Additional Item and List-Based Metrics

Goals for Today

- To understand metrics that relate more closely to business goals and/or user experience, specifically:
  - Coverage
  - Popularity / Novelty
  - Personalization
  - Serendipity
  - Diversity
- To understand why it may be worth trading off accuracy for other purposes …

A couple of stories …

- The perfect supermarket recommender …

Buy Bananas

Maybe Bread Too!

Ziegler's Inspiration …

- Let's see what books Amazon.com recommends that I buy …

Coverage

- Coverage is the measure of the percentage of products for which a recommender can make a prediction
  - Or a prediction that's personalized
  - Or a prediction above a confidence threshold
- Computed as a simple percentage
  - Inconsistent averaging over user/unrated item
  - Easiest is to "hide" every item and compute for entire data set

Use of Coverage

- Directly relevant in cases where predictions are displayed
  - What percentage of movies will I get a star-score for?
- Often used as "background" metric when comparing top-n recommenders
  - Some have extended to ask what percentage of items will appear in *someone's* top-n
- Business interest: reach entire catalog …

Side Note …

- Differences in coverage may lead to a need to adjust computation of other metrics:
  - How do we compare an algorithm with high accuracy and low coverage against one that has low accuracy but high coverage?

Popularity / Novelty

- A simple metric that can be applied to recommendations (and list of recommendations) is the popularity of items recommended:
  - Popularity can be measured as percentage of users who buy/rate the item, or used from external sources.
  - Novelty is often expressed as the inverse of popularity.

Personalization

- Several ways to measure how personalized a recommender is:
  - Variance of predictions per item across users
  - Average difference in top-N lists among users
    - Related coverage metrics
  - User-perceived personalization (survey)

Serendipity

- Definition: "the occurrence and development of events by chance in a happy or beneficial way"
- In recommender systems: surprise, delight, not the expected results
- Several ways to operationalize, such as:
  - $serend = \frac{1}{N}\sum_{i=1}^{N} \max(\Pr(s_i) - Prim(s_i), 0) * isrel(s_i)$
  - Key concept – need prior "primitive" estimate of obviousness, one such metric is overall popularity.

Serendipity in Practice

- Don't need an overall metric to increase serendipity
  - Simply downgrade items that are highly popular (or otherwise obvious)
  - This tends to require experimentation and tuning
- Business goal – get people to consume less popular items

Diversity

- Measure of how different the items recommended are
  - Applied to a top-n list
- Start with a pairwise similarity metric
  - E.g., Ziegler used book categories, others have used tags, keyword vectors
- Intra-list similarity is the average pairwise similarity, lower score is higher diversity

Diversification …

- Common approach is to penalize/remove the items from the top-n list that are too similar to prior items already recommended (never touch #1)
  - Replace with n+1st or later items – first ones that don't exhibit too high a similarity
  - Diversification factor limits how many substitutions will be made

Alternatives to Diversification

- Clustering approaches allow "diversification through bundling"
- Scatter-gather interfaces allow user-controlled diversification
- Business goal: don't turn away customers who are not currently interested in a narrow portion of your catalog

Business Objectives and Metrics

- Users of recommenders have a much broader set of objectives than the metrics we've discussed:
  - Immediate lift
  - Net lift (subtract out cost of returns)
  - Time to next transaction
  - Long-term customer value (lifetime value)
  - Referrals
  - And much more ….

Wrap up …

- Metrics can address "fuzzier" goals such as producing a diverse, delightfully surprising set of recommendations; or assessing whether recommendations go deep into the "long tail" of the catalog and not just a few oft-recommended items.
- Experimental data shows that these objectives may even be worth sacrificing some accuracy.

Looking forward …

- Now that we have an extensive set of metrics, a couple of sessions on how to use them rigorously, and how to apply them to special cases …

Additional Item and List-Based Metrics