Information Filtering and User Modelling

2 - Methodologies & Approaches

Fabio Gasparetti
gaspare@dia.uniroma3.it

Dipartimento di Informatica e Automazione

Overview

- Content Based & Collaborative Filtering
- How to Exploit User Profiles
- How to Build User Profiles
  - User Identification
  - Collecting User Data
    - Implicit/Explicit Feedbacks
  - Initial Profile Generation
- How to show Results

Overview

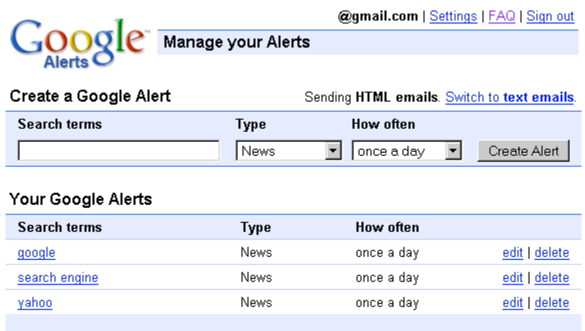
- Content Based & Collaborative Filtering
- How to Exploit User Profiles
- How to Build User Profiles
  - User Identification
  - Collecting User Data
    - Implicit/Explicit Feedbacks
  - Initial Profile Generation
- How to show Results

Content-based IF

- Representations of documents are based on document contents
- Each user operate independently
- The matching is between models of user needs and document representations

Esempio: Google Alerts

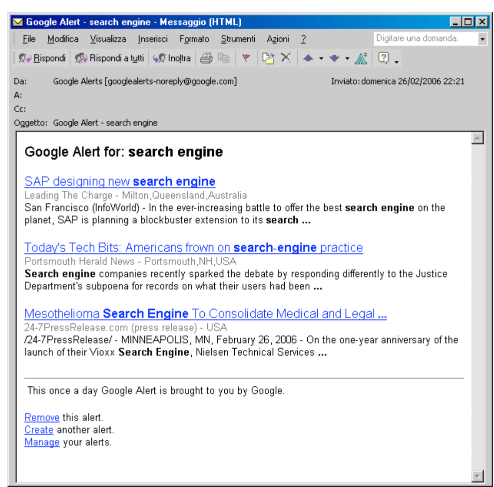
L'utente suggerisce esplicitamente i termini di cui è interessato (needs)



ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

Esempio: Google Alerts

Il motore lancia periodicamente la query su News e/o Web e i risultati vengono inviati via email



Content-based IF

- Svantaggi :
  - Stessi problemi della ricerca basata su keywords (comune a molti sistemi content-based), vedi *vocabulary problem*.
  - Occorre prevedere che i needs cambino col tempo (e.g.,Google Alerts non ha adattività)
  - Occorre definire una rappresentazione basata sul contenuto, difficile per alcuni tipi di risorse, e.g., video, audio, etc

Collaborative-based IF

- Un documento viene analizzato in base al comportamento/ratings di altri utenti del sistema senza considerare il contenuto
- "*Collaborative filtering* uses the assumption that people with similar tastes will rate things similarly; *Content-based* filtering uses the assumption that items with similar objective features will be rated similarly" (Schafer et al.'06)

Collaborative-based IF

- Tipi di *ratings*:
  - *Espliciti:*
    - *Scalar*: numerical vote, e.g., 1-5
    - *Binary*: e.g., like/dislike
  - *Impliciti:*
    - *Unary*: observed/purchased/read or not

Esempio: MovieLens

Ratings dell'utente

Suggerimenti dal sistema

Collaborative-based IF

- *User-based Nearest Neighbor Prediction*
  based on ratings from similar users

rating matrix

		Items			
		The Matrix	Speed	Sideways	Brokeback Mountain
Users	Amy	4	4	3	5
	Matt		3		4
	Paul	5	5	?	1
	Cliff	5	5	5	5

Collaborative-based IF

- *Item-based Nearest Neighbor Prediction*
  based on similarities between items

rating matrix

		Items			
		The Matrix	Speed	Sideways	Brokeback Mountain
Users	Amy	1	4	3	5
	Matt		3	5	4
	Paul	5	5	?	4
	Cliff	5	5	5	5

Collaborative-based IF

- *User-based Nearest Neighbor*:

$$pred(u,i) = \frac{\sum_{n \subset neighbors(u)} userSim(u,n) \cdot r_{ni}}{\sum_{n \subset neighbors(u)} userSim(u,n)}$$

  - *pred(u,i)* prediction rating item *i* for user *u*,
  - r_{ni} is neighbor *n*'s rating for item *i*,
  - *userSim(u,n)* is a measure of the similarity between a target user *u* and a neighbor *n* based on ratings on same items

ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

Collaborative-based IF

- Per valutare la similarità tra utenti si utilizza la *Pearson correlation* (=1 perfect agreement, -1 complete disagreement, 0 no correlation)

$$userSim(u,n) = \frac{\sum_{i \subset CR_{u,n}} (r_{ui} - \bar{r}_u)(r_{ni} - \bar{r}_n)}{\sqrt{\sum_{i \subset CR_{u,n}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \subset CR_{u,n}} (r_{ni} - \bar{r}_n)^2}}$$

  - $CR_{u,n}$ denotes the set of co-rated items between *u* and *n*
  - $\bar{r}_u$ $\bar{r}_n$ average of $\{r_u\}$ and $\{r_n\}$ ratings

Collaborative-based IF

- *Item-based Nearest Neighbor*:

$$pred(u,i) = \frac{\sum_{j \in ratedItems(u)} itemSim(i,j) \cdot r_{ui}}{\sum_{j \in ratedItems(u)} itemSim(i,j)}$$

$$itemSim(i,j) = \frac{\sum_{u \subset RB_{i,j}} (r_{ui} - \bar{r}_u)(r_{uj} - \bar{r}_u)}{\sqrt{\sum_{u \subset RB_{i,j}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{u \subset RB_{i,j}} (r_{uj} - \bar{r}_u)^2}}$$

  - $RB_{i,j}$ denotes the set of users who have rated both item *i* and item *j*

Collaborative-based IF

- Provare a predire il rating dell'item #2 per l'utente #3 con i due metodi:

Items

Users	1	2	3	4
1	5	4	3	
2	4	5	5	3
3	3	?	4	2
4	5	3	3	4
5	2		1	

$$pred(u,i) = \frac{\sum_{n \subset neighbors(u)} userSim(u,n) \cdot r_{ni}}{\sum_{n \subset neighbors(u)} userSim(u,n)}$$

$$pred(u,i) = \frac{\sum_{j \in ratedItems(u)} itemSim(i,j) \cdot r_{ui}}{\sum_{j \in ratedItems(u)} itemSim(i,j)}$$

$$userSim(u,n) = \frac{\sum_{i \subset CR_{u,n}} (r_{ui} - \bar{r}_u)(r_{ni} - \bar{r}_n)}{\sqrt{\sum_{i \subset CR_{u,n}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \subset CR_{u,n}} (r_{ni} - \bar{r}_n)^2}}$$

Collaborative-based IF

- Collaborative Filtering major tasks:
  1. Help me find new (or already seen) items I might like
  2. Advise me on a particular item:
     I have a particular item in mind; does the community know whether it is good or bad?
  3. Help me find a user (or some users) I might like (e.g., for discussion groups, matchmaking)

Collaborative-based IF

- Svantaggi:
  - Occorrono molti utenti per filtrare efficacemente le risorse
    (# ratings >> # items)
  - Ogni utente deve analizzare più items

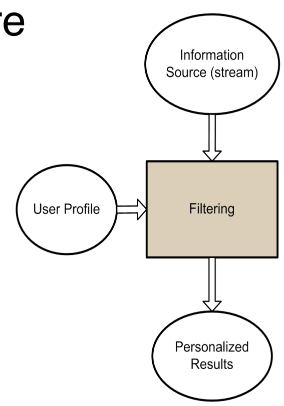
Overview

- Content Based & Collaborative Filtering
- How to Exploit User Profiles
- How to Build User Profiles
  - User Identification
  - Collecting User Data
    - Implicit/Explicit Feedbacks
  - Initial Profile Generation
- How to show Results

How to Exploit User Profiles

- Nel'IF i profili sono solitamente sfruttati per filtrare uno stream di informazioni (e.g., mailing-list, news)
- Ma alcune sorgenti non sono stream (e.g., Web)

→ Si può combinare user profiles e IR (con periodici update sulla sorgente)

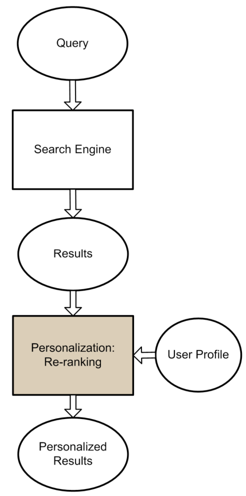


How to Exploit User Profiles

- Re-ranking
  - user profiles take part in a second step, after evaluating the corpus ranked via non-personalized scores.

    + customizable filtering

    - time to download/analyze resources



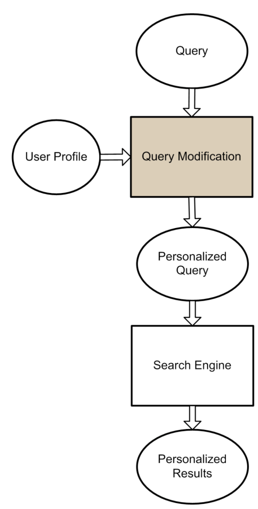
Dipartimento di
Informatica e Automazione

How to Exploit User Profiles

- Query modification
  - user profiles affect the submitted representation of the information needs, e.g., query, modifying or augmenting it.

    + same time to retrieve resources

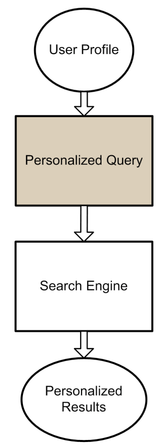
    - cannot employ complex filtering algorithms



How to Exploit User Profiles

- Recommendation
  - user profiles generate queries based on the current user behavior

    +/- same as query modificaiton



Overview

- Content Based & Collaborative Filtering
- How to Exploit User Profiles
- How to Build User Profiles
  - User Identification
  - Collecting User Data
    - Implicit/Explicit Feedbacks
  - Initial Profile Generation
- How to show Results

How to Build User Profiles

- 3 sub-tasks:
  1. User Identification
     per i sistemi multi-utente
  2. Collecting User Data
     queries, doc history, etc.
  3. Data Exploitation
     update the profile representation by the collected data

User Identification

- Identificazione utente:
  1. cookies (Web)
     tag che identificato il computer
  2. session ids (Web)
     tag che identificato la session
  3. logins
     username & password
  4. software agents
     software ad-hoc da installare lato client
  5. proxy servers (Internet/Intranet)
     sistemi che monitorano il traffico rete

User Identification

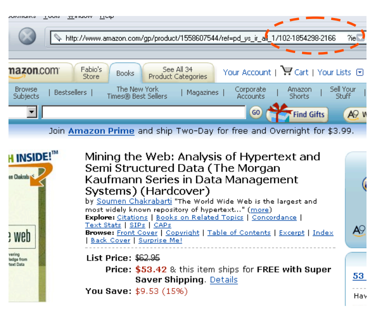
- *Cookies (Web)*
  - Alla 1a interazione col sistema viene rilasciato un tag che identifica univocamente il computer dell'utente (cookie)
  - Viene memorizzato lato client

    - tag identifica computer e non utente
    - solo per browser
    - non permanente
    - poco sicuro (se non via https)
    + trasparente all'utente

User Identification

- *session ids (Web)*
  - Viene assegnata una chiave che identifica l'interazione corrente col sistema
  - Può essere combinato a cookies
  - Utile per e-commerce

    - temporaneo
    - visibile sulla net
    - poco sicuro (se non via https)
    + trasparente all'utente

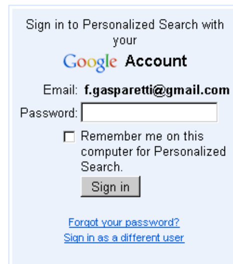


User Identification

- *logins*
  - Viene richiesto username e password prima di interagire col sistema
  - se abbinato ai cookie (vedi “remember me”) evita di fare login ogni volta

    \- password

    \+ sicuro



ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

User Identification

- *software agents*
  - l’utente interagisce con software ad-hoc
  - solitamente il software permette interazioni più complesse con l’utente rispetto al browser

\- software da installare

\- connessioni di rete non controllate

\- problemi con le network policy

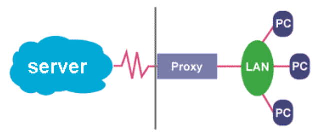
\+ maggiore HCI

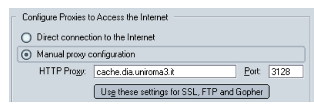
User Identification

- *proxy-server*
  - il traffico rete viene rediretto su un sistema intermedio

    - configurazione client per interagire col proxy

    - privacy





ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

Overview

- Content Based & Collaborative Filtering
- How to Exploit User Profiles
- How to Build User Profiles
  - User Identification
  - Collecting User Data
    - Implicit/Explicit Feedbacks
  - Initial Profile Generation
- How to show Results

Collecting User Data: Explicit Feedback

- Il metodo più semplice è basato su *explicit feedback*
  - l'utente segnala le informazioni utili per il user profile
  - *user data*: informazioni direttamente relative alle preferences, needs, background, knowledge dell'utente

ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

Collecting User Data: Relevance Feedback

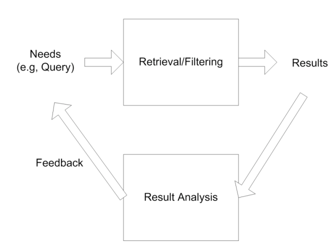
(Salton & Buckley '90)

- Introdotto nel IR: la query iniziale viene modificata con il contenuto estratto da un set di documenti valutati dall'utente

  + il processo di query re-formulation è automatizzato

  + la ricerca (i.e.,formulazione query) si decompone in una sequenza di step

  + buone performance in IR



Collecting User Data: Relevance Feedback

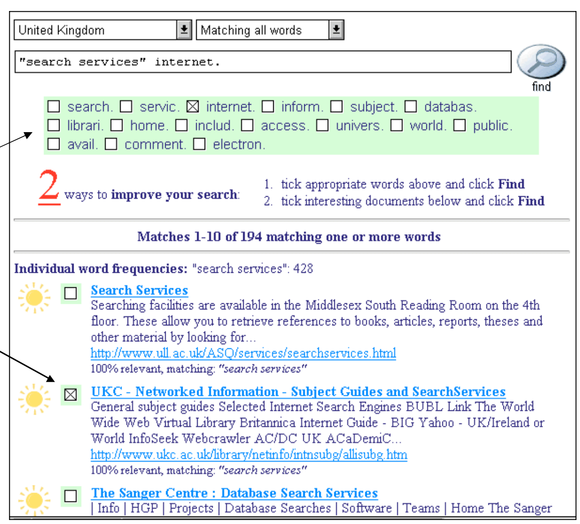
- Nell'interazione con un sistema IR la query formulation può risultare difficile (quali termini utilizzare?)
- Inoltre l'utente può non sapere cosa cercare al principio (cosa cercare?)
- L'interazione e l'analisi dei documenti permette di raffinare la query e scoprire i termini più corretti

Collecting User Data: Relevance Feedback

EuroFerret (1998): Search Engine for EU domains

Suggest keywords related to queries

Relevance feedback on specific documents



Collecting User Data: Relevance Feedback

- Nel VSM il RF "muove" il vettore query verso i documenti giudicati positivamente
- Rocchio ('66, '71) costruisce la query utilizzando 2 set di docs:

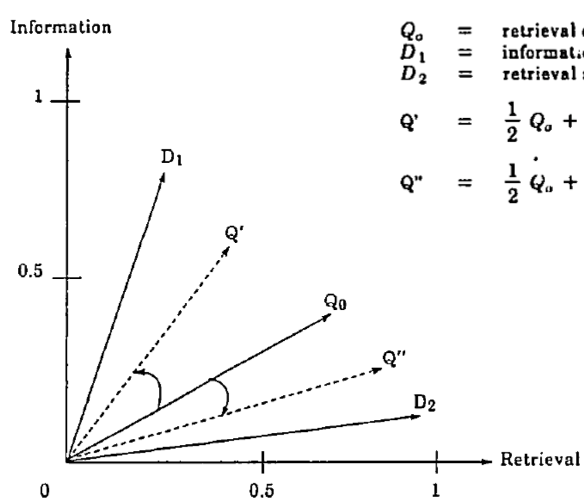
la query iniziale può essere inclusa

$$Q = \frac{1}{n} \sum_{relevant\ items} \frac{D_i}{|D_i|} - \frac{1}{N-n} \sum_{non-relevant\ items} \frac{D_i}{|D_i|} + \alpha Q_0$$

dove D_i è il vettore del documento *i*
N = #docs, n = #docs rilevanti

- Nel RF si prendono i primi 10-20docs e si pesano i vettori in base al giudizio dell'utente

Collecting User Data: Relevance Feedback



Q_o = retrieval of information (0.7, 0.3)
D_1 = information science (0.2, 0.8)
D_2 = retrieval systems (0.9, 0.1)

$$Q' = \frac{1}{2} Q_o + \frac{1}{2} D_1 = (0.45, 0.55)$$

$$Q'' = \frac{1}{2} Q_o + \frac{1}{2} D_2 = (0.80, 0.20)$$

Collecting User Data: Explicit Feedback

- Vantaggi & svantaggi:
  + solitamente alta precisione sui dati forniti
  - tempo impiegato per valutare i dati
  - a volte difficile da comprendere e sfruttare proficuamente per l'utente, soprattutto se non c'è una buona interfaccia

Collecting User Data: Implicit Feedback

- Nel *Implicit Feedback* il profilo viene costruito senza l'intervento diretto dell'utente
  - *usage data*: dati estratti durante il comportamento dell'utente e l'interazione col sistema, e.g., server access logs, query & browsing history, mouse/keyboard tracking

Collecting User Data: Implicit Feedback

- Tipi di usage-data:
  - Browser Cache/History (Web)
  - Proxy Server Logs (Web)
  - Desktop Agents
  - Low level Interactions
  - Web Server Logs (Web)
  - Query Search Logs

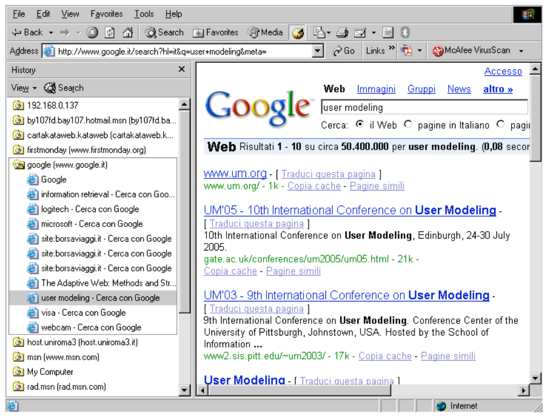
maggiormente sfruttate dai prototipi di IF fin'ora

Collecting User Data: Implicit Feedback

- *Browser Cache/History*
  - I browser Web salvano una history delle pagine visitate, e a volte una copia locale per limitare le connessioni Internet
  - Information Collected: Browsing history
  - Information Breadth: Any Web site

Collecting User Data: Implicit Feedback

- *Browser Cache/History*



Collecting User Data: Implicit Feedback

- *Browser Cache/History*
  + dati facilmente recuperabili lato client
  + alcune delle pagine Web sono molto attinenti ai needs dell'utente
  - se i dati vengono manipolati in remoto, occorre upload della history
  - molte pagine di scarso interesse (ludiche, needs estemporanei, browsing casuale)
  - il Web non è l'unica sorgente con cui l'utente interagisce
  - occorre estrarre il contenuto di interesse dalle pagine Web

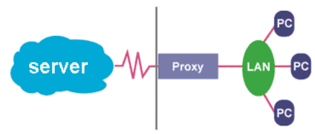
Collecting User Data: Implicit Feedback

ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

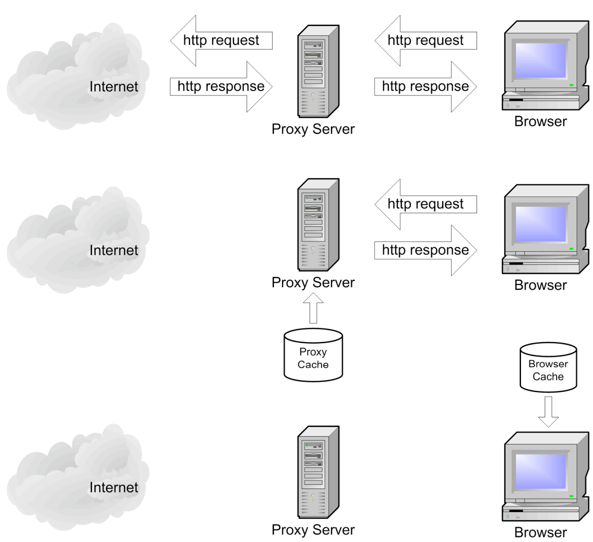
Collecting User Data: Implicit Feedback

Collecting User Data: Implicit Feedback

- Proxy Servers
  - sistema software che monitora il traffico Internet (e.g., protocollo HTTP)
  - Information Collected: Browsing history
  - Information Breadth: Any Web site



Proxy-server Interaction



1. Il proxy redireziona la richiesta del browser verso l'host su Internet

2. Il proxy restituisce la propria copia cache della risorsa richiesta in modo trasparente al browser

3. Il browser usa la propria copia cache senza connettersi al proxy

Collecting User Data: Implicit Feedback

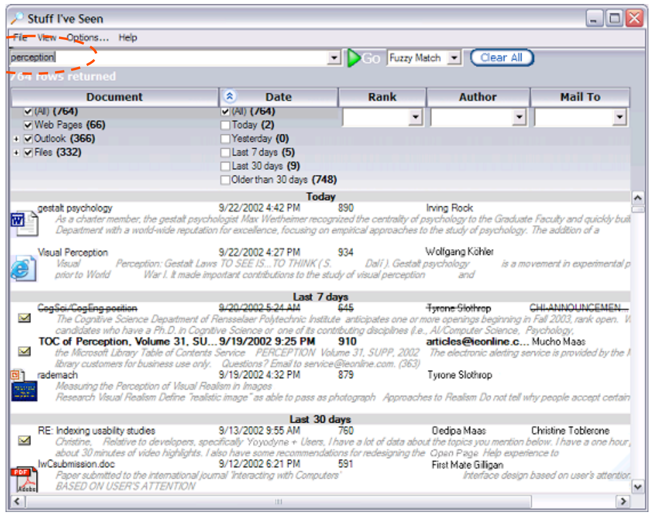
- *Proxy Servers*
  - stessi problemi login (privacy e configurazione lato client necessaria)
  - stessi problemi browser cache/history (alcune pagine Web poco attinenti e/o difficile da estrarre il contenuto di interesse)
  - il browser non inoltra tutte le richieste al proxy
  - occorre installare un proxy server

Collecting User Data: Implicit Feedback

- *Desktop Agents*
  - software che indicizzano i documenti (MS Word, Excel, PDF, etc.) sul computer dell'utente
    e.g., Copernic, Google Desktop Search, Mac Tiger, Windows Desktop Search, Yahoo! Desktop Search or X1.
  - a volte includono anche la history del browser, emails, video, audio, etc.
  - Information Collected: All user activity
  - Information Breadth: PC software data

Collecting User Data: Implicit Feedback

Microsoft Stuff I've Seen (Dumais *et al.* 2003): it records also all user actions with the interface, e.g.,: query text, use of filters, and the number of results returned by each query…



Collecting User Data: Implicit Feedback

- *Desktop Agents*
  + molti documenti sul computer sono attinenti gli user needs
  + indicizzano anche la history dell'utente (ed eventualmente altro)
  + permettono di avere un search engine locale
  - occorre installare software sul client
  - i dati vengono trasmessi fuori dal pc?

Collecting User Data: Implicit Feedback

- *Low-level Interactions*
  - actions by the user (such as pressing the mouse button, typing on the keyboard, or inserting a disk)
  - changes in windows on the screen
  - Information Collected: All user/software interaction
  - Information Breadth: Any PC software
  - Ex.: Win32 events

```
1. Click on fmEvents.fEv.ck29
2. MouseMove on fmEvents.fEv.ck29 (Œwarg=0.375 0.75 0 0 0)
3. MouseMove on fmEvents.fEv.ck29 (Œwarg=0.375 2.5 0 0 0)
4. Exit on fmEvents.fEv.ck29 (Œwarg=fmEvents.ckSync)
5. Unfocus on fmEvents.fEv.ck29
6. Focus on fmEvents.ckSync
7. MouseUp on fmEvents.ckSync (Œwarg=0.625 1 1 0 0)
8. Click on fmEvents.ckSync
9. MouseMove on fmEvents.ckSync (Œwarg=0.625 1 0 0 0)
10. Paint on fmEvents.fEv.ck29
```


Collecting User Data: Implicit Feedback

- *Low-level Interactions*

  \+ si possono catturare tutte le informazioni relative all'interazione (browser Web, Word, Outlook, etc)

  \- difficile risalire alla semantica delle azioni dell'utente

  \- occorre installare software sul client

Collecting User Data: Implicit Feedback

- *Web Server Logs*
  - file di log generati dal server HTTP (remoti) a cui l'utente accede (e.g., http://*www.dia.uniroma3.it*) che contengono informazioni sulle pagine Web visitate
  - le informazioni salvate nel log possono essere parametrizzate dall'admin del server
  - Information Collected: Browsing activity
  - Information Breadth: Logged Web site
  - HTTP log example:

```
1020816229.231 516 61.87.2.67 TCP_MISS/304 333 GET
http://www.creationent.com/.../new_sidebar.jpg - DIRECT/216.122.237.6
1020816267.836 193 61.87.2.67 TCP_MISS/302 644 GET http://home-
13.tiscali.nl/.../honeyz01.jpg - DIRECT/195.241.76.80 text/html
1020816304.598 55 226.90.141.125 TCP_REFRESH_HIT/304 203 GET
http://ar.atwola.com/...-pkCJqF1Tww$7aol - DIRECT/152.163.226.185 -
```


Collecting User Data: Implicit Feedback

- *Log popular fields:*
  - *Timestamp:*
    The time when the client connection is closed
  - *Elapsed Time:*
    The elapsed time of the request, in milliseconds
  - *Client Address*:
    IP address identifying the client
  - *HTTP Code*:
    200=OK, 301=Moved Permanently, 404=Not Found, …
  - *Size:*
    The number of bytes written to the client.
  - *Request Method:*
    The HTTP request method: GET/POST
  - *URL*
  - *Content Type:*
    text/html, image/gif, …

Collecting User Data: Implicit Feedback

- Come riconoscere il traffico generato da uno specifico utente?

```
1020816229.231 516 61.87.2.67 TCP_MISS/304 333 GET
http://www.creationent.com/.../new_sidebar.jpg - DIRECT/216.122.237.6 -
1020816267.836 193 61.87.2.67 TCP_MISS/302 644 GET http://home-
13.tiscali.nl/%7Eti017329/honeyz01.jpg - DIRECT/195.241.76.80 text/html
1020816304.598 55 226.90.141.125 TCP_REFRESH_HIT/304 203 GET
http://ar.atwola.com/.../aol - DIRECT/152.163.226.185 -
1020816320.249 130 134.202.51.180 TCP_REFRESH_MISS/200 2226 GET
http://disney.go.com/.../background.gif - DIRECT/63.70.47.83 image/gif
1020816488.105 36 191.212.159.184 TCP_CLIENT_REFRESH_MISS/304 297 GET
http://www.dailyjolt.com/../usericon_pippi.gif - DIRECT/66.70.39.30 -
1020816531.633 293 134.202.51.180 TCP_MISS/304 261 GET
http://www.traveldocs.com/images/nav_over-1x5.gif - DIRECT/63.148.100.225 -
1020816532.277 75 61.87.2.67 TCP_REFRESH_HIT/304 254 GET
http://dl.www.juno.com/../arrow.gif - DIRECT/64.136.25.24 - 1020816537.022
239539 134.202.51.180 TCP_MISS/504 1130 GET http://www.ht.ee/link.htm - NONE/- -
1020816610.488 10 134.202.51.180 TCP_MEM_HIT/200 6166 GET
http://www.realmadrid.com/../logo_madrid.gif - NONE/- image/gif
1020816668.215 6 61.87.2.67 TCP_IMS_HIT/304 267 GET http://www.hnd.com/1.gif -
NONE/- image/gif
```


Collecting User Data: Implicit Feedback

- User Tracking from HTTP server logs:
  - *IP/hostname*
    +semplice, -(NAT) più utenti stesso IP
  - *Host-munging*
    chop suffix off IP addresses, e.g., 151.47.66.170→ 151.47.*.*
    -meno accuratezza
  - \+ *Agent field*
    e.g., Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.1; .NET CLR 1.1.4322)
  - \+ *Session timeouts*
    9.5mins between requests on average, after 25mins timeouts
  - \+ *Tracking cookies when available*
  - \+ *Truncate long sessions*

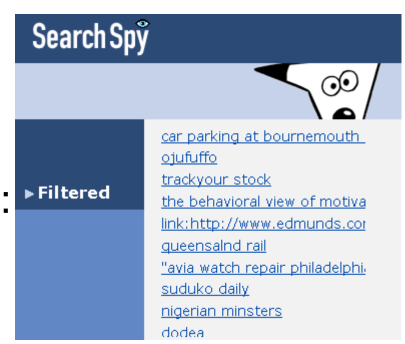
Collecting User Data: Implicit Feedback

- *Web Server Logs*
  - \+ utile per analizzare il comportamento di più utenti (e.g., quale pagine vengono più visitate dopo l'hp?)
  - \- il Web non è l'unica sorgente con cui l'utente interagisce
  - \- si può monitorare l'utente all'interno di 1 solo sito
  - \- occorre avere accesso ai file di log remoti
  - \- la cache del browser maschera alcuni accessi
  - \- difficile con più utenti per indirizzo IP
  - \- occorre escludere i robots

ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

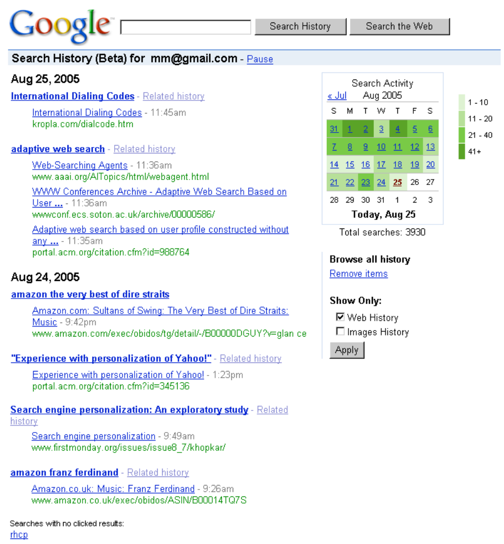
Collecting User Data: Implicit Feedback

- *Query Logs*
  - file di log delle query sottomesse in un motore di ricerca di un sito Web (e.g, portale verticale, AltaVista)
  - Information Collected: Search Queries
  - Information Breadth: Search engine site



Collecting User Data: Implicit Feedback

Google Personalized Search (2005):
records queries and visited urls



Collecting User Data: Implicit Feedback

- *Query Logs*
  - \+ le query sono buoni dati per identificare gli user needs
  - \+ i motori di ricerca collezionano già queste informazioni → personalizzazione lato server
  - \- il Web non è l'unica sorgente con cui l'utente interagisce
  - \- l'utente può interagire con più motori di ricerca
  - \- senza cookie/login non si può identificare l'utente

Collecting User Data: Implicit Feedback

(Kelly & Teevan '03)

Minimum Scope

Behavior Category	Segment	Object	Class
Examine	View Listen	Select	
Retain	Print	Bookmark Save Purchase Delete	Subscribe
Reference	Copy / paste Quote	Forward Reply Link Cite	
Annotate	Mark up	Rate Publish	Organize

ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

Collecting User Data: Implicit Feedback

- Software quali:
  - *Copernic, Google Desktop Search, Mac Tiger, Windows Desktop Search, Yahoo! Desktop Search or X1…*

  permetto già di monitorare l'utente, salvando copie di e-mails, e documenti visualizzati, nonché di analizzare l'attività corrente
- Quando si crea un account su un motore di ricerca, tutte le query e le pagine selezionate vengono salvate

→ Privacy?

Collecting User Data: Implicit Feedback

- AOL Search Query Data (August 2006)
  - 20,000,000 queries from 650,000 users
  - Uncensored queries for three months of AOL search service
  - Essentially public domain
  - Contains dangerous private information
- grep -i -e "[0-9]\{4\}-[0-9]\{4\}-[0-9]\{4\}-[0-9]\{4\}" *.txt
  - 9006-0512-xxxx-xxx
  - 1550-0905-xxxx-xxxx
- grep -i -e "\b[0-9]\{3\}-[0-9]\{2\}-[0-9]\{4\}\b" *.txt
  - kristy nicole vega hammond la. social secruity number 437-67-xxxx birth date 03 08 xx drivers license number l 00765xxxx address 41178 rene dr. hammond la.
  - pamela button 079-60-xxxx
  - thomas j finney socsec 370-40-xxxx

Collecting User Data: Implicit Feedback

- Vantaggi & Svantaggi:

  + trasparente all'utente che non è coinvolto nel processo di formulazione dei propri needs (e.g., come costruire una query?)

  + meno tempo impiegato per la ricerca

  - usage data spesso con rumore o informazione poco attinente → tecniche più sofisticate

  - usage data fa riferimento solo a positive examples

  - monitoring utente non sempre realizzabile a meno di installare software

  - privacy

Implicit vs Explicit Feedback

- Performance Evaluation
  - Quiroga and Mostafa 2000:
    precision explicit feedback 63%, implicit 58% → combined 68%
  - White et al. 2001:
    implicit and explicit feedback are interchangeable
  - Teevanet al. 2005:
    profiles built from text collected implicitly from the user's desktop index could perform better than profiles built from explicit feedback

ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

Overview

- Content Based & Collaborative Filtering
- How to Exploit User Profiles
- How to Build User Profiles
  - User Identification
  - Collecting User Data
    - Implicit/Explicit Feedbacks
  - Initial Profile Generation
- How to show Results

Initial Profile Generation

- Explicit & Implicit Feedback take time to collect user/usage data to build profiles
- Besides empty profiles that are updated as the user interacts with the systems, there're other approaches:
  - Manual
  - Stereotypes
  - Training Set

Initial Profile Generation

- *Manual*:
  - A system asks users to register their interests in the form of keywords, topics etc.
  - A form of explicit feedback, so it shares same issues:

    - burden on users if they want an accurate profile

Example: 1st Google Personalized Search

- The results are personalized according to a set of categories explicitly selected by the user
- Results which falls under the pre-selected categories (and subcategories) will be marked by an icon

Initial Profile Generation

- *Stereotypes*: “a mean of representing partial descriptions of frequently occurring situations” (Rich ’79);
- It’s aimed at generating initial predictions about the user as a classification problem;
- Stereotype represents the features of classes of users.
  E.g., if a user does not use advanced commands, he is probably a “novice" stereotype.
- Useful for group of users no-highly homogeneous

Initial Profile Generation

- The system must recognize the events whose occurrence signals the appropriateness of particular stereotypes, e.g., many *UNDO*s → novice;
- *Trigger:* an object associated with a particular situation
- The *predictions* about user characteristics are incorporated into the UM
- Because of the uncertain nature of behaviorally inferred knowledge, the UM usually include a rating representing how *confident* it is the information

Initial Profile Generation

- Typical Input Data for classification:
  - record data (name, address, etc.)
  - geographic data (area code, city, etc),
  - user characteristics (age, sex, ethnic origins, etc.),
  - psychographic data (e.g., lifestyle),
  - other (income, crime, married, children, education, etc.)
  - domain specific data

Initial Profile Generation

Grundy: a system that plays librarian

- If a librarian knows you, she will be able to provide some suggestions right away.
- If the librarian doesn't know you, she will first size you up quickly:
  - How old is he?
  - How well educated does he appear to be?
  - Is he a man or a woman?
  - What the last book that he read and liked was?
  - Who his favorite author is…
- Based on the reaction to the first few suggestions, the librarian will modify her view of you if necessary and continue making suggestions until you're satisfied

Initial Profile Generation

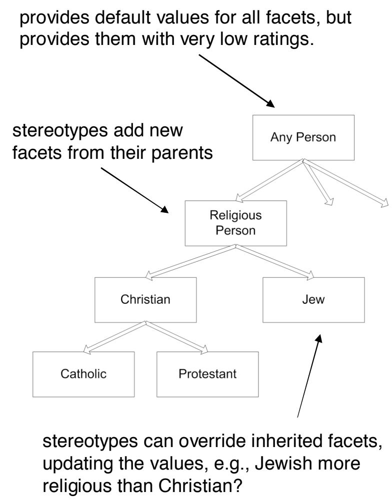
- *Grundy*'s stereotype:
  - *facet*: caratteristica
  - *value*: simple linear scale [-5..+5]
  - *rating*: degrees of certainty [0-1000]

FACET		*VALUE*	*RATING*
Activated-by		*Athletic-w-trig*	
Genl		*ANY-PERSON*	
Motivations			
	Excite	3	600
Interests			
	Sports	4	800
Thrill		5	700
Tolerate-violence		4	600
Romance		-5	500
Education		-2	500
Tolerate-suffering		4	600
Strengths			
	Physical-streng	4	900
	Perseverance	3	600

SPORTS-PERSON

Initial Profile Generation

- A directed acyclic graph (DAG) allows information not to have to be represented identically in many different stereotypes



Initial Profile Generation

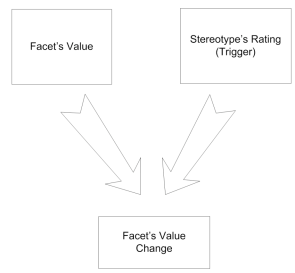
- *Grundy*'s trigger:
  - *name* of the stereotype to be activated
  - *rating* (a number between 0 and 1000) to be assigned to the stereotype.

SCI-ED-TRIG (This trigger is associated with the SCIENTIST stereotype and will be activated whenever the SCIENTIST stereotype is activated.)

FACET	*VALUE*
Stereotype	EDUCATED-PERSON
Rating	900
Reasons	SCIENTIST

Initial Profile Generation

- The amount a stereotype influences the value of a facet is a function:
  - the strength with which the stereotype predicts the value of the facet
  - the rating of the stereotype



Initial Profile Generation

- How stereotypes affect UM?
  - Going through each of the facets for which information is predicted by the stereotype, seeing what information is already there and why, and deciding what the value should be, based on both the new and the old information.
  - If the prototype was already activated, then it may be necessary to propagate its change in rating to all the predictions it makes (maybe with a rating increases). But it is only necessary to do that if the rating change is significant.

Initial Profile Generation

- *Adaptivity*: It's hard to collect real data on which to base the initial stereotypes; It is important for a UM system to be able to modify its data base of stereotypes
  - If the user behaves in a way predicted by the stereotype → prediction and triggers appropriate
  - If the user exhibits a behavior that conflicts with a stereotype prediction → prediction and/or triggers inappropriate
- What changes?
  - Stereotype' values + Triggers' ratings

Initial Profile Generation

- Esercizio: contemplare un esempio (e.g., servizio Web) dove gli stereotypes possono dare un beneficio nella creazione di un modello utente, e schematizzare alcune classi (e relative caratteristiche) della gerarchia

Overview

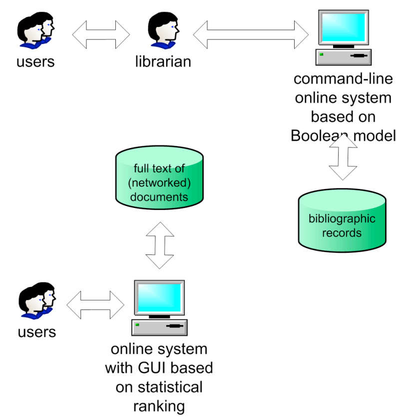
- Content Based & Collaborative Filtering
- How to Exploit User Profiles
- How to Build User Profiles
  - User Identification
  - Collecting User Data
    - Implicit/Explicit Feedbacks
  - Initial Profile Generation
- How to show Results

How to Show Results

- Once information related to the user needs has been collected, it must be organized and shown to the user
- *Goal*: provide visual depictions of very large information spaces through UIs
- *Design Principles*:
  - Provide users with feedback about relationship query-retrieved docs
  - Reduce working memory load during search (e.g., remembering search strategies, suggesting keywords)
  - Alternative interfaces novice/expert users

How to Show Results: Earlier UI

- online systems based on bibliographic records, Boolean models; users interact with a librarian; textual documents; command-line interaction



How to Show Results

- The UI should also take into consideration
  (see *berry-picking* model, Bates '89)
  - shifts of user needs during the interaction with the information sources different priorities of same goal during the search
    - new directions/goals during the search
  - a goal could be accomplished only after a sequence of distinct interactions

How to Show Results

- Berry-picking interactions with sources:
  - *browsing*: casual/undirected exploration
  - *querying* with keywords or categories selected during scanning
  - *navigating*: following a chain of links toward some goal
  - *scanning* of information structures, e.g., titles, thesaurus terms, hyperlinks, category labels; user selects/displays item for some purpose (read in detail, as a new query, surf new pages)

Browser Interface

Traditional Browser Interface

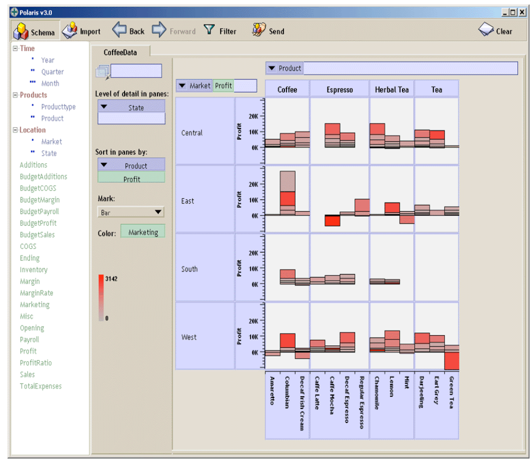
Back/History Button

Visit a URLs

Click a link

How to Show Results

- Other visualization techniques for databases aim at manipulating data to meet complex queries, e.g., visualization of data warehouses in order to: finding trends, making comparisons, aggregating inforomation, identifying crtical subsets.
- IF manipulate information objects but not data



How to Show Results

- Major approches to provide an overview on a large collection
  - *category hierarchies*
  - *clustering techniques*
  - *co-citation analysis*

How to Show Results

- *Category hierarchies*: organize Web pages into pre-defined categories (offline), e.g., MEDLINE, ACM, Yahoo!, ODP
  - Users can navigate the hierarchy and/or refine the query selecting specific subcategories

  + intuitive GUI

  + many categorization algorithms available

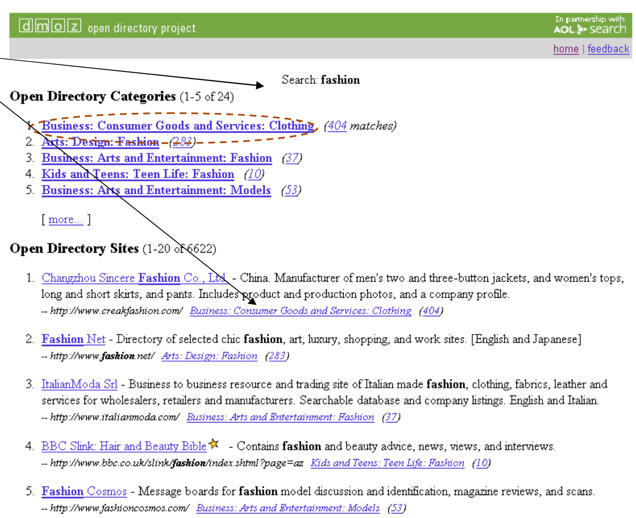
  - complex hierarchy not easy to navigate and display to users

  - automatic categorization is sometimes inaccurate while human categorization takes time

How to Show Results: ODP

User submits a query and results are shown with the related category

Results are narrowed selecting specific categories…



How to Show Results: ODP

dmoz open directory project — In partnership with AOL search

about dmoz | suggest URL | help | link | editor login

Search advanced

Arts — Movies, Television, Music...
Business — Jobs, Real Estate, Investing...
Computers — Internet, Software, Hardware...
Games — Video Games, RPGs, Gambling...
Health — Fitness, Medicine, Alternative...
Home — Family, Consumers, Cooking...
Kids and Teens — Arts, School Time, Teen Life...
News — Media, Newspapers, Weather...
Recreation — Travel, Food, Outdoors, Humor...
Reference — Maps, Education, Libraries...
Regional — US, Canada, UK, Europe...
Science — Biology, Psychology, Physics...
Shopping — Autos, Clothing, Gifts...
Society — People, Religion, Issues...
Sports — Baseball, Soccer, Basketball...
World — Deutsch, Español, Français, Italiano, Japanese, Nederlands, Polska, Dansk, Svenska...

Become an Editor — Help build the largest human-edited directory of the web

over 4 million sites - 74,222 editors - over 590,000 categories

ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

How to Show Results : ODP

dmoz open directory project — In partnership with AOL search

home | feedback

Search: **fashion**
Restricting Search to Category: **Business/Consumer_Goods_and_Services/Clothing**

Open Directory Categories (1-2 of 2)

1. **Business: Consumer Goods and Services: Clothing** (*492 matches*)
2. **Business: Consumer Goods and Services: Clothing: Men's: Wholesale and Distribution: Fashion**

Open Directory Sites (1-20 of 492)

1. Changzhou Sincere **Fashion** Co., Ltd. - China. Manufacturer of men's two and three-button jackets, and women's tops, long and short skirts, and pants. Includes product and production photos, and a company profile.
   -- *http://www.creakfashion.com/ Business: Consumer Goods and Services: Clothing (492)*
2. A.R. **Fashion** Outfit Ltd. - India. Manufactures and exports men's, women's and children's wear, and towels. Includes product photos, a list of equipment and manpower.
   -- *http://www.arfashion.com/ Business: Consumer Goods and Services: Clothing (492)*
3. Alpine Fashions International - Bangladesh. Manufactures men's, women's and children's apparel. Includes product details, photos, a company profile and a query form.
   -- *http://www.alpine-fashions.com/ Business: Consumer Goods and Services: Clothing (492)*
4. DI **Fashion** World - Myanmar. Manufacturer of mens' jackets, ladies' tops, skirts and pants, and childrens' wear. Includes product photos.
   -- *http://difashionworld.com/ Business: Consumer Goods and Services: Clothing (492)*
5. Natalia Naftalieva **Fashion** House - Russia. Manufacturer of haute couture collections, corporate clothes and children's wear. Includes photographs and sketches.
   -- *http://naftalieva.com/ Business: Consumer Goods and Services: Clothing (492)*

How to Show Results

- *Clustering Techniques*: techniques to extract general topics from the collection (online/offline)
  - Users can submit a query and see the results clustered by topics, or they can navigate through clusters

  + intuitive GUI

  + some clustering algorithms adaptable to this domain (online clustering is a time-consuming task)

  - sometimes clustering doesn't recognize important topics

ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

How to Show Results: Scatter & Gather

- Scatter/Gather (Cutting et al. '92)
  - clusters documents into topically-coherent groups
  - presents textual summaries that characterize each cluster extracted by the related docs
  - users may select an interesting cluster reclustering its content on-the-fly

  Evaluation shows that without a search facility the prototypes is less effective than traditional query-based search

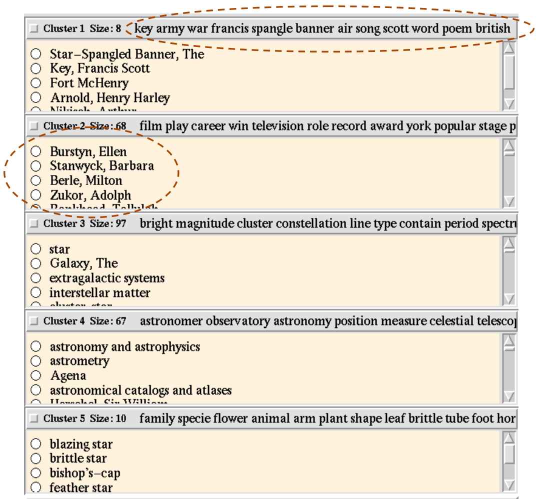
ROMA TRE UNIVERSITÀ DEGLI STUDI

Dipartimento di Informatica e Automazione

How to Show Results: Scatter & Gather

Cluster summary

Documents

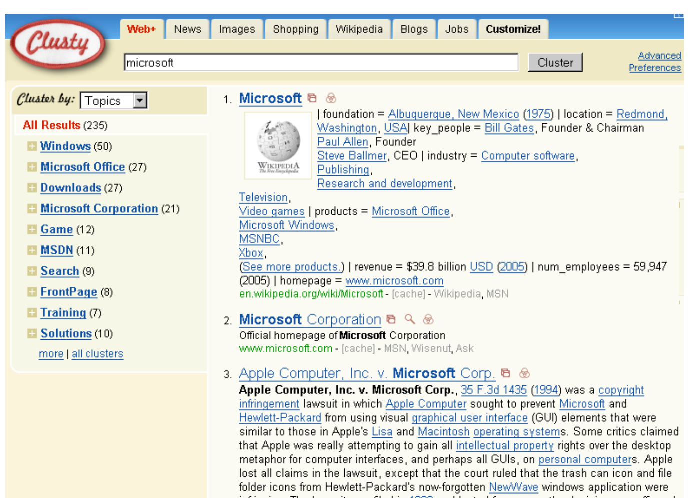


97

ROMA TRE UNIVERSITÀ DEGLI STUDI

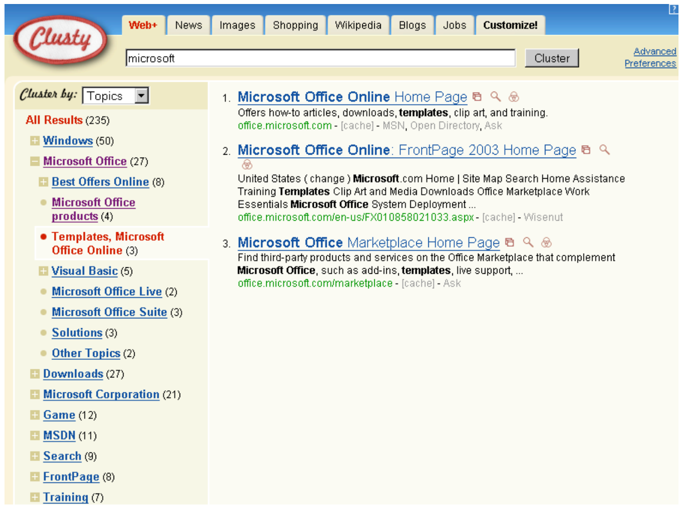
Dipartimento di Informatica e Automazione

How to Show Results: Clusty.com



98

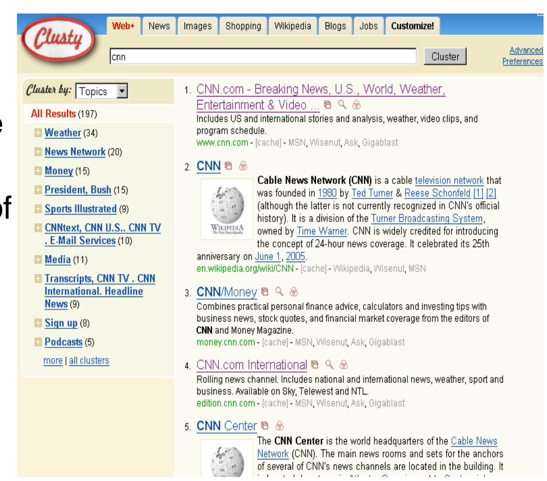
How to Show Results: Clusty.com



ROMA TRE UNIVERSITÀ DEGLI STUDI
Dipartimento di Informatica e Automazione

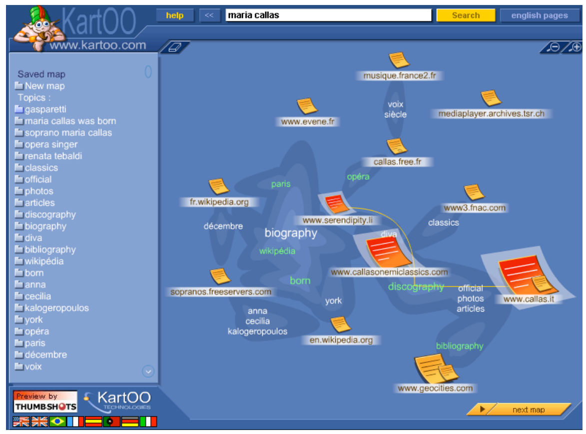
How to Show Results: Clusty.com

- Besides show less important topics, result clustering have one more problem…
- very different levels of description within a single display



How to Show Results: KartOO

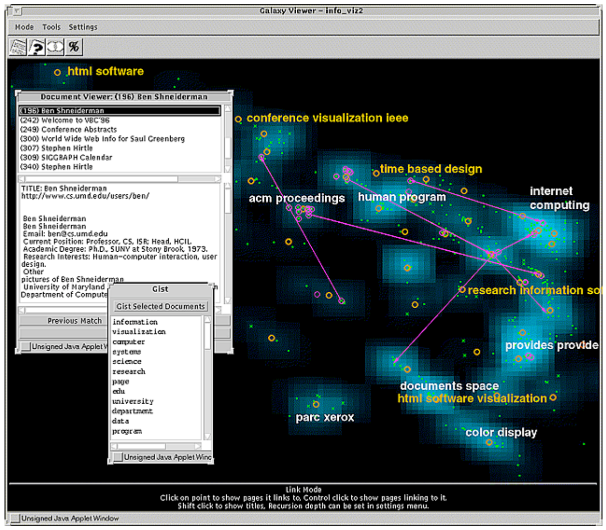
- Meta search engine which presents its results on a map.
- Sites are represented by more or less important size pages, depending on their relevance.



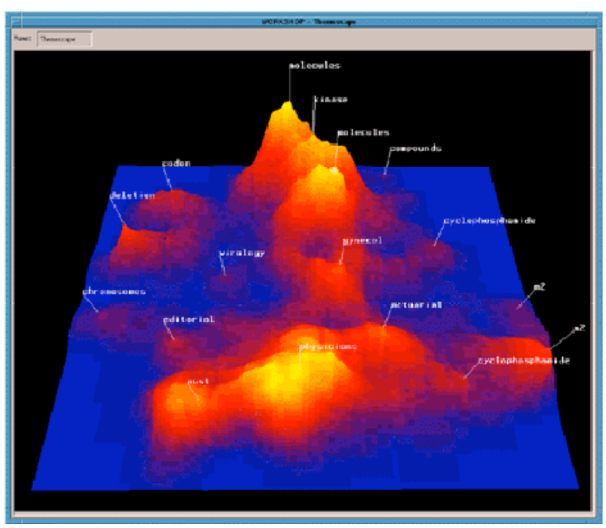
How to Show Results:

- Other clustering techniques show results in 2D/3D maps to emphasize:
  - how many documents per cluster
  - affinity between documents
  - cluster-document affinity

How to Show Results: WebTheme



How to Show Results: ThemeScapes



How to Show Results

- *Co-citation analysis*: show centrally-located or related documents based on co-citations patterns (see also *bibliometric*).
  e.g., times 2 articles are cited by a third, times pair of articles cite the same one
  - On the Web citations become links

  \+ based on link structure and not on the content

  \- it doesn't work if there're no link information

How to Show Results: CiteSeer (Bollacker *et al.* '98)

- Scientific literature digital library (750.000 docs) and search engine
- Allows browsing the database using citation links

How to Show Results: CiteSeer (Bollacker *et al.* '98)

Locates related documents using citation and word based measures, e.g.,

- *active bibliography:* lists documents that have citations in common with the current document
- *co-citations*: documents that are being cited by the documents citing the current document

Cited by: More
Generalizing the Notion of Support - Michael Steinbach Dept (Correct)
Hyperclique Pattern Discovery - Hui Xiong Hui (Correct)
Mining Strong Affinity Association Patterns in Data Sets.. - Distribution Hui Xiong (Correct)

Similar documents (at the sentence level):
56.4%: Fast Algorithms for Mining Association Rules - Agrawal, Srikant (1994) (Correct)

Active bibliography (related documents): More All
0.6: An Efficient Algorithm for Mining Association Rules in Large.. - Ashok Savasere (1995) (Correct)
0.4: Mining Association Rules between Sets of Items in Large.. - Agrawal, Imielinski, Swami (1993) (Correct)
0.3: Active Data Mining - Agrawal, Psaila (1995) (Correct)

Similar documents based on text: More All
0.4: Literaturverzeichnis - Dm-Seminar Juli (Correct)
0.3: Parallel Mining of Association Rules - Agrawal, Shafer (1996) (Correct)
0.3: Searching with Numbers - Agrawal, Srikant (2002) (Correct)

Related documents from co-citation: More All
59: Mining association rules between sets of items in large databases - Agrawal, Imielinski et al. - 1993
35: Discovery of multiple-level association rules from large databases - Han, Fu - 1995
33: Mining Generalized Association Rules - Srikant, Agrawal - 1995