Content-Based Filtering

• Outline
• Introduction to content based filtering
  – Applications
  – Main research problems
• Main framework
• Learning a threshold
Content Based Filtering

- Use when information needs are stable
  - System should make a delivery decision on the fly when a document “arrives”

User Profile:

Information Needs are Stable
- System should make a delivery decision on the fly when a document “arrives”
- User profiles are built by some initialization information and the historical feedback information from a user (e.g., relevant concept extracted from relevant documents)
- The delivery decision is made by analyzing the content information of a document
Content Based Filtering: History-Based

User History
Description: In the near future, a computer hacker named Neo (Keanu Reeves) discovers that all life on Earth may be nothing more than an elaborate facade created by a malevolent cyber-intelligence...........
Rating: 

What to Recommend?
Description: Lazslo arrives with Ilsa, Rick's one time love. Rick is very bitter towards Ilsa, who ran out on him in Paris, but when he learns she had good reason to ....
Rating: ★★★★

Description: A secret government project to create genetic mutants results in them being released into the general population. One of the scientists responsible.....
Recommend: ? No

Content Based Filtering Filtering

Many Applications
• Stock trader who is interested in specific financial news (e.g., news about big oil companies)
• Intelligent agent who is interested in foreign news about terrorists (i.e., maybe cross-lingual filtering)
• Researchers who are interested in call for papers, call for proposals
• You are interested in job postings!
Content Based Filtering

Many Applications
• Or if you are lazy….

Key Steps for Content-Based Filtering

1. Initialization (e.g., text)
2. User Profile:
3. Filtering System Decision Making
4. Learning

yes

no
Three Main Problems in CBF

- **Initialization**: Initialize the user profile with several key words or very few examples
- **Delivery Decision Making**: new documents → yes/no based on current user profile
- **Learning**: utilize relevance feedback information from users to update user profile (only on “recommended” documents)

Evaluation

- **F measure**
  \[
  F = \frac{(1 + \beta^2) \text{Precision} \times \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}}
  \]

- **Utility function (utility gained by the user)**
  - Each delivered doc gets a utility value
  - Relevant doc gets a positive value
  - Irrelevant doc gets a negative value
  - E.g., Utility = 5* #good - 1 *#bad (linear utility)
• Content Based Filtering as retrieval
  – Rank the incoming documents
  – Select the top k ranked docs to delivery for a user
  – Problems?

• Content Based Filtering as categorization
  – Binary classification of a document to relevant or not-relevant
  – Delivery docs classified as relevant to a user
  – Problems?
Differences

- Content Based Filtering as retrieval
  - Select the top k ranked docs to delivery for a user?
  - *Have to make a decision for each document, there is no ranked list*

- Content Based Filtering as categorization
  - Binary classification of a document to relevant or not-relevant
  - *Very limited amount of training data at the initial stage; what if the learned rule is to strict at the beginning?*

Content Based Filtering

- Content Based Filtering as retrieval
  - Use retrieval method and query (profile) to score a document
  - Use a threshold to make delivery decision
  - Improve the query (i.e. profile) with feedback information
  - Different approaches for threshold setting and learning

- Content Based Filtering as categorization
  - Use unbalanced, binary categorization
  - Use a threshold to make delivery decision
  - Trained with unbalanced data
  - Different approaches for initialization
General Framework

Filtering System
Decision Making
User Profile
Threshold
Learning

User Judgment
yes
no

Difficulties in Threshold Learning

<table>
<thead>
<tr>
<th>retrieval model</th>
<th>Classification model with probability output</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5 R</td>
<td>E.g., Utility = 5 * #good - 1 * #bad</td>
</tr>
<tr>
<td>10.4 N</td>
<td></td>
</tr>
<tr>
<td>10.1 R</td>
<td></td>
</tr>
<tr>
<td>8.2 ?</td>
<td>0=10.0</td>
</tr>
<tr>
<td>7.1 ?</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Difficulties
- Very little or even no training data
- Exploitation and exploration
Threshold Learning: Logistic Regression (Robertson & Walker. 00)

Estimate the probability of relevance for each doc

\[
\log \frac{P(\text{Rel} | d)}{1 - P(\text{Rel} | d)} = \beta_0 + \beta_1 s(d)
\]

- Big step beyond heuristic thresholding

But

- Unbalanced data
- Few or even no positive feedback at beginning
- Does not address the issue of exploration

Directly Optimize the Utility

- What is the utility function?
  - We should consider the current document.
    But what about exploration? How can we represent that factor?
- Empirical utility optimization
Direct Utility Optimization

• Given:
  – A utility function $U=\{U_R^+, U_R^-, U_N^+, U_N^-\}$
  – Training data $D_i=\{d_i, \{R,N,?\}\}$
  – Empirical utility can be represented as a function of threshold (e.g., $U=F(\theta)$)
  – Choose the threshold to maximize empirical utility
    $\theta^* = \arg \max_{\theta} F(\theta)$

Difficulties in Threshold Learning

Threshold setting example

Classification model with probability output

E.g., Utility = $5\times \#\text{good} - 1\times \#\text{bad}$

Relevant Delivered
Irrelevant Delivered

How to set $\theta$?
Empirical Utility Optimization

• Maximize Empirical Utility
  – Compute empirical utility on training data
  – Choose the threshold that gives the maximum empirical utility

• Problems
  – The training data is biased; more positive documents and less negative documents; often the learn threshold is an upper bound for the true optimal one (why?)

• Solutions
  – Heuristic adjustment (lower the threshold)
  – More sophisticated modeling

Score Distribution Approaches
(Aramptzis & Hameren 01; Zhang & Callan 01)

➢ Training data \( D_i = \{ d_i, \{ R, N, ? \} \} \)

\[
\begin{align*}
\arg \max \sum_i \log \left( P(D_i \mid H) \right) \\
= \arg \max \sum_i \log \left( P\left( \text{Score} = \text{Score}_i, R_i \mid H, \text{Score} > \theta \right) \right)
\end{align*}
\]

Use two types of score distribution of relevant and irrelevant documents to approximate
Exploration versus Exploitation

- The models introduced only consider utility for the current documents
- What about the current model is wrong? Always delivery all documents? Always delivery nothing?
- It is very important to explore in the early stage
- Incorporate model uncertainty into the utility