商品推薦系統

-評估推薦系統

Lecture 5

講師 :周光宇博士

2018.08

簡介

- 協同過濾的評估與分類和回歸有許多相似之處， 然而，也有獨特的方面。
- 基於內容的方法的評估更加類似於分類和回歸建模。
- 推薦系統的評估往往是多方面的，單一的標準不能捕捉設計師的許多目標。
- 推薦系統可以使用線上方法或離線方法進行評估
- 離線方法是從研究和實踐的角度來看最常用的方法；但是，在設計系統時必須小心謹慎，以使測量的指標從用戶的角度真實反映系統的有效性。

重要的設計問題

- **評估目標：** 單獨的準確度測量是不完整的。 新奇、信任、覆蓋面和驚喜性也很重要。 但是，其中一些因素的實際量化往往是相當主觀的。
- **實驗設計問題：** 即使只評估準確性，我們也需要避免高估或低估。
- **準確性指標：** 預測評分或排名的準確性是最重要的指標。 諸如平均絕對誤差(Mean-Absolute Error)和均方誤差(Mean-Squared-Error)之類的度量常常用於評級預測，而基於效用(Utility-Based)的計算，排序相關係數(Rank Correlation Coefficient)和接收者操作特徵曲線(Receiver Operating Characteristic Curve)被用於評估排序。

評估方式

- 用戶研究
- 線上評估
- 使用歷史數據集進行離線評估

用戶研究

- 用戶被積極招募，他們被要求與推薦系統進行互動，以執行特定的任務。
- 蒐集在與系統互動之前、中、後的回饋訊息。
- 用戶研究的一個重要優點是它們允許蒐集關於用戶與系統實際互動的訊息。
- 可以測試各種情況下關於改變推薦系統對用戶互動的影響，包括改變算法和用戶界面。
- 然而，招募大量用戶是昂貴的，而招募的用戶可能不能代表總用戶人數。
- 因此，用戶評估的結果不能完全信任。

線上評估

- 它也被稱為A / B測試，它是以利潤等目標, 測試系統長期性能之最準確的直接測試。
- 它也是利用用戶研究, 然而用戶常常是完全部署或商業系統中的真實用戶。
- 這種方法有時候比較不容易受到用戶研究招聘過程帶來的偏見，因為用戶通常直接使用目前使用的系統。
- 一個評估效率的典型例子是轉換率 (Conversion Rate)。 如果需要，可以將預期成本或利潤加到商品中，使效率評估包含了每個商品的重要性。
- 除非大量用戶已經註冊了目前使用的系統，否則這種方法不能實際部署。

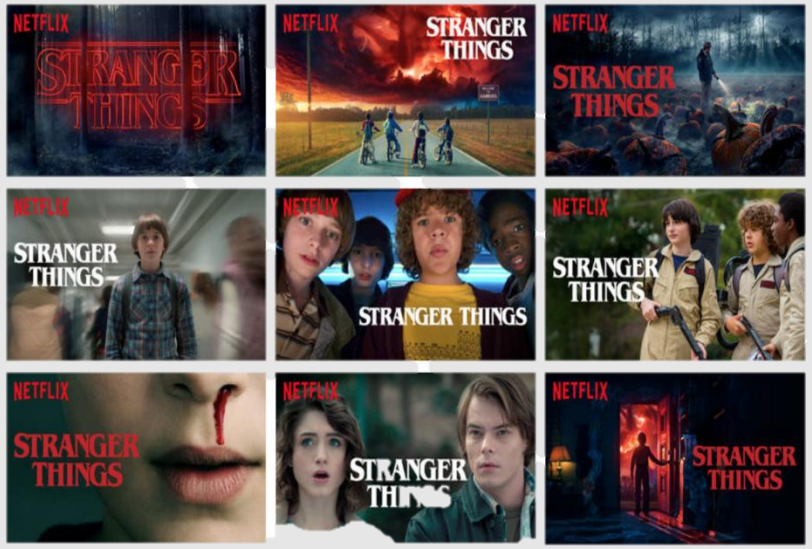
A/B 測試的基本思路

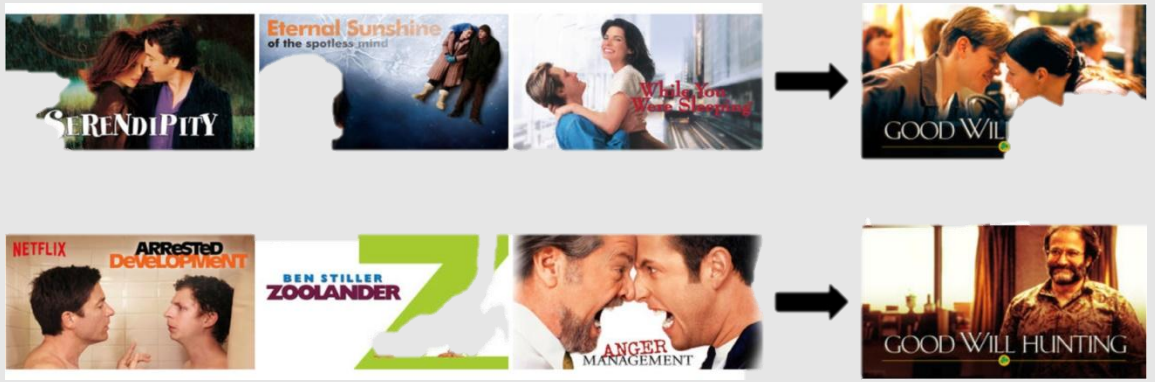
- 比較兩種算法的性能
- 將用戶分為兩組, A和B.
- 在一段時間內對組A使用一種算法，對於組B使用另一種算法，同時保持兩組中的所有其他條件盡可能相似。
- 在過程結束時，比較兩組的轉化率（或其他收益指標）。
- 一個觀察是，在用戶和推薦者之間的每個互動之回報信息可以被分別地測量的情況下，沒有必要嚴格地將用戶分組。 在這種情況下，同一用戶可以隨機顯示其中一種算法，並且可以測量來自該互動的回報信息。

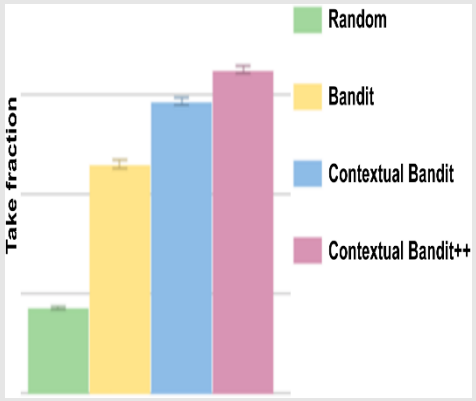
多臂強盜算法 (Multi-Arm Bandit Algorithms)

- 這些A / B 測試的方法被推廣到開發出更有效的推薦算法。他們叫多臂強盜算法。
- 其基本思想類似於賭場（推薦系統），一個賭徒面臨在賭場中選擇一個收益最好的老虎機（推薦算法）。
- 例如，賭徒懷疑其中一台機器比其他機器有更好的回報（轉換率）。 因此，賭徒在10％的時間內隨機嘗試一個老虎機，比較收益。剩下的90％時間選擇收益最好的老虎機。
- 這是探勘和開採(exploration and exploitation) 之間的權衡。

個性化電影海報由算法決定







https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76

使用歷史數據集進行離線評估

- 這是測試推薦算法最受歡迎的技術，因為它們依賴於歷史數據，不需要訪問龐大的用戶群。
- 來自各個領域（例如音樂，電影，新聞）的多個數據集可用於測試推薦系統被廣泛應用的可能性。
- 針對這些情況已經制定了標準化的框架和評估措施。
- 然而，這些評估並沒有衡量用戶對未來推薦系統反應的實際傾向，也沒有對於驚喜性和新穎性的評估。

評估設計一般目標

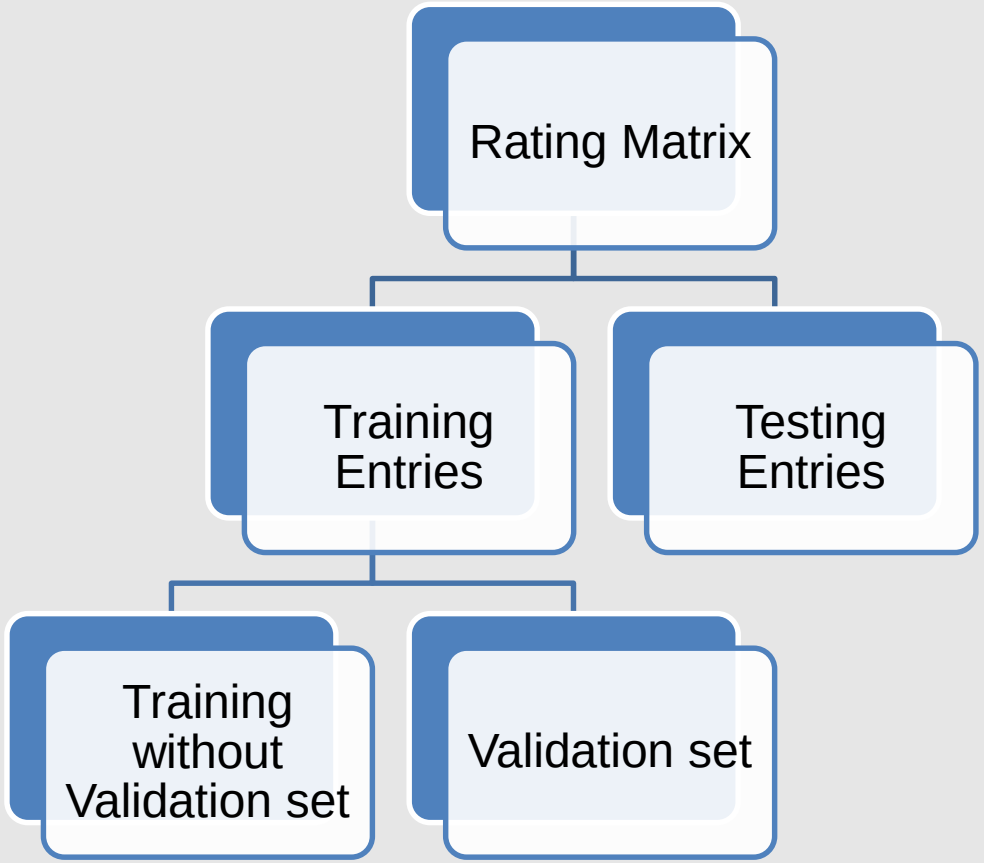
- 準確度
- 覆蓋度
- 置信度 (系統)
- 可信度 (用戶)
- 多樣性
- 驚喜性
- 新奇性
- 穩健性
- 可擴展性

其中一些目標可以具體量化，而另一些則是基於用戶體驗的主觀目標，只能通過用戶調查來衡量。

準確性

- **設計準確性評估:**
  - 使用個別的數據集進行訓練和評估。 例如 將觀察到的評分中的一小部分用來作為評估集合，並使用其餘部分來訓練推薦者系統。
- **準確度指標:**
  - Accuracy of estimating ratings
    - $e_{uj} = \bar{r}_{uj} - r_{uj}$ for user u and item j
    - MSE = $\frac{\sum_{u,j \in E} {e_{uj}}^2}{|E|}$ and RMSE = $\sqrt{\frac{\sum_{u,j \in E} {e_{uj}}^2}{|E|}}$
    - Mean Absolute Error (MAE): $\frac{\sum_{(u,j) \in E} |e_{uj}|}{|E|}$
  - Accuracy of estimated rankings
    - Rank-correlation measures
    - Utility-based measures
    - Receiver operating characteristic
- 準確性指標的主要問題在於，它們通常不會在實際環境中測量推薦系統的真實效果（例如，推薦用戶最終會購買甚至沒有推薦的東西）。

將評級細分為培訓和測試數據集



離線推薦系統中的設計問題

- 在推薦系統準確性評估中常犯的錯誤是使用相同的數據進行參數調整和測試。 這種做法會高估準確性和過度擬合。
- 為了防止這種可能性，通常將數據分為三個部分:
  - 訓練數據 Training data
  - 驗證數據 Validation data
  - 測試數據 Testing data
- 將評級矩陣劃分為比率2：1：1是常見的。
- 當評級矩陣的大小很大時，可以減少驗證和測試的比例。

Model Building	Validation – Tuning, Model Selection	Test Set
50%	25%	25%

分割方法

Hold-Out (隱藏部分數據)

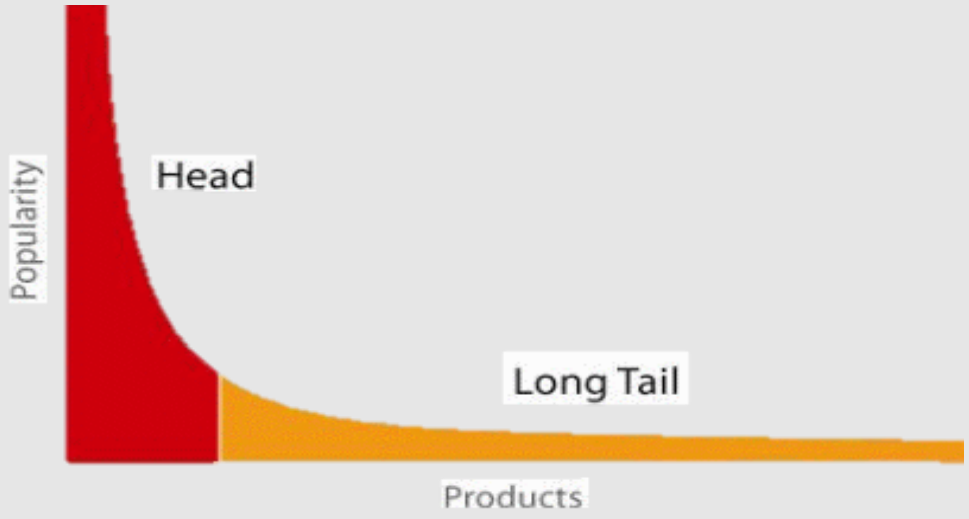
- 隱藏評分矩陣中的一小部分數據來做評估，其餘數據用於構建訓練模型。
- 然後利用訓練好的模型預測隱藏數據的評分並與原來的評分進行比較,它的準確性報告做為整體準確度。
- 這種做法通常低估了真實的準確性。

Cross-Validation (交叉驗證)

- 將評分數據分成q個相等的數據集
- 隱藏其中一個數據集來做評估，其餘（q-1）個數據集用於訓練和驗證。
- 通過使用q個數據集中的每個數據集作為評估集來估準確度，並重複該過程q次。
- 取q個不同評估集的平均準確度做為整體準確度。

長尾現象的影響

- 準確性指標存在的一個問題是，它們深受流行商品評分的影響。
- 然而，長尾上的商品往往會貢獻絕大部分的利潤。
- 處理這個問題的一種方法是分別計算所有測試商品的RMSE或MAE，然後根據這些商品對零售商的相對重要性、利潤、或 效用以加權方式進行平均來推薦。



通過相關係數評估排序

- 這是衡量排名順序準確性的最常見方法。
- 我們想衡量評分的真實排序與推薦系統預測的排序有多類似。
- 由於評分通常是從一些預設的評分中選擇的，並且在真實排序存在許多相同的評分，所以在衡量排名順序準確性時不要懲罰系統在將兩個相同評分的商品排列在前後。
- 兩個最常用的排名順序相關係數是：
  - **Spearman排序相關係數：** 先將真實集和測試集中的所有商品排序，然後計算Pearson相關係數以確定排名的準確性。 如果有相同的評級，則使用相同的平均排名。
  - **Kendall相關係數**：這是衡量測試集中每一對商品的預測排序和真實排序之間的差異度。

Spearman排序相關係數 (Spearman Rank Correlation Coefficient)

- 第一步是對用戶未購買(測試數據)的所有商品根據真正的評分和推薦系統預測分別進行排序。
- Spearman排序相關係數等於應用Pearson相關係數於這些排序。
- 所得到的計算值總是在（-1，+1）範圍內，並且越大的正值越好。

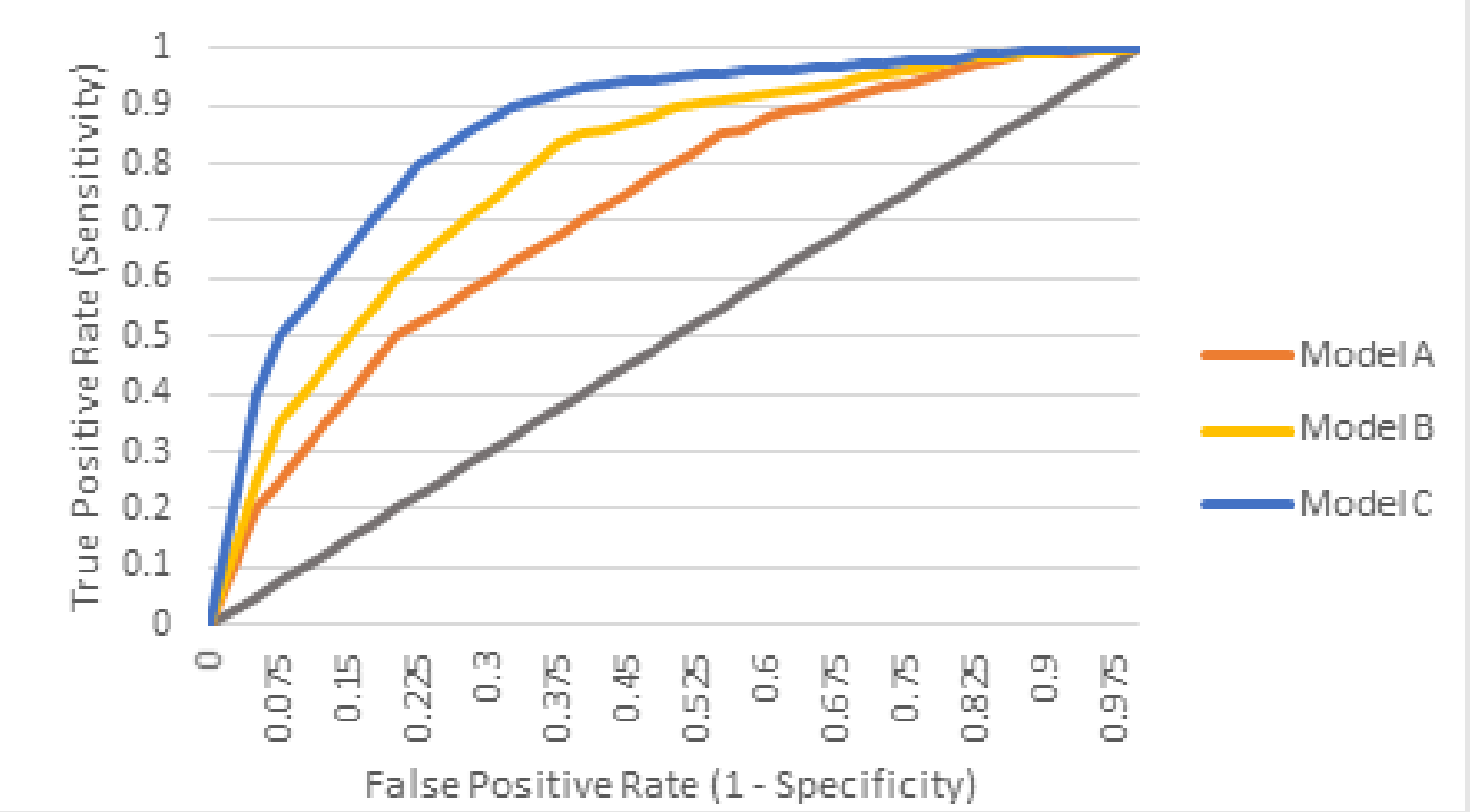
Spearman Coefficient

Items	A	B	C	D	E
Predicted Ratings	4	4	5	3	2
True Ratings	4	5	4	3	1

Items	A	B	C	D	E
Predicted Ranking	2.5	2.5	1	4	5
True Ranking	2.5	1	2.5	4	5

1. Use average ranking to break the ties
2. Calculate the Pearson Correlation Coefficient of (2.5, 2.5, 1, 4, 5) and (2.5, 1, 2.5, 4, 5)

通過ROC評估排名



Source: dataminingapps.com

覆蓋度

- 在實際情況下，推薦系統可能無法推薦某個部分的商品，或者可能無法向某個比例的用戶推薦。（為什麼？）
- 可以透過預置值(default values)替代不可能預測的評級，以犧牲準確性為代價來增加覆蓋率。
- 用戶空間的覆蓋:
  - 預測至少k個商品評分的用戶比率
  - 用戶空間覆蓋應評估準確性和覆蓋範圍之間的權衡。
- 商品空間的覆蓋:
  - 預測至少k個用戶評分的商品比率
  - 這個指標很少使用，因為推薦系統通常向用戶提供商品推薦，很少用於為商品推薦可能會感興趣的用戶。
  - 目錄覆蓋率：推薦給至少一個用戶的商品比率。（為什麼這是一個更適用的指標？）

置信度(Confidence)和可信度(Trust)

- 置信度是衡量系統對所提供推薦的信心。 許多推薦系統可能會連同置信度的估計和評分一起報告。
- 可信度衡量用戶對推薦的信心。 較小的推薦信賴區間(confidence Interval)有助於加強用戶對系統的信任。
- 訓練數據和演算方法對預測評分有顯著的影響，並導致用戶對預測準確性的不確定性。
- 信任與準確性密切相關，但不完全相同。 當推薦系統提供解釋時，用戶更可能相信系統，特別是如果解釋合乎邏輯。
- 衡量信任的最簡單方法是進行線上用戶調查。 通過離線實驗很難衡量信任。

多樣性 (Diversity)

- 多樣性是衡量一個推薦列表中推薦的多樣性。 例如，如果推薦3個具有相同類別和類似演員的電影的列表，則幾乎沒有多樣性。
- 增加多樣性常常會增加推薦的新穎性和驚喜性。 此外，還可以增加銷售的多樣性和系統的目錄覆蓋面。
- 可以根據一堆商品之間，以內容為中心的相似性來衡量多樣性。 每個商品描述的向量代表可以用於相似性的計算。 例如，如果向用戶推荐一組k個商品，則可以在列表中計算每對商品之間的相似度，並且可以將所有商品之間的平均相似度報告為多樣性。
- 為增加多樣性, 通常可能導致不利於準確性指標的結果。

驚喜性

- 驚喜性是衡量成功推薦的驚喜程度。 這是一個比新奇更強的指標。 所有驚喜的推薦都是新奇的，但反之則不然。
- 例如：推薦一家巴基斯坦餐館給印度美食愛好者可能是新奇的（如果用戶不知道這家餐館），而向這些用戶推薦伊索比亞餐館可能是驚喜的，因為它不是顯而易見的。
- 測量驚喜性的方法：
  - **線上方法：** 蒐集推薦商品的有用性和顯而易見性的回饋。有用和不顯而易見推薦的比率是一個驚喜性的指標。(用戶調查)
  - **離線方法：** 比較這個推薦和來自低多樣性的推薦（如基於內容的系統）之推薦內容。top-k列表中推薦商品的部分是正確的，也不被低多樣性系統推薦，這是衡量驚喜性的一個指標。
- 驚喜性對於提高推薦系統的轉換率具有重要的長期影響，雖然它與提高準確性的目標相反。

新奇性

- 新奇性是一個衡量推薦系統向用戶提供他們不知道或者以前從未見過商品的指標。 這比提供他們已經知道但尚未評分的商品更重要。
- 推薦用戶沒見過的商品往往會增加他們發現以前所不知道之喜歡而且非常不一樣的商品。
- 衡量新奇性的最自然的方法是透過線上實驗，其中用戶被明確詢問他們以前是否知道這個商品。
- 使用離線方法,有些假設受歡迎的商品新奇性較低。有些假設推薦將來更有可能選擇的商品新奇性較高，利用時間戳(time stamp)，估計新奇性。但他們並不完全真實反映系統帶給顧客的新奇性。

穩健性

- 一個推薦系統是穩健的，如果遭受攻擊(例如:假評分)或當數據的模式隨著時間顯著演變，然而推薦的準確性並沒有被顯著地影響。

可擴展性

- 近年來，蒐集大量的評分和隱式反饋訊息變得越來越容易，數據集的規模也在不斷增加， 因此，設計可擴展的推薦系統已變得越來越重要。
- 使用各種措施來確定系統的可擴展性：
  - 訓練時間：在大多數情況下，訓練是離線完成的，如果訓練時間在幾個小時內，則是可以接受的。
  - 預測時間：預測時間低是至關重要的。
  - 內存要求：當評分矩陣較大時，設計算法以最大限度地減少內存需求至關重要。
- 由於“大數據” 的重要性日益增強，近年來可擴展性的重要性已經變得尤為突出。