Adaptive Information Filtering
& its Application in Medical Domain

Lanbo Zhang, Yi Zhang, Carla Kuiken

Motivation

- Some users may want to be kept in touch with up-to-date information of a particular topic.
  - IT financial analysts: reports/news/discussions related to companies in their stock portfolio
  - Researchers: most recent papers in related fields
- Some users may want to be alerted of sensitive information
  - FBI investigators: documentation that contains terrorism information

Motivation - In Medical Domain

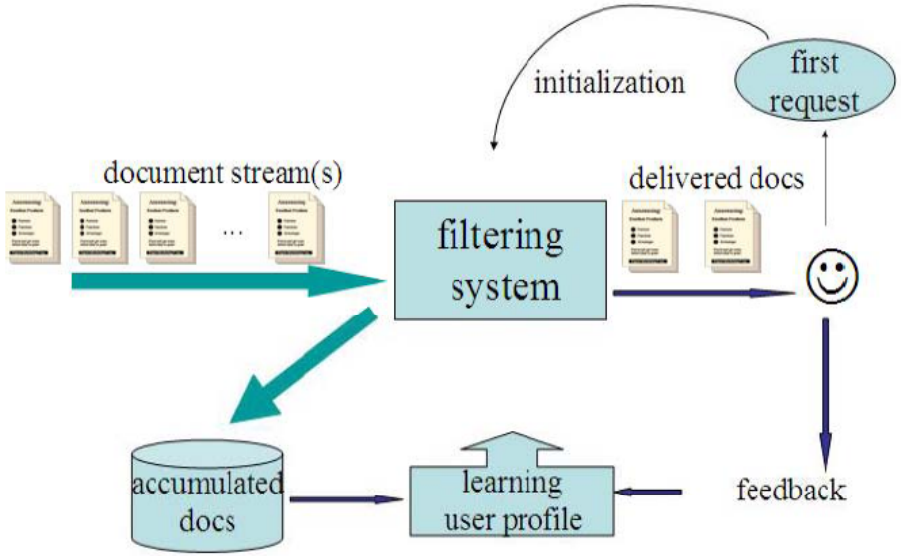
- A physician m
  - New disc
  - Patien
- A me
  - Most
- CDC p
  - Reports of
  - Up-to-date infor

In these cases, users want an intelligent system that can **push** interesting information to them.

Lanbo Zhang, Yi Zhang (UCSC advisor), Carla Kuiken (LANL mentor)

Information Filtering System

- An autonomous agent that delivers information to interested users whenever it is available.
  - Users input some keywords or document examples to represent their information needs.
  - Users can give some feedback on delivered information to help filtering system perform better.



Lanbo Zhang, Yi Zhang (UCSC advisor), Carla Kuiken (LANL mentor)

User Profile in Filtering

- A filtering system maintains a profile for each user, which stores the user's information needs (interests)
- The user profile is learnt based on
  - Initial keywords/document examples
  - User feedback
- A filtering system determines whether to deliver a document to a user based on
  - Comparison of the user profile and the document content.

Interactive User Profile Initialization

- Challenge
  - How to learn good user profiles with limited user feedback (training data)
- Two mechanisms we tried
  - Relevance feedback
  - Faceted feedback

Relevance Feedback

- User gives a Boolean answer ("yes" or "no") of whether a document is relevant to his/her information need.

Keywords: **lung cancer**

Treatment of stage II lung cancer (T1N1 and T2N1)… — Yes

….

Transconjunctival oxygen monitoring as a predictor of hypoxemia during helicopter transport… — No

….

Diagnosis and treatment of small cell carcinoma of the larynx: a critical review… — Yes

How to Use Relevance Feedback (1): Vector Space Model

- User profiles and documents are represented as term(word) vectors
  - Each dimension: TF*IDF
  - TF: term frequency
  - IDF: inverted document frequency $idf(t) = \log \frac{N}{n(t)}$
- Rocchio Algorithm

$$\mathbf{Q}' = \alpha \cdot \mathbf{Q} + \beta \frac{\sum_{\mathbf{x_i} \in R} \mathbf{x_i}}{|R|} - \gamma \frac{\sum_{\mathbf{x_i} \in NR} \mathbf{x_i}}{|NR|}$$

  - Q: initial user profile
  - X_i: document vector
  - R: a set of labeled relevant documents
  - NR: a set of labeled non-relevant documents

How to Use Relevance Feedback (2): Bayesian Logistic Regression Model

- The prior is estimated with Rocchio algorithm
- Performs better than Rocchio algorithm when more user feedbacks are available

$$\alpha^* = \underset{\alpha}{argmax} \sum_{i=1}^{t} \frac{1}{1+\exp(-y_i \alpha \mathbf{w_R^T x_i})}$$

$$\mathbf{m_w} = \alpha^* \cdot \mathbf{w_R}$$

$$\begin{aligned} \mathbf{w}_{MAP_t} &= \underset{\mathbf{w}}{argmax} P(\mathbf{w}|D_t) \\ &= \underset{\mathbf{w}}{argmax} \prod_{i=1}^{t} P(y_i|\mathbf{w}, \mathbf{x_i}) P(\mathbf{w}) \\ &= \underset{\mathbf{w}}{argmax} \sum_{i=1}^{t} log(1+\exp(-y_i \mathbf{w}^T \mathbf{x_i})) - v_w(\mathbf{w}-\mathbf{m_w})^2 \end{aligned}$$

Faceted Feedback

- Let users give feedback on document metadata
- Each type of metadata is a facet, like source, MeSH, Author

```
.I 1
.U
87049087
.S
Am J Emerg Med 8703; 4(6):491-5
.M
Allied Health Personnel/*; Electric Countershock/*;
Emergencies; Emergency Medical Technicians/*; Human;
Prognosis; Recurrence; Support, U.S. Gov't, P.H.S.; Time
Factors; Transportation of Patients; Ventricular
Fibrillation/*TH.
.T
Refibrillation managed by EMT-Ds: incidence and outcome
without paramedic back-up.
.P
JOURNAL ARTICLE.
.W
Some patients converted from ventricular fibrillation to
organized rhythms by defibrillation-trained ambulance
technicians (EMT-Ds) will refibrillate before hospital
arrival. The authors analyzed 271 cases of ventricular
...
.A
Stults KR; Brown DD
```

source journal

MeSH

authors

Lanbo Zhang, Yi Zhang (UCSC advisor), Carla Kuiken (LANL mentor)

Faceted Feedback

- A user checks relevant values of a facet type

Which of the following MeSH(s) are you interested in?

- ☐ Pregnancy Complications/*TH
- ☐ Carbon Monoxide Poisoning/CO/*TH
- ☐ Pregnancy Trimester, Third
- ☐ Respiration, Artificial
- ☐ Respiratory Distress Syndrome, Adult/ET/*TH

To restrict your interested articles to be from specific journal(s), please identify the journal name(s).

- ☐ Am J Emerg Med
- ☐ ASAIO Trans
- ☐ Br J Anaesth
- ☐ Burns Incl Therm Inj
- ☐ Cardiovasc Clin

How to Use Faceted Feedback (1): Boolean Model

- A document must contain all user clicked facet values to be delivered.

$$s(d) = \begin{cases} s_o(d) & d \text{ contains all the facet values selected by user} \\ -\infty & \text{otherwise} \end{cases}$$

Lanbo Zhang, Yi Zhang (UCSC advisor), Carla Kuiken (LANL mentor)

How to Use Faceted Feedback (2): Soft Model

- Motivations
  - document metadata might have noise
  - Some types of metadata are vague so that users have difficulty to identify them correctly.
- α_f reflects quality of a type of metadata (facet)

$$s(d) = s_o(d) + \sum_f \alpha_f \sum_v \delta(d, f, v)$$

$$\delta(d, f, v) = \begin{cases} 1 & d \text{ contains facet - value pair } f{:}v \\ 0 & \text{otherwise} \end{cases}$$

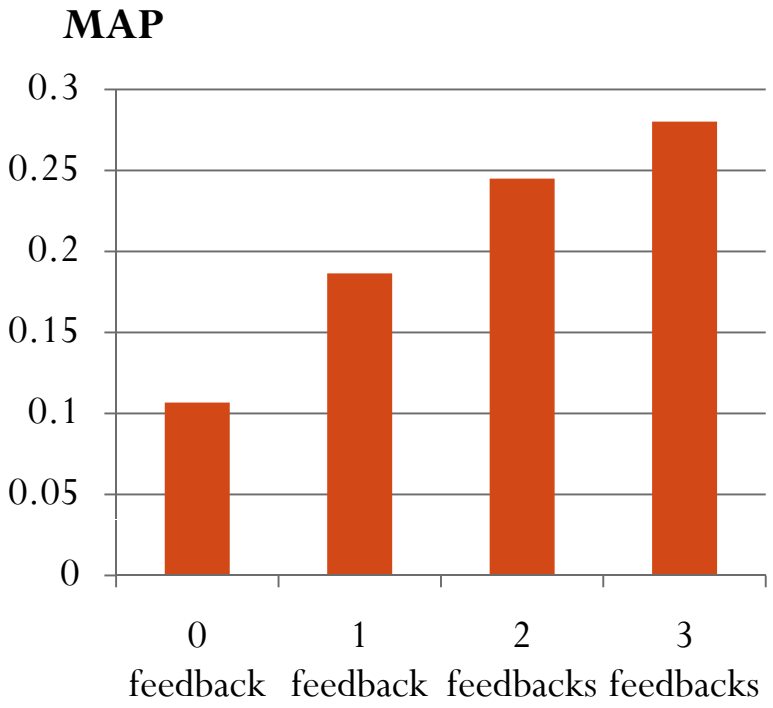
Experimental Dataset

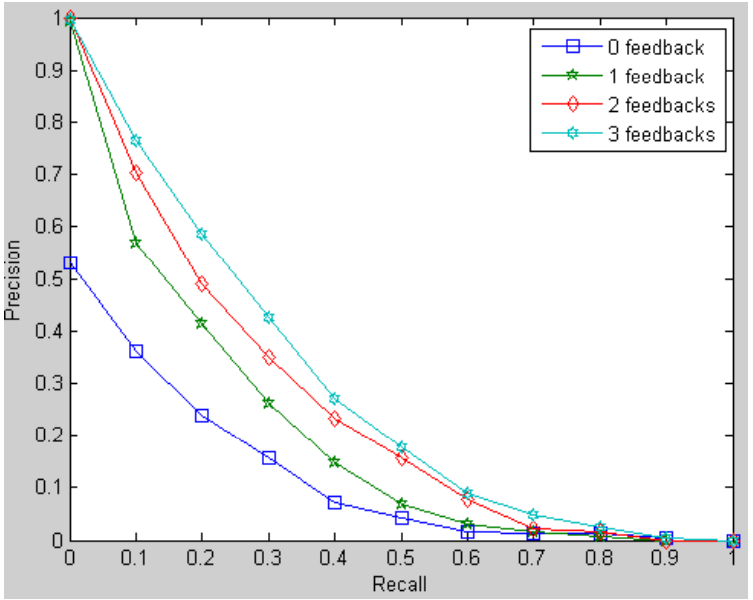
- Corpus: OHSUMED
  - Contains 348566 articles from 270 famous medical journals.
- 63 topics (user information needs)
  - Created by physicians who are experienced in using MEDLINE
- Relevance judgments
  - Documents were assessed for relevance by a different group of physicians
  - Relevance in a three point scale: definitely, possibly, or not relevant.

Evaluation Measures

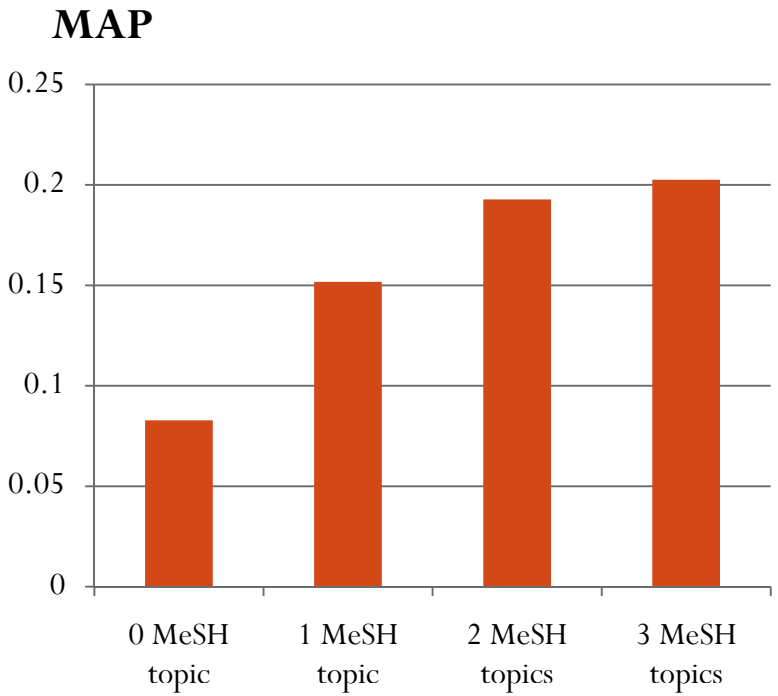
- Precision
  - Proportion of delivered documents that are relevant
- Recall
  - Proportion of relevant documents that are delivered
- MAP
  - Mean of the precisions after each relevant document delivered

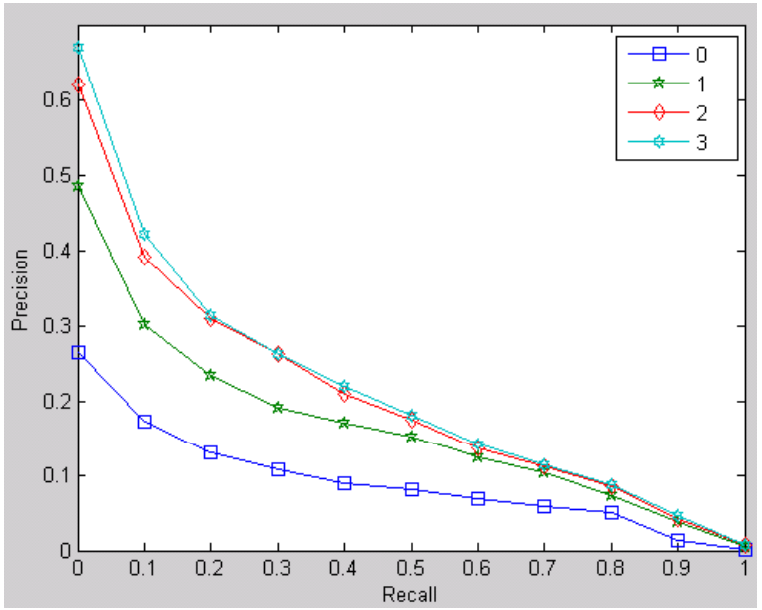
Experimental Results (1)
Relevance Feedback on Documents





Experimental Results (2)
Faceted Feedback on Metadata





Summary and Future Work

- Two interactive mechanisms for learning user profiles
- Evaluate our approaches on a medical dataset
- Next steps
  - Facet value recommendation: active learning methods
  - User profile: feature vector instead of word vector

Questions?