invia* la stessa query a diversi motori e si *collezionano* i risultati

La battaglia dei motori

- Tre motori di ricerca (luglio 2003)
 - google
 - yahoo/inktomi + overture/altavista/fast/alltheweb (acquisizione di luglio 2003)
 - teoma/askjeeves
- ...e Microsoft?
 - ha cercato di comperare google (ottobre 2003)
 - gossip: sta cercando di comperare askjeeves ...

Situazione a settembre 2003

- [...] Business che si configura come uno dei mercati a maggior crescita su internet, considerando che solo per il 2003 il fatturato relativo al posizionamento a pagamento sui motori di ricerca è stimato intorno ai 2 miliardi di dollari. Una torta enorme su cui intendono mettere le mani. Google, in febbrile sforzo di aggiornamento di un prodotto leader a livello mondiale, Microsoft, che lavora in gran segreto per sviluppare in casa una tecnologia in grado di competere con i concorrenti (ed inserirla in Longhorn, nome in codice del prossimo sistema operativo, con lancio previsto non prima del 2005). Yahoo!, che con l'acquisto di Inktomi e Overture, operatori leader nella fornitura di tecnologie e servizi per i "paid listings", posizionamenti a pagamento, ha prodotto l'ultimo terremoto nel mondo di internet. Ora, IBM fa capolino e lancia Web Fountain. Difficile dire quale sviluppo avrà questo software e le reazioni del mercato. Certamente, considerato il nome in gioco, non passerà inosservato.
- (da Repubblica, 26 settembre 2003)

Situazione a maggio 2003

- Google 55.2%
- Yahoo 21.7%
- MSN Search 9.6%
- AOL Search 3.8%
- Terra Lycos 2.6%
- Altavista 2.2%
- Askjeeves 1.5%

Il problema dell'ECommerce

- I potenziali acquirenti vogliono trovare *prodotti* (non *testo!*) che sono:
 - Descritti in modo impreciso
 - Presentati in formati differenti
 - Nascosti all'interno di database (WEB profondo)
 - Costantemente modificati
- Fallimento dei motori?
 - Non possono cercare database interni ai siti
 - Non hanno conoscenza di prodotti
 - Non si focalizzano su siti di mercanti
 - Lasciano all'utente le attività di raffinare la query
 - Non sono in grado di usare informazione imprecisa

Shopping Bots

- Sono motori di ricerca specializzati, con orientamenti specifici per categorie di prodotti
 - Libri, elettronica, viaggi, computer, giochi, film, musica, software, ...
 - mySimon.com , StreetPrices.com
 - Copernic Shopper (www.copernic.com)
- il progetto froogle di google
 - www.froogle.com

Strumenti

- Uno strumento, al momento disponibile solo per le macchine fotografiche digitali su shopping.yahoo.com, è quello per capire quale è il modello giusto per le vostre esigenze.
 - Dopo aver indicato il prezzo minimo-massimo che siete disposti a spendere, basterà rispondere a una serie di domande usando una sorta di righello che varia da "molto importante" a "per niente importante".
 - Finito il test il servizio vi presenta una lista di 10 modelli che soddisfano le vostre preferenze.

Modelli di business per gli shopping bots

- “Un tanto al click” (*pay for performance*)
 - Il sito guadagna una piccola somma (da qualche decina di centesimi a un dollaro) quando l'utente clicca sul prodotto venduto dal suo negozio online
- “Commissioni”
 - Una percentuale variabile dal 5 al 15 per cento, che si incassa solo se il rivenditore vende la merce consigliata.

La situazione negli USA...

- Lo shopping comparativo - ha spiegato al "New York Times" Daniel Ciporin, presidente di Shopping.com - non si limita più ai prodotti tecnologici. Abbigliamento (200 mila capi che diventeranno 370 mila entro metà ottobre), arredamento e altri generi diventano oggi fondamentali per attrarre il mercato di massa al quale puntiamo".
- Con l'elettronica la comparazione automatizzata era più facile: poche marche, criteri oggettivi tipo numero di watt, di megabyte e così via. Adesso trovare un comun denominatore tra golf di lana è un'operazione delicata, anche per i software più sofisticati.
- Ma l'allargamento merceologico è stato imposto anche dalla modificata demografia delle rete se è vero, come conferma anche un rilevamento di BizRate, che dal 1998 al 2002 si è passati dal 61 per cento di acquirenti online maschi alla stessa percentuale di femmine.
- E sempre più persone si rivolgono ai siti specializzati per i consigli per gli acquisti: dal 9 per cento degli internauti nel 2002 al 15 per cento censito nel 2003 da uno studio Nielsen/Netratings.

...e in Italia

- Seppur con meno categorie e funzionalità di ricerca, anche l'Italia conferma la tendenza alla crescita di servizi analoghi.
- Costameno.it sembra uno dei più efficaci: "Confronta e compra in più di 1000 siti" recita il sottotitolo.
 - Si può navigare all'interno delle categorie oppure puntare dritti sul prodotto, digitandone il nome nella griglia di ricerca. La lista di risultati è ordinata dal più economico al più caro e cliccando si arriva sul sito dei vari rivenditori.
- Kelkoo.it e BuyCentral.it (versioni italiane di un network internazionale), Sconti.it