Information Filtering

Lecture 14

Information Filtering

- Given
  - a stream of documents (news articles, CDs)
  - a set of users (with stable and specific interests)
- Recommend documents to users who will be interested in them
  - "Tell me when a good jazz CD comes out."
  - "Tell me when an airplane crash is reported."
  - "Send me abstracts on learning text classifiers."
  - "Only show me the good articles on USENET."

Features of a Filtering System

- handle structured and unstructured data
  - primarily text, but also multimedia
- handle large streams of data (e.g. newswire)
- based on a user profile that describes information interests
  - profile is learned from user feedback
  - as interests change, profiles are modified

Selection and removal

- Panning for gold
  - "I only want things like this."
  - Out of a large stream, return best matching documents
  - Perhaps further sorted into categories
- Throwing out the trash
  - "Don't give me more of this."
  - Remove data from the stream
  - e.g. Spam removal
- A profile can express what the user wants, and what he doesn't want.

Filtering Email with Procmail

```
:0 :
* ^Sender:.*linux-kernel
linux

:0 :
* ^Sender:.*csee-grad
/dev/null

:0 B:
* funny|joke|lightbulb|bar
!ian@home.com
```

- Separating out mail from a list into a folder "linux"
- Getting rid of spam
- Forwarding mail to another address

Filtering Email with Gnus

```
(setq nnmail-split-methods 'nnmail-split-fancy
      nnmail-cache-accepted-message-ids t)
(setq nnmail-split-fancy
  '(| (: nnmail-split-fancy-with-parent)
      ("subject" "\\[webir\\]” "mail.research.ir-lists")
      (any "p2p-hackers" "mail.research.p2p")
      (any "jobs@cra\\.org" "mail.cra")
      ;;; ... other specific rules for lists ...
      ("X-Spam-Flag" "YES" "mail.spamgate")
      (to "nist" "mail.misc")
      (: ifile-spam-filter
        "mail.misc")))
```

Filtering in TREC 2002

- Adaptive Filtering task
  - 50 search topics
  - Reuters Corpus vol. 1
- Given a topic title and three relevant examples,
  - Process each document in timestamp order
  - Make a binary decision on whether to retrieve the document
    - If retrieved, you get the relevance judgment if it exists
    - Profile may then be modified

Related Ideas

- **Selective Dissemination of Information** (SDI): deliver documents to individuals based on information profiles
- **Topic Detection and Tracking**: notice when a new event happens, and get all subsequent stories on it
- **Text Routing**: given a starter profile, send relevant incoming documents to the right users
- **Text Categorization**: sort documents into defined categories
- **Extraction**: identify and extract facts from documents

Routing in TREC-8

- Routing task: part of the Filtering track
  - 50 search topics
  - Financial Times (FT) collection, 1992-4
- Given:
  - Topic descriptions
  - Any FT relevance judgments from 1992
  - Any TREC data not in FT93-4
- Rank the top 1000 documents in FT93-4 for each topic

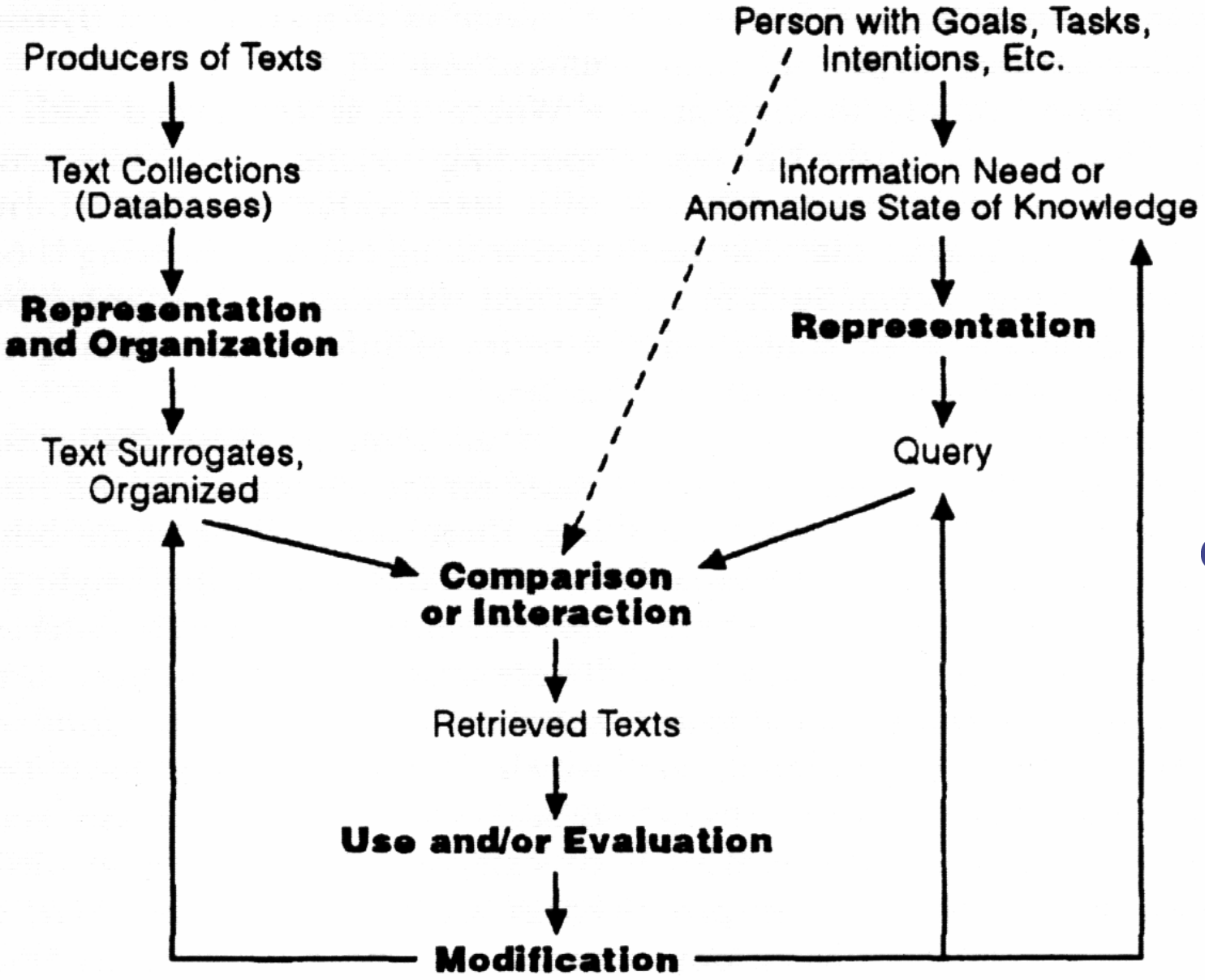
Categorization with Reuters-21578

Goal: predict category labels for new documents

<REUTERS>
<DATE>26-FEB-1987 15:01:01.79</DATE>
<TOPICS><D>**cocoa**</D></TOPICS>
<PLACES><D>**el-salvador**</D><D>**usa**</D><D>**uruguay**</D>
</PLACES>
<TEXT>
<TITLE>BAHIA COCOA REVIEW</TITLE>
<DATELINE> SALVADOR, Feb 26 - </DATELINE>
<BODY>Showers continued throughout the week in
the Bahia cocoa zone, alleviating the drought since early
January and improving prospects for the coming temporao,
although normal humidity levels have not been restored,
Comissaria Smith said in its weekly review …

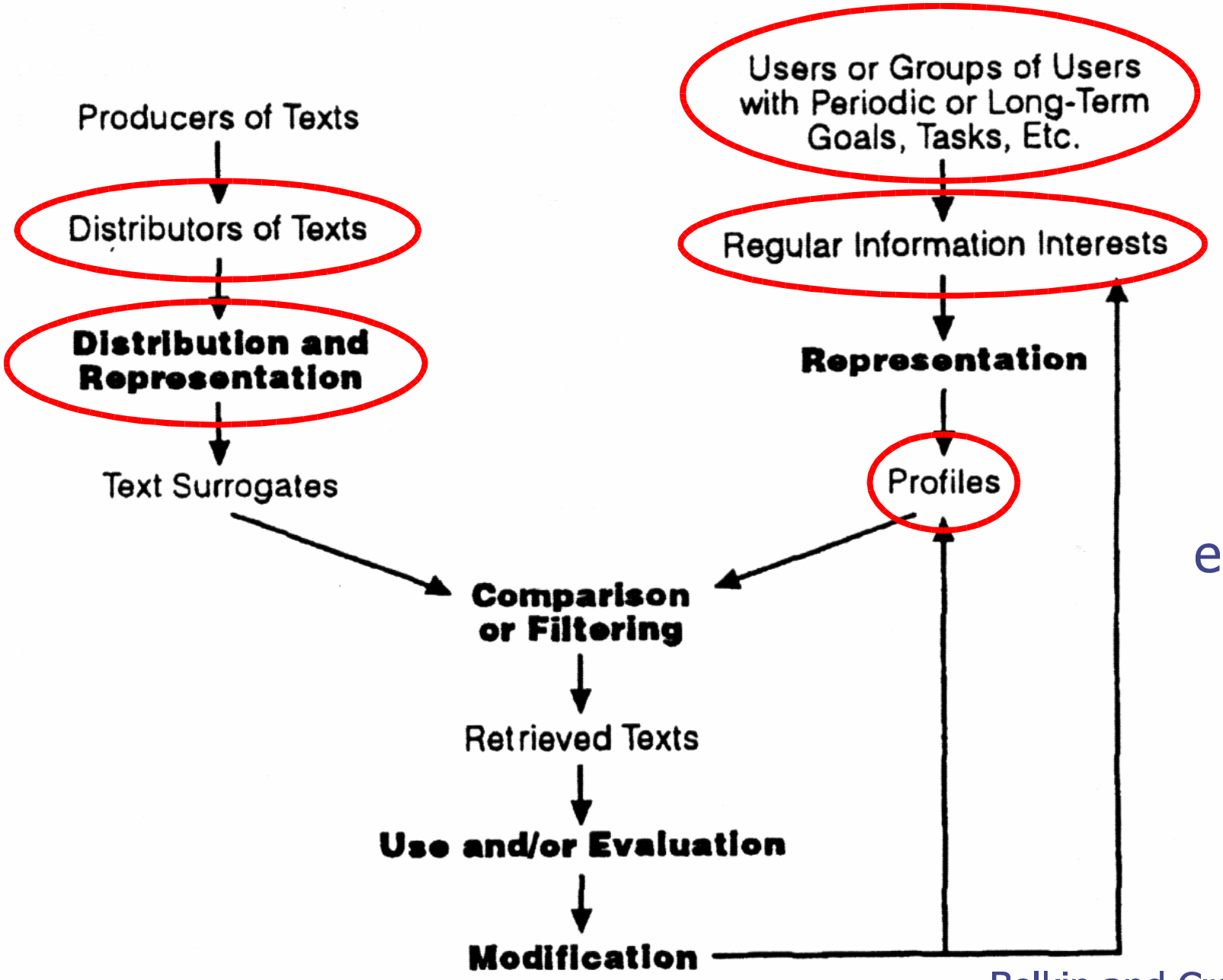
Retrieval and Filtering

- Retrieval
  - **Select** relevant documents
    - from a static collection
    - which has been indexed, categorized…
  - In response to an **ad hoc query**
    - formed from a current need
- Filtering
  - **Deliver** incoming documents
    - from a high-volume stream
  - To a set of **stable user interests**
    - represented by profiles
    - which are learned over time



Model of **retrieval** entities and process

Belkin and Croft '92



Model of **filtering** entities and process

Belkin and Croft '92

Retrieval vs. Filtering

1. one-time goal	1. evolving need
2. ad hoc queries	2. user profiles
3. query representation	3. profile representation
4. document organization	4. document distribution
5. static collection	5. document stream

Unique needs of filtering

1. Timeliness of results
2. Different user model
   - Needs are more loosely defined
   - Information may be less critical
   - User is not motivated to actively search
3. Privacy and security
   A. What would you do with the profiles of 10,000 customers?

Implementing Filtering

- retrieval: compare query to multiple documents
- filtering: compare document to multiple profiles
- Key ideas
  - maintain a collection of profiles
    - built from need statement or sample documents
    - collect user feedback
    - learn profiles using Rocchio's algorithm
  - As documents arrive…
    - Compare incoming document to each profile
    - Recommend if similarity exceeds a certain threshold
  - System needs to learn both profile and threshold

Building Profiles

- IR view: relevance feedback
  - compute optimal query given relevance information
  - select terms and adjust weights
- Relevance feedback is a learning algorithm
  - learn function that optimizes retrieval results
  - supervised learning
- Relevance feedback is not the only learning algorithm!

Learning Text Profiles

- Machine learning approaches
  - Neural networks
  - Rule-based learning
  - Bayesian learning
  - Boosting
  - Support Vector Machines
- LSI
  - identifying document features
  - clustering to create profiles

Evaluation

- Can we evaluate filtering like retrieval?
- Documents are in a stream
  - no ranked list, unless system batches documents
  - no retrieved "set" if we filter each document alone
- Profiles are learned over time
  - learning rate
  - change in error rate
- User interests evolve over time
  - does the system keep up?
  - what is the average performance? best?

Filtering Evaluation Measures

- Utility
- Set-based metrics
  - set precision
  - F measure
  - mean absolute error
- Detection error tradeoff
  - misses vs. false alarms