Offline Metrics

Fernando Diaz, Google

what are offline metrics?

system evaluation that replaces real user behavior with some proxy for that behavior.

proxies

- user models
- annotations
- log data

why use offline metrics

- faster than running a real experiment
- repeatable
- avoids exposing users to ineffective systems

example: search and recommendation

task: given an a query/context and a repository of items, find relevant information

example: search and recommendation

task: given an a **query/context** and a repository of items, find relevant information

example: search and recommendation

task: given an a query/context and a **repository of items**, find relevant information

example: search and recommendation

task: given an a query/context and a repository of items, find **relevant information**

instances

$$\mathcal{D}_q = \{\langle x, y \rangle_1, \ldots, \langle x, y \rangle_n\}$$

x query-document features
(e.g. term matches,
item popularity, context)

y relevance

all metric definition for a single query or context

metric: accuracy

$$\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$$

f	document scoring function
y_i	relevance label for document i
x_i	query-document features for document i

example: search and recommendation

task: given an a query/context and a repository of items, find relevant information

questions

- *how is information presented?*
- *how do users consume the information?*

metric: precision

$$\mathrm{Prec}(\mathcal{Y}^+, \mathcal{F}_\theta^+(X)) = \frac{|\mathcal{Y}^+ \cap \mathcal{F}_\theta^+(X)|}{|\mathcal{F}_\theta^+(X)|}$$

$\mathcal{Y}^+$ relevant document set

$\mathcal{F}_\theta^+(X)$ predicted relevant document set

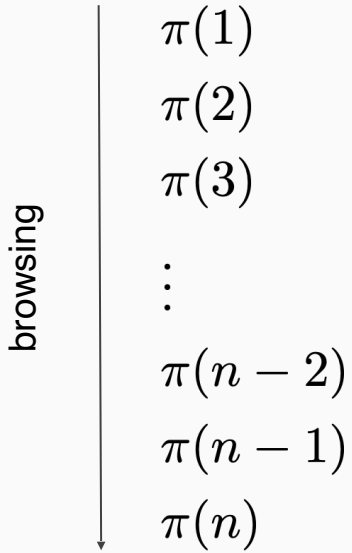
Cyril W. Cleverdon, Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems, 1962

metric: recall

$$\mathrm{Rec}(\mathcal{Y}^+, \mathcal{F}_\theta^+(X)) = \frac{|\mathcal{Y}^+ \cap \mathcal{F}_\theta^+(X)|}{|\mathcal{Y}^+|}$$

Cyril W. Cleverdon, Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems, 1962

ranking



metric: expected search length

user model: in-order traversal of a ranked list, collecting up to *k* items.
metric: number of nonrelevant documents skipped before reaching *k* relevant items.

$$\mathrm{ESL}(\mathcal{Y}^+, \pi, k) = \text{min-}k_{i \in \mathcal{Y}^+} \overline{\pi}(i)$$

$$\begin{aligned} \text{min-}k & \quad k\text{th smallest value} \\ \overline{\pi}(i) & \quad \text{rank position of item } i \end{aligned}$$

William S. Cooper. Expected search length,. American Documentation, 19(1):30–41, 1968.

metric: R-precision

user model: in-order traversal of a ranked list, collecting all relevant items.
metric: precision when recall is 1.

$$\mathrm{RPrec}(\mathcal{Y}^+, \pi) = \mathrm{Prec}(\mathcal{Y}^+, \pi_{1:k^*})$$

$$k^* = \max_{i \in \mathcal{Y}^+} \overline{\pi}(i)$$

metric: R-precision

user model: in-order traversal of a ranked list, collecting all relevant items.
metric: precision when recall is 1.

$$\begin{aligned}\mathrm{RPrec}(\mathcal{Y}^+, \pi) &= \mathrm{Prec}(\mathcal{Y}^+, \pi_{1:k^*}) \\ &= \frac{|\mathcal{Y}^+|}{\mathrm{ESL}(\mathcal{Y}^+, \pi, |\mathcal{Y}^+|)}\end{aligned}$$

metric: reciprocal rank

user model: in-order traversal of a ranked list, satisfied by one item.
metric: inverse of the number of documents skipped before reaching the relevant item.

$$\mathrm{RR}(\mathcal{Y}^+, \pi) = \max_{i \in \mathcal{Y}^+} \frac{1}{\pi(i)}$$

E. Voorhees and D. Tice. The trec-8 question answering track evaluation. TREC, 1999.

metric: reciprocal rank

user model: in-order traversal of a ranked list, satisfied by one item.
metric: inverse of the number of documents skipped before reaching the relevant item.

$$\begin{aligned}\mathrm{RR}(\mathcal{Y}^+, \pi) &= \max_{i \in \mathcal{Y}^+} \frac{1}{\overline{\pi}(i)} \\ &= \frac{1}{\mathrm{ESL}(\mathcal{Y}^+, \pi, 1)}\end{aligned}$$

E. Voorhees and D. Tice. The trec-8 question answering track evaluation. TREC, 1999.

metric: average precision

user model: in-order traversal of a ranked list, collecting all relevant items.
metric: precision averaged over all recall levels.

$$\mathrm{AP}(\mathcal{Y}^+, \pi) = \frac{1}{|\mathcal{Y}^+|} \sum_{i \in \mathcal{Y}^+} \mathrm{Prec}(\mathcal{Y}^+, \pi_{1:\overline{\pi}(i)})$$

metric: average precision

user model: in-order traversal of a ranked list, collecting all relevant items.
metric: precision averaged over all recall levels.

$$\mathrm{AP}(\mathcal{Y}^+, \pi) = \frac{1}{|\mathcal{Y}^+|} \sum_{i \in \mathcal{Y}^+} \mathrm{Prec}(\mathcal{Y}^+, \pi_{1:\overline{\pi}(i)})$$

$$= \frac{1}{|\mathcal{Y}^+|} \sum_{r=1}^{|\mathcal{Y}^+|} \frac{r}{\mathrm{ESL}(\mathcal{Y}^+, \pi, r)}$$

metric: rank-biased precision

user model: in-order traversal of a ranked list; utility independent of stopping probability.
metric: expected utility given simulated user behavior.

$$\mathrm{RBP}(y, \pi) = (1 - \gamma) \sum_{r=1}^{n} y_{\pi(r)} \gamma^{r-1}$$

γ patience parameter

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. TOIS, 2002.
Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. TOIS, 2008.

metric: expected utility

user model: in-order traversal of a ranked list, gains utility of 1 for each relevant item.
metric: expected utility given simulated user behavior.

$$\mathrm{EU}(y, \pi) = \sum_{r=1}^{n} y_{\pi(r)} \phi(y_{\pi(r)}) \underbrace{\gamma^{r-1} \prod_{r'=1}^{r-1} (1 - \phi(y_{\pi(r')}))}_{\text{examination probability}}$$

$\phi(i)$ probability that user stops given relevance of item i

Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. SIGIR, 2009.

metric: time-based gain

user model: in-order traversal of a ranked list; utility independent of stopping probability; utility based on time to reach the position.

metric: expected utility given simulated user behavior.

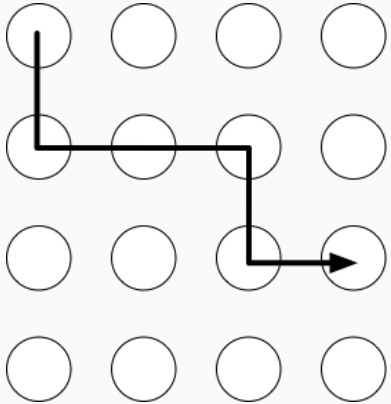
$$\mathrm{TBG}(y, \pi) = \sum_{r=1}^{n} y_{\pi(r)} \Delta(y, \pi, r)$$

$\Delta(y, \pi, r)$ discount based on time to reach rank r

Mark D. Smucker and Charles L.A. Clarke. Time-based calibration of effectiveness measures. SIGIR, 2012.

grid metrics

user model: model of ordered user traversal of grid items; utility independent of stopping probability.
metric: expected utility given simulated user behavior.



X Xie, J Mao, Y Liu, M de Rijke, Y Shao, Z Ye, M Zhang, S Ma. Grid-based evaluation metrics for web image search. WWW 2019.

aggregating metrics

- traffic-weighted
- unique query/context
- usage-segmented
- group-segmented

Rocío Cañamares, Pablo Castells, and Alistair Moffat, Offline evaluation options for recommender systems, Information Retrieval Journal, 2020

further reading

IR evaluation: Modeling user behavior for measuring effectiveness
Charles L.A. Clarke, Mark D. Smucker, and Emine Yilmaz
SIGIR, 2015

Ian Soboroff. Building Test Collections, SIGIR 2017.

Offline evaluation options for recommender systems
Rocío Cañamares, Pablo Castells, and Alistair Moffat
Information Retrieval Journal, 2020