Admin

- Paper review due Friday next week

- 2 versions on one A4 page, side by side
  - ChatGPT version, your version

- Each is about 250 words

  - Why this paper
  - What does it do (academic writing)
  - Why is it relevant to you (academic writing)
  - Prompt for version1, comparason/Evaluation for version2

Personalised search

Personalised information retrieval

A related area is called Adaptive Hypermedia

Also closely related to Web Usage Data Mining

- Web logs, search history
- Common search queries
- Popular pages, dwell time on page

Also closely related to recommender system

recommender: more on item-based

Personalised search: more on user-based

- Two directions: Query adaptation or result adaptation

Information gathering

- Information gathering approach
  - Explicit, Implicit, Both
- Type of information
  - User supplied information
  - User’s categorical interests
  - Queries, clicked documents, snippets of documents
  - Cashed web pages, dwell time on page, desktop documents
  - Email, calendar items
  - Tags and bookmarks on online social applications
- Source of information
  - Server side, Client side, user intervention

Information representation

User model

- Short-term interests, long-term interests
- Static, dynamic, periodic
- Terms or conceptual terms (use WordNet, ontology)
- Vector-based
  - Models where user’s interests are maintained in a vector of weighted keywords (concepts).
- Semantic network based
  - Models where user’s interests are maintained in a network structure of terms and related terms (concepts and related concepts)

Query expansion/adaptation

Resources

- explicit
  - individual relevance feedback, interactive query expansion
- implicit
  - individualised
    - User model
  - Aggregate
    - Usage information (search logs)
  - Not user-focused
    - Pseudo-relevance feedback
    - Thesaurus based (Static or term correlation, co-occurrence)

Query Reformulation

- Revise query to account for feedback:
  - Query Expansion: Add new terms to query from relevant documents.
  - Term Reweighting: Increase weight of terms in relevant documents and decrease weight of terms in irrelevant documents.

- Pseudo-relevance feedback
  - Assume the top N are relevant

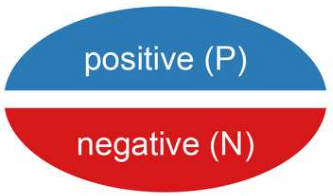
Search results filtering/adaptation

- Different applications: individual, aggregate, web search or recommendation, databases search
- Typically use supervised machine learning
  - Relevant, not relevant: binary classification
  - Training data:
    - Labeled data
    - Assume clicked docs are relevant
  - Machine learning methods
    - KNN: K nearest neighbour
    - Naïve Bayes
    - SVM: Support vector machines
    - Deep learning
- Challenges: time issue, dynamic environment, multiple profiles, new tasks, etc.

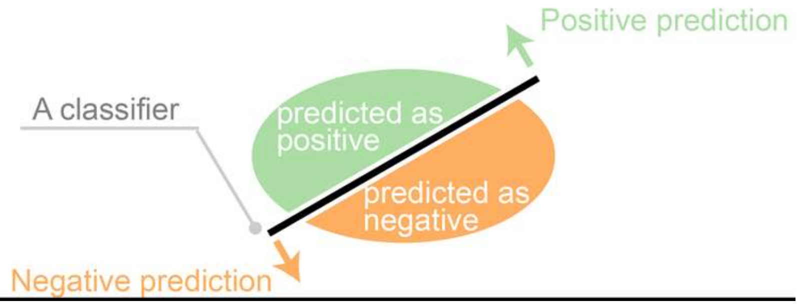
EVALUATION

Classification Systems Evaluation

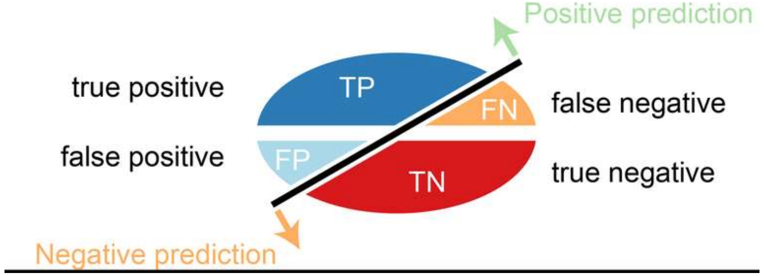
Two actual classes or observed labels



Predicted classes of a classifier



Four outcomes of a classifier



		Predicted class P	Predicted class N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

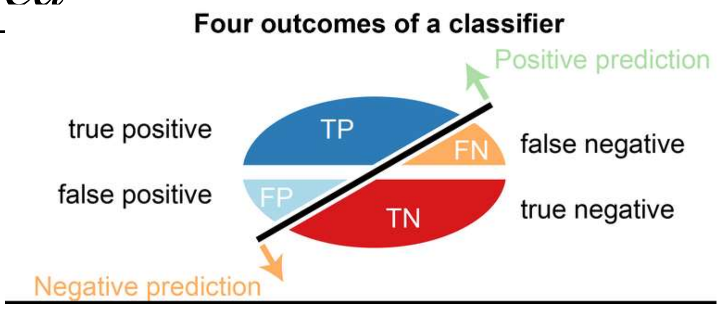
$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$$

$$\mathrm{ERR} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{P} + \mathrm{N}}$$

Information Retrieval Evaluation

- Data collection
  - TREC
  - Queries; documents labelled as relevant and not-relevant
- Evaluation criteria
  - Precision: Percentage of retrieved documents that are relevant

$$P = \frac{\#\,of\ RelevantItems\ Retrieved}{\#\,ofItem\ Retrieved}$$



P = TP/ (TP + FP)

  - Recall: Percentage of all relevant documents that are found by a search

$$R = \frac{\#\,of\ RelevantItems\ Retrieved}{\#\,of\ RelevantItemsInCollection}$$

  - R = TP / (TP + FN) = TP/ P

IR evaluation discussion

- Exercise: calculate precision and recall
  - For a query, If a system finds 200 results, among them 50 are relevant.
  - The human labels (model solutions) have 120 relevant documents.
- Why not use Accuracy or Error rate in IR?

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$$

- Which is more important in Web search: precision or recall?
- How to compare two IR systems

Evaluation: F measure, MAP, AUC

- F-score is a harmonic mean of precision and recall.

$$F_1 = \frac{2 \cdot \mathrm{PREC} \cdot \mathrm{REC}}{\mathrm{PREC} + \mathrm{REC}}$$

- AUC: Area under the precision and recall curve
- Top N precision
- MAP: consider ranking, precision, recall
  - Mean of the Average Precision for all queries
  - Average Precision: the mean of the precision when each relevant document is retrieved. (M is the No of relevant documents)

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \mathrm{Precision}(R_{jk})$$

  - Average precision is roughly the area under the precision and recall curve
- ARR: the average rank of the documents rated as “relevant”

Evaluation in general

- Information retrieval evaluation methods can be used for evaluation in many other areas

- Recommender can be binary: change rates to positive or negative
    - Precision
    - Top N precision
    - Recall
    - F-measure

Personalized Search Evaluation

- In lab setting

  10-500 users

- Quantitative & Qualitative
- System performance
- User evaluation, system usability
- Data sets

  open web corpora, in-lab generated logs,

  TREC collection, search engine query logs

  subset of annotated documents from specific sites

Clustering systems Evaluation

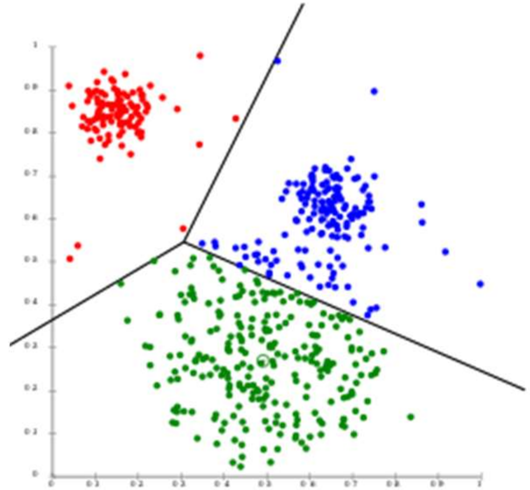
No labels

Labels are not used in training

Use labels only for evaluation

Rand Index = (TP + TN)/ (TP+ TN + FN + FP)

- Typically consider document pairs rather than individual document
- Pair of documents: same class label in the same cluster TP



Recommender Systems Evaluation

- Consider ranking score

- MAE: mean absolute error

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|.$$

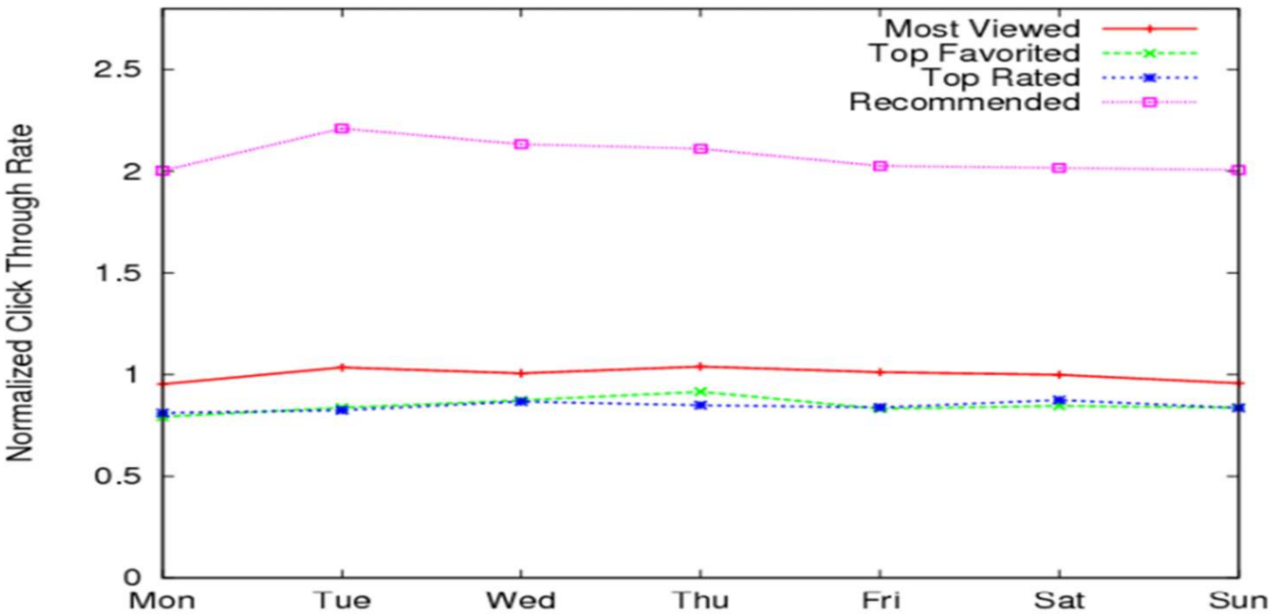
Evaluating Recommendation Quality

- CTR (Click Through Rates): is the ratio of the number of clicks on a video to the number of times that video was seen
- Long CTR: only counting clicks that led to watches of a substantial fraction of the video
- Session Length
- Time until first long watch
- Recommendation Coverage: The fraction of logged in users with recommendation.

The CTR for recommended videos exceeded Most Views, Top Rated, etc

Evaluation

Per-day average CTR for different browse page types over a period of 3 weeks



Evaluation in reality, in practise

- A/B testing
- A/B testing (sometimes called split testing) is experimenting and comparing two types or variations of an online or offline campaign such as a landing page, ad text, a headline, call-to-action or just about any other element of a marketing campaign or ad.
- By displaying two variations of your campaign, you can see which one attracts more interaction and conversions from your customers.
  - e.g. CTR (clickthrough rate): the number of clicks that your ad receives divided by the number of times your ad is shown