Content Based Filtering

Luo Si

Department of Computer Science
Purdue University
Content Based Filtering

Outline

- Introduction to content based filtering
  - Applications
  - main research problems

- Main framework

- Learn threshold
Content Based Filtering

Information Needs are Stable
System should make a delivery decision on the fly when a document “arrives”
Content Based Filtering

Google News

bailout plan

EU approves German financial bailout plan
International Herald Tribune - 7 hours ago
The commission has to approve such aid plans to ensure they do not distort competition in the EU's single market. German lawmakers passed the rescue plan ...
EU Gives Green Light to German Bank Bailout Plan Deutsche Welle
EU approves German bailout scheme Financial Times
Straits Times Xinhua Euroalert.net More (25) »

Hackett discusses bailout plan, energy independence and the War on ...
Towanda Daily Review - 11 hours ago
BY JAMES LOEWENSTEIN In an interview, congressional candidate Chris Hackett said that he would have "led a charge" for a better bill than the $700 bailout ...
Carney, Hackett favor reducing federal debt Towanda Daily Review
More (10) »

No easy bailout plan for Devils
Arizona Republic - Oct 26, 2008
Staring at his fourth losing season in 20 years as a college head coach, the 2007 Pac-10 Coach of the Year is confident in his fix-it ability, even if his ...
Early line has OSU favored by nearly two TDs ... ASU hasn't ... The Oregonian - OregonLive.com
The Oreg... More (210) »
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
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: ★★★★★

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: ? Yes
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 Filtering

Many Applications

- Even if you are lazy….
Content Based Filtering Filtering

Key steps for CBF

Initialization

User Profile:

Filtering System
Decision Making

Learning

Initial Profile (e.g., text)

yes

no
Three Main Problems in CBF

Three problems

- **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 of CBF System

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 Filtering

Difference between CBF and Retrieval, Categorization

Content Based Filtering as retrieval

- Rank the incoming documents
- Select the top k ranked docs to delivery for a user
- Problems?
Content Based Filtering Filtering

Difference between CBF and Retrieval, Categorization

Content Based Filtering as categorization

- Binary classification of a document to: relevant or not-relevant
- Delivery docs classified as relevant to a user
- Problems?
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 lists....

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 about the learned rule is very strict at the beginning?
Content Based Filtering 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
Content Based Filtering Filtering

General Framework

User Profile

Threshold

Filtering System
Decision Making

Learning

User Judgment

yes

no
**Difficulties in Threshold Learning**

### Classification Model with Probability Output

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>10.4</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>10.1</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>8.2</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

\[ \theta = 10.0 \]

- **Classification model with probability output**
  - E.g., Utility = 5* #good - 1 * #bad

### Difficulties

- Very few or even no training data
- Exploitation and exploration
Threshold Learning

Logistic Regression (Robertson & Walker. 00)

Estimate the probability of relevance for each doc

\[
\log \frac{P(\text{Rel} | \vec{u})}{1 - P(\text{Rel} | \vec{u})} + \beta_s(\vec{u})
\]

- Big step beyond heuristic thresholding

But

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

Direct 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
Given information

- 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* #good - 1 *#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
Empirical Utility Optimization

Score Distribution Approaches

( Aramptzis & Hameren 01; Zhang & Callan 01)

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

$$\text{arg max } \sum_i \log \left( \frac{P(D_i | H)}{\sum_h P(D_i | H)} \right)$$

$$= \text{arg max } \sum_i \log \left( P(\text{Score} = \text{Score}_i, R_i | H, \text{Score} > \theta_i) \right)$$

Use Two types of score distribution of relevant and irrelevant documents to approximate
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.
Outline

● Introduction to content based filtering
  ➢ Applications
  ➢ main research problems

● Main framework

● Learn threshold