Collaborative Computing

- Today's lecture:
  - Information Filtering (and Information Retrieval)
  - Collaborative Filtering and Content based filtering
  - Filtering Techniques
  - User Modelling and levels of adaptation
  - Adaptive Systems in general
  - Basic Evaluation of Information Filtering Systems
  - Research Trends

Information Filtering (IF)

- The advent of the Web has exposed users to a huge amount of information.
- Information Filtering is a technique that tries to reduce the information overload and filter information that is relevant to users.
- Information is filtered with the help of a profile of the user, also called User Model (UM) or User Profile (UP)

Differences between Information Filtering and Information Retrieval

- Old Definitions:
  - Information Retrieval is concerned with retrieving information to a user on the basis of user questions/queries
  - Information Filtering is concerned with building a long term profile of the user information needs and sort out incoming information to the user

  (Belkin and Croft 1992)
- Techniques in IF are similar to the techniques utilized in IR.
- Today: Personalized Information Retrieval is also considered in the domain of IF (Waern 04)

Where is Information Filtering utilized?

- Recommender Systems, i.e. systems that make a personalized selection of information items or products.
- News filtering, i.e. systems filter incoming streams of news information
- Email filtering (e.g. filters out SPAM)

Information Filtering main categorization

- Balabanovic and Shoham (97) categorize IF into two major topics: *Content Based Filtering* and *Collaborative Filtering*
- *Content Based Filtering:* representations of the information items are compared to the representation of the user (user model) in order to find the information items that are relevant.
- *Collaborative Filtering* aims at predicting user preferences, based on the preferences of a group of users (with similar interests).
- Information can be gathered from both restricted domains or open domains.

Other Categorizations of IF systems

- Initiative of Operation: Active vs Passive
- Location of Filtering Operation: at the information source, at a filtering server (3-tier architecture), or at the user's site (locally)
- Methods for acquiring information about the user: explicit, implicit or a combination of both (more about this later)

More on Collaborative Filtering

- The focus is on the opinions of user groups rather than on the content of a document or item.
- Can be defined as a *Social Navigation Technology* (Munro, Hook, Benyon 99).
- *Social Navigation*: human beings are social animals and tend to follow other people's advice or judgment when looking for information or buying items.

Techniques utilized in Information Filtering

- Filtering techniques are usually divided in:
- Knowledge-based techniques: e.g. rules and semantic nets
- Statistical techniques: data based (e.g. user profiles are weighted vectors of terms that are compared to weighted vector of terms of information items or other user profiles)

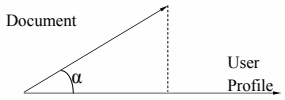
Example of Knowledge-based techniques

- Filtering Rules: "if the mail comes from an unknown sender, give it low rank". Usually expressed in some formal language e.g. Logical expressions.
- Semantic Nets: "a structure which is used to represent associations between concepts" (Gray 84). Nodes are concepts and arcs are the relations between the concepts. Every node and arc has a weight that reflects their co-occurence relation.

(from Arezki et al. 04)

Example of Statistical Techniques: Vector Space Model (VSM)

- This technique come originally from IR (Salton and Buckley, 1988)
- Main idea: documents and user profiles are represented as vectors in a n-dimensional space. The closer the vectors are, the more similar the user profiles and the documents are.
- We count how close the vectors are with the help of cosine similarity.



Term Vector

- d_a, d_b, d_c, d_d are documents
- t_1, t_2, t_3, t_4 are terms that occur in those documents

	t_1 (rent)	t_2 (car)	t_3 (travel)	t_4 (bank)	…
UP	1	1	0	0	
d_a	1	0	0	0	…
d_b	0	1	1	1	…
d_c	0	0	1	1	…
d_d	1	1	1	0	…

n dimensions of the vector space

Value for the n^{th} position of a document vector

- $<t_1, t_2, \ldots, t_n>$ for a particular document is a term vector, a.k.a. document / user profile vector

UP = <1, 1, 0, 0>, d_a = <1, 0, 0, 0>, d_b = <0, 1, 1, 1>

Cosine Similarity i VSM

- Query vector q and document vector d, both of length n. Cosine similarity between them is defined as:

Original definition of scalar product

$q \bullet d = |q| \cdot |d| \cdot \cos \alpha$

$$sim(q,d) = \cos\alpha = \frac{q \bullet d}{|q| \cdot |d|}$$

$$q \bullet d = \sum_{i=1}^{n} q_i \cdot d_i$$

Values of the i^{th} position in vectors q and d

$$sim(q,d) = \frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$

$$|v| = \sqrt{\sum_{i=1}^{n} v_i^2}$$

Length of a vector in *n*-dimensional space

Examples of Cosine Similarity

Terms	rent	car	travel	bank	Stockholm	*sim* (*q*, *d*)
UP	1	1	0	0	1	
d_a	1	0	0	0	1	0.816
d_b	0	1	1	1	1	0.577
d_c	0	0	1	1	0	0
d_d	1	1	1	0	1	0.866

UP = user profile
Documents = d_a , d_b , d_c , d_d

Values of the i^{th} position in vectors q and d

$$sim(q,d) = \frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$

sim (UP, d_a)= $\frac{(1+0+0+0+1)}{\sqrt{1^2+1^2+0^2+0^2+1^2}*\sqrt{1^2+0^2+0^2+0^2+1^2}} = \frac{2}{\sqrt{3}*\sqrt{2}} = 0.816$

sim (UP, d_b)= $\frac{(0+1+0+0+1)}{\sqrt{3}*\sqrt{4}} = 0.577$

- Relevance:
  1. d_d 2. d_a 3. d_b 4. d_c

Term Weights

- Up to now we considered binary term weights:
  - 1 : a term occurs in the document
  - 0 : a term does not occur in the document
- Better using term weights based on Term Frequency (TF) and Inverted Document Frequency (IDF)

TF and IDF

- tf_{ij} is the frequency of the j^{th} term in the i^{th} document (how many times a term occurs in the document)
- IDF helps to individuate the topic of the document finding the most "informative" words for the document
- idf-score of the j^{th} term measures the distribution of the term over the entire collection of documents:

$$idf_j = \log(\frac{N}{n_j})$$

  - N is the total number of documents in the collection
  - n_j is the number of documents that contain the j^{th} term

Tf*Idf-score Individual for Each Document

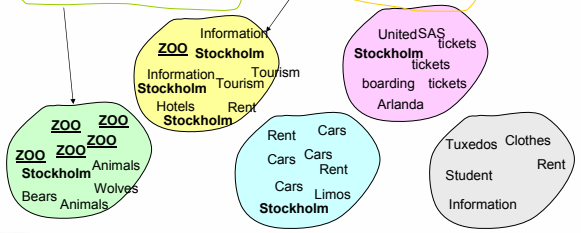
$w_{ZOO} = 5 \cdot \log_2(\frac{5}{2}) \approx 6.6$

$w_{STOCKHOLM} = 1 \cdot \log_2(\frac{5}{4}) \approx 0.32$

$w_{ZOO} = 1 \cdot \log_2(\frac{5}{2}) \approx 1.32$

$w_{STOCKHOLM} = 3 \cdot \log_2(\frac{5}{4}) \approx 0.96$

Which words best dinstinguish the document topic?



tf.idf-score in Cosine Similarity

- For best results, consider the combined tf.idf-score as the **term weight** for the j^{th} term in the i^{th} document :

$$w_{ij} = tf_{ij} \cdot idf_j$$

$$sim(q,d) = \frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$

Example of Statistical Techniques : Pearson Correlation Coefficient (Foltz and Dumais, 92)

- Given two sets (vectors) with values, it counts how close/correlated those sets are. Mainly utilized in Collaborative Filtering.
- The result ranges between -1 and 1.
- For example users are classified as more similar if their ratings are similar.
- Given two users "a" and "b" we have the formula:

$$w(a,b) = \frac{\sum_j \left(v_{a,j} - \overline{v_a}\right) * \left(v_{b,j} - \overline{v_b}\right)}{\sqrt{\sum_j \left(v_{a,j} - \overline{v_a}\right)^2 * \sum_j \left(v_{b,j} - \overline{v_b}\right)^2}}$$

- Where $v_{a,j}$ is the rating for item j by user "a", $v_{b,j}$ is the rating for item j of user "b".
- $\overline{v_a}$ is the mean value of user "a" ratings and $\overline{v_b}$ is the mean values of user "b" ratings

Example of Pearson Correlation value

	User A ratings	User B ratings
Item A	5	4
Item B	4	3
Item C	3	3

$$corr(a,b) = \frac{\sum_j \left(v_{a,j} - \overline{v_a}\right) * \left(v_{b,j} - \overline{v_b}\right)}{\sqrt{\sum_j \left(v_{a,j} - \overline{v_a}\right)^2 * \sum_j \left(v_{b,j} - \overline{v_b}\right)^2}}$$

$\overline{v_a} = 4$

$\overline{v_b} = 3.33$

$v_{a,j}$ is the rating for item j by user "a"

$$corr(a,b) = \frac{(5-4)(4-3.33)+(4-4)(3-3.33)+(3-4)(3-3.33)}{\sqrt{((5-4)^2+(4-4)^2+(3-4)^2)*((4-3.33)^2+(3-3.33)^2+(3-3.33)^2)}}$$

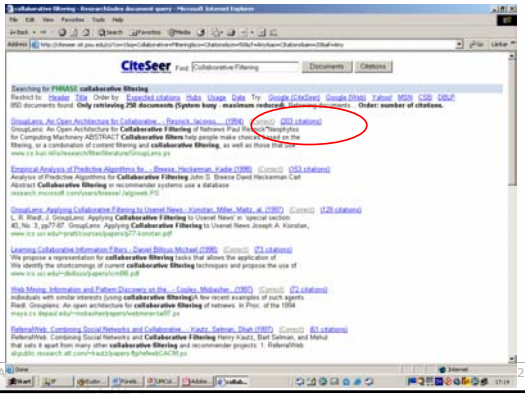
$$corr(a,b) = \frac{(1)(0.67)+(0)+(-1)(-0.33)}{\sqrt{(1+0+1)*(0.448+0.108+0.108)}} = \frac{1}{\sqrt{(2)*(0.664)}} = \frac{1}{\sqrt{1.328}} \approx 0.86$$

How does Collaborative Filtering provide suggestions?

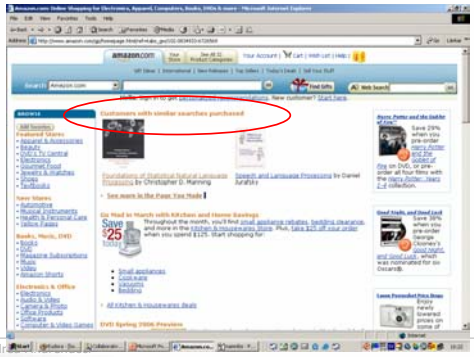
- Two basic classes of CF (Breese, Heckerman, Cardie 98)
  - Individual items are presented one-at-a-time with its rating
  - Ordered lists of recommended items, where highest ranked items are predicted to be most preferred. Remind of search engines.

Daily Collaborative Filtering Examples: Google

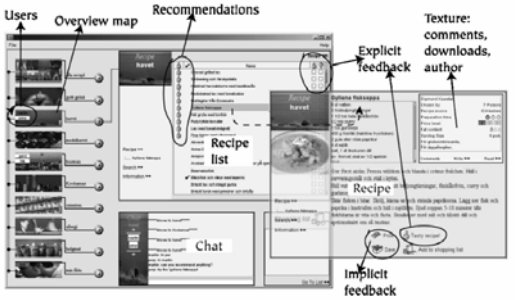
Daily Collaborative Filtering Examples : CiteSeer



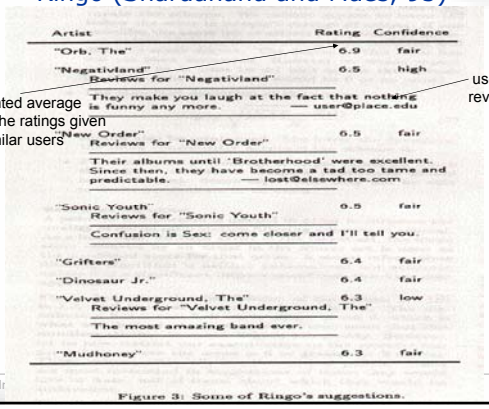
Daily Collaborative Filtering Examples : Amazon



Collaborative Filtering – Research Examples: Kalas (Svensson et al. 05)



Collaborative Filtering – Research Examples: Ringo (Shardanand and Maes, 95)



Artist	Rating	Confidence
"Orb, The"	6.9	fair
"Negativland"	6.5	high
Reviews for "Negativland"		
They make you laugh at the fact that nothing is funny any more. — user@place.edu		
"New Order"	6.5	fair
Reviews for "New Order"		
Their albums until 'Brotherhood' were excellent. Since then, they have become a tad too tame and predictable. — lost@elsewhere.com		
"Sonic Youth"	6.5	fair
Reviews for "Sonic Youth"		
Confusion is Sex: come closer and I'll tell you.		
"Grifters"	6.4	fair
"Dinosaur Jr."	6.4	fair
"Velvet Underground, The"	6.3	low
Reviews for "Velvet Underground, The"		
The most amazing band ever.		
"Mudhoney"	6.3	fair

Figure 3: Some of Ringo's suggestions.

Weighted average of all the ratings given by similar users

user reviews

Content-based filtering research example – Persival (McKeown et al., 2001)

- Personalized Search Engine for HealthCare articles
- Extract user profiles from patient records and utilizes the information in the UP to rank the documents retrieved from online medical resources
- Represent user profiles and documents as term-value vectors and utilize cosine similarity
- Utilize Natural Language Processing techniques (syntactic parsing) to parse medical documents and create summaries that are tailored to the patients background
- Helps practitioners to find evidence for treatment or diagnosis of patient diseases

Content-based Filtering - Drawbacks

- Require machine-readable/parsable items, e.g. text-based documents, since it creates a formal representation of the content of the information items
- It is more difficult to automatically create a representation of images, speech or sound
- News filtering poses real-time constraints (the system cannot process the information too long)
- It is difficult to judge the quality of the information items

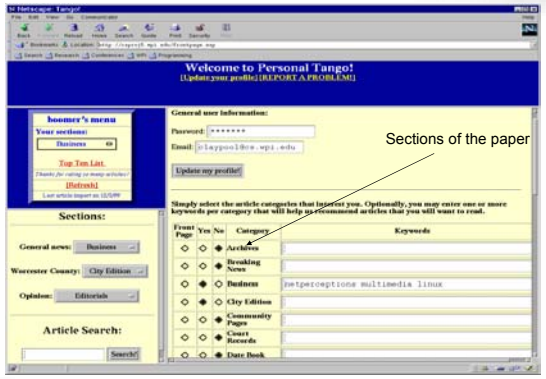
Collaborative Filtering - Drawbacks

- Requires bootstrapping: recommendations cannot be done if there is not sufficient amount of data, i.e. user ratings
- Sparsity problem: users may rate small sets of all available items or different sets of items, which make comparison of user preferences more difficult.
- Early rater problem: prediction cannot be made for an item when it first appears, since there are not user ratings for that item.
- Changing interestes problem: what happens if our interests change? Do we have to re-rate all the items?

How to overcome those problems? Combine!

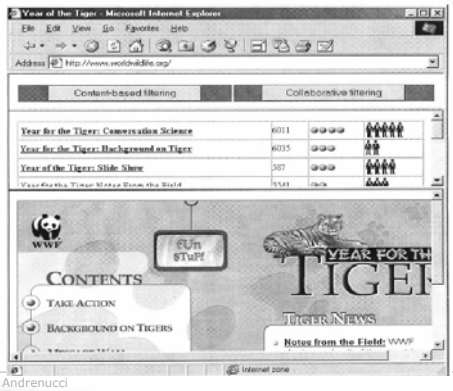
- Claypool et al. (99) combines both approaches in a system called P-Tango, that filters news articles. The prediction of articles relevance is based on the average of the content-based predictions and the collaborative predictions.
- Grouplens (Sarwar, Konstan et al. 98) also combines the approaches. The system provides a content-based evaluation of news articles and computer ratings are treated just like ratings of human beings.
- ProfBuilder (Wasfi, 99) recommends Web pages in two lists: one generated by content-based Filt. and another generated by Collaborative Filtering.

P-Tango (from Claypool et al. 99)



P-Tango (from Claypool et al. 99)

ProfBuilder, Profile Builder, (Wasfi 99)



User Modeling (UM)

- User Model is ”the knowledge about the user, either explicitly or implicitly encoded, that is used to improve the interaction between the user and a system” (Kass, Finin 98)
- UM usually i symbiosis with Adaptive Systems (which include filtering systems).
- Adaptive Systems: programs that adapt their behaviour to user characteristics and background (Hypertext: *Adaptive presentation and Adaptive Navigation Support*).

Adaptive Presentation

- Adaptive Presentation (Content Level Adaptation):
  - Text adaptation
  - Adaptation of Modality (Which medium to choose in order to present information: audio, video or text?)

”Content adaptation” Example
Text for doctors and patients: Opade system
from (De Carolis et al. 96)

Patient	**Doctor**
Comments to the drug prescription of Mr Fictif. You have been diagnosed as suffering from a mild of what we call “angina pectoris”, that is a spasm of chest resulting from overexertion when heart is diseased. In addition you have elevated cholesterol…	As you certainly remember, Mr Fictif is a 62 years old man. He is overweight. He is suffering from a mild form of angina and he has got elevated cholesterol…

Adaptive Navigation Support

- Adaptive Navigation Support (Link Level Adaptation):
  - Direct Guidance (e.g. educational systems)
  - Link Hiding
  - Link Sorting
  - Link Annotation (e.g. icons or markers)
  - Link Generation (generates new links in real time)

"Link Level Adaptation" Example
Syskill & Webert, (Pazzani et al. 96)

Methods for acquiring the UM

- There are mainly three methods for gathering information about the user:
  - Explicitly
  - Implicitly
  - A combination of both

Explicit Acquisition of UM

Users are asked to:

- provide keywords
- choose from lists of topics and grade their interests on a fixed scale

Implicit Acquisition of UM

- Focuses on monitoring the user-system interaction:
  - the time spent a Web page and how users scroll the content of the Page, do they just skim?
  - repeated visits on a Web page
  - mouse clicks and link selections
  - bookmarks
  - types of documents and information items saved or printed
  - relevance feedback (judging the relevance of retrieved data items)

A combination of both

- Users explicitly create their profiles and the profile is dynamically updates by "looking over users' shoulders" (i.e. which documents were ranked as relevant by the user).
- Users explicitly submit information about themselves and are assigned to a *stereotype*. Further information about their profile is inferred with help of *the stereotype* (Ex: Male, mid twenties, work-out -> most likely to enjoy action films)

Drawbacks of Explicit User Modeling

- It can be a tedious process and users may skip it.
- Users may omit important information or forget to update their profile, reducing the accuracy of the retrieved/filtered information.

Drawbacks of Implicit User Modeling

- User models are difficult to correct when they are not accurate or have ”aged” → user interests change
- Users are monitored and tracked, which can arise privacy and integrity as well as ethical issues.
- Users tend to dislike being tracked.
- What if the user needs information on someone else’s behalf, e.g. a friend or a relative?

Exampels that combine both explicit and implicit UM

- P-Tango (Claypool et al.) recommends news articles for an online newspaper. The UM both *explicit,* with keywords entered by the user, and *implicit*, with keywords gathered from articles the users rated as interesting.
- ConfCall (Waern et al. 04) recommends relevant Conference calls. Through a profile editor, users submit keywords about their interests. The system then monitors which incoming documents users read or discard, updating the profile.

Example: ConfCall (Waern 04)

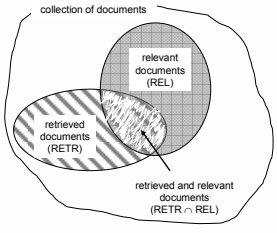
Some Basic Evaluation Techniques: Precision and Recall (Salton and McGill, 1983)

- Precision: is the fraction of relevant documents retrieved from the total number retrieved (Accuracy)
- Recall: is the fraction of relevant documents retrieved from the set of total relevant documents in the collection. (Completeness)

Precision of Retrieval

- **Precision** characterizes the fraction of relevant documents retrieved from the total number retrieved :

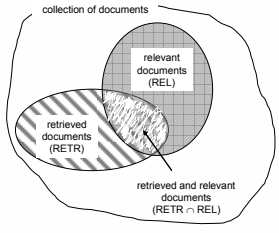
$$P = \frac{|RETR \cap REL|}{|RETR|}$$



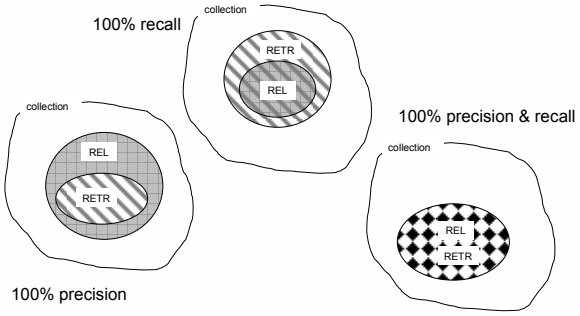
Recall of Retrieval

- **Recall** characterizes the fraction of relevant documents retrieved from the set of total relevant documents in the collection :

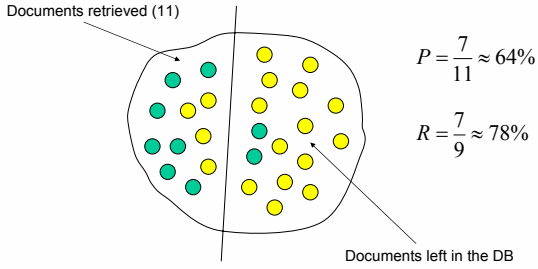
$$R = \frac{|RETR \cap REL|}{|REL|}$$



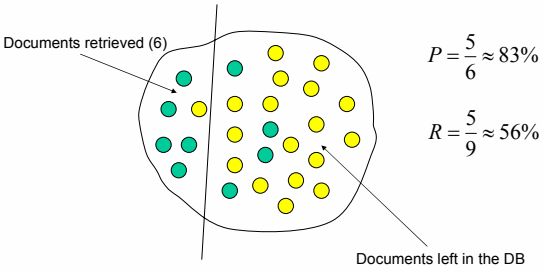
100%



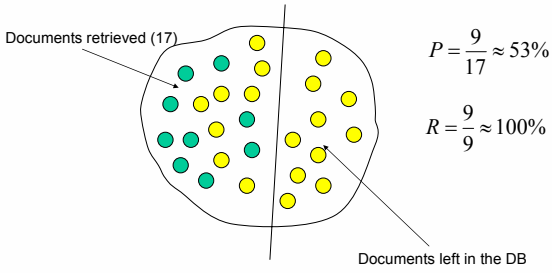
Visual Example



Moving the Cut-off Line



Moving the Cut-off Line



Some Research Issues and Trends in IF

- Make IF wearable: palm Tops and Mobile Phones
- Combining the UM approaches (Both explicit and implicit)
- Combination of filtering techniques (Content based and Collaborative)
- Improvement of visualization techniques and metaphors
- Privacy protection
- Multilingual IF
- Portability of IF systems

The End

Thank you for listening