Email Agent: Information Filtering

After the Presentation, You can Talk
- What are differences between information filtering and information retrieval?
- What is an ideal text filtering system?
- Where are the machine learning techniques used in the Email agent?
- Why is the social filtering needed?

Content
- Email agent
- Information filtering systems
- Information filtering vs information retrieval
- Four basic components in IF
- User modeling
- Machine learning
- Case study
- Conclusion

Role of Email Agent
- Drop unwanted Emails (junk mails)
- File incoming Emails into proper folders
- Additional works
  - automatic responses if it can be autonomous
  - alarms for important messages
  - forward an urgent message to SMS
  - and other useful things that you can think of

Process of Email Agent
- Analyze an incoming Email
  - how to represent the result of analysis?
- Compare it with user’s interest
  - how to describe user’s interest?
  - how to compare between them?
- Do proper actions for the email
- Get feedback from the process
  - how to feedback?

Information Filtering Systems
- are typically designed to sort through large volumes of dynamically generated information and present the user with sources of information that are likely to satisfy his or her information requirement
Information Filtering

- Deals primarily with text information
- Involves large amount of data
- Involves streams of incoming data
- Is based on descriptions of individual or group information preferences (profile)
- Is often meant to imply the removal of data from incoming stream

Information Filtering VS. Information Retrieval

User preference
- filtering: long-term interest
- retrieval: short-term interest

Applied data
- filtering: streams of incoming data
- retrieval: static data like items in database

Object
- filtering: removing information from a stream
- retrieval: finding information from a stream

Four Basic Components in IF

- Representation of the document
- Representation of the information need (i.e., profile construction)
- Comparing the profiles with the document representation
- Using the result of that comparison

Text Filtering System Model

In an ideal text filtering system, $d(f(\text{info need}), d(\text{doc})) = \frac{1}{\text{Number of documents}} \sum \text{Occurrences of term } t \text{ in document } d(\text{doc})$.

Representation of the Document

- Information derived from the document
  - simple word-based
    - lexical scanning to identify words, morphological analysis to reduce different word forms to common “stems”, counting occurrences of those stems
    - term-frequency–inverse document frequency
  - natural language-based
    - recognizing noun phrases using syntactic or stochastic parsing
- Annotations made by others
Comparison Function c

- Binary judgement
  - whether a document satisfy the profile
- Vector space method
  - the degree to which the content of two document is similar
- Probabilistic method
  - the probability that a document satisfies the information need represented by the profile
  - bayesian inference networks

User Modeling

- A discipline which is concerned with both how information about users can be acquired and used by automated systems
- Methods to describe a person’s interests
  - provide a set of words to describe their interests
  - articles they have read in the past
  - what organization they work in
  - which books they have ordered
  - familiarity, novelty, importance or urgency

Difficulties in Modeling User’s Interests

- Vocabulary problem
  - the information available are different from those the user would use to specify his or her interest
- Conceptualization problem
  - the concepts used to represent available information are different from the concepts the user has for the domain

Inference Network

- Explicit models
  - constructed explicitly by the user
  - increase the cognitive load on the user
- Implicit models
  - abstracted by the system on the basis of the user’s behavior

Machine Learning

- Supervised learning problem
  - a machine is learning from differences between prediction and user’s judgement
- Machine learning approaches
  - rule induction, instance based learning, statistical classification, regression, neural networks, genetic algorithm
TREC evaluation

- Precision
  - the fraction of the selected documents which are actually relevant to the user’s information
- Recall
  - the fraction of the actual set of relevant documents that are correctly classified
- Fallout
  - the fraction of the non-relevant document that are selected

<table>
<thead>
<tr>
<th>Selected as</th>
<th>Actually is</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>Relevant</td>
<td>Not Relevant</td>
</tr>
<tr>
<td>Not Relevant</td>
<td>Not Relevant</td>
<td>Relevant</td>
</tr>
<tr>
<td>Precision:</td>
<td>Found</td>
<td>Found and false alarms</td>
</tr>
<tr>
<td>Recall:</td>
<td>Found</td>
<td>Found and misses</td>
</tr>
<tr>
<td>Fallout:</td>
<td>False alarms</td>
<td>False alarms correctly rejected</td>
</tr>
</tbody>
</table>

Table 2: Measures of text selection effectiveness.

Information Filtering Systems

- SDI (Selective Dissemination of Information)
  - automatically inform new documents published in their areas of specification to scientists
  - use the profile to match the keywords against new articles if it is relevant to scientist’s interests
- Allen’s predicting preferences experiment
  - articles a person would read based on a measure of overlap of nouns between new and old articles
  - the average correlation is fairly low (r=0.44)

Content-Based Filtering

- SIFT
  - manually construct profiles by specifying words to prefer or avoid, and updated by the user
    - accepting or reject the present day’s article
    - select additional words to be avoided
  - no distinction among the words in articles

- InfoScope
  - automatic profile learning by observing user actions, which minimizes the cognitive load of managing the information filtering system
  - key contribution
    - machine-assisted profile learning,
    - the addition of user-controlled levels of abstraction
    - implicit feedback
Social Filtering
- Tapestry by Nicols and others at PARC
  - users manually construct profile based on document content and on annotations made regarding those documents by other users
  - users associate a score with each rule
  - implementation: client-server model
    - server applies simple rules to determine whether each document may be of interest to each user
    - more sophisticated rules in each profile are executed in each user’s workstation to develop the ranked list

GroupLens
- content servers: standard news server
- annotation servers
  - register user’s annotation for the article
  - collect annotations and predict unseen articles
- contributions
  - distributed annotation servers
  - profile learning for social filtering

Maxim
- desperation based communication
  - invoked when a particular agent has insufficient experience to make confident prediction
- exploratory communication
  - find the best set of peer agents to ask for help in certain classes of situation
  - registration, locating peers, collaboration

Limitation of Social Filtering
- User motivation
  - no incentive for the first user for annotation
  - “chicken and the egg” problem
  - mixing content-based and social filtering
- URN: Internet News Filtering
  - content-based judgements about the document
  - collaboratively improve the system’s initial representation of the document by adding or deleting words

Case Study in Real World
- ZyALERT
  - routing and delivering the information of interest to user’s email
  - content-based filtering
- Procmail, mailfilt, elm filter, maildrop
  - support languages which filter messages or do actions if user specified rules are met

Outlook Express 4.0
Cont’d

Netscape Composer 4.5

Eudora 4.3

Conclusion

- Filtering is difficult to use and not practical
- Client part
  - implicit user action modeling
  - learning and making profile from user actions
    - read, delete, move, forward, reply
    - check time consumption at or until a certain action
  - compare predictions with user actions

- Server part
  - email server collects user profiles
  - if a new user is arrived, a common profile in the same domain is used as the basic profile
  - junk mails are automatically removed
  - prioritize the messages and forward them to multiple accounts by considering urgency
    - email, instant message, PDA, hand phone