VALUES TO METRICS TOOLKIT

INTRODUCTION

The goal of this exercise is to build evaluation metrics for recommender systems that go beyond how good the system is at predicting a click. To accomplish this, we will first discuss at an abstract level what a recommender system needs to do. Then, we will translate this step-by-step into concrete evaluation metrics.

STEP 1

DEFINING OBJECTIVES (20 MIN)

Align with your discussion partners on the recommender system you need to evaluate. Discuss the following points:

- What does the recommender system do?
- Why does it do this?
- Who does it do it for?

We call the answers to these questions the objectives of your recommender systems. Write the objectives down on post-it notes, as we will refer to them later in the process. Identify as many objectives as you can.

STEP 2

VOTE (10 MIN)

Which objectives are most important? Each group member gets 5 votes; distribute your votes among the objectives you find most important. One person can assign multiple votes to an objective.

STEP 3

BUILDING A METRIC (20 MIN)

The goal of this exercise is to build evaluation metrics that reflect the objectives you defined in the previous step. What would you need to evaluate or optimize in order to determine whether the recommender system behaves as you want?

A metric consists of three parts:

1. *What* needs to be measured.
2. *Where* you want to measure.
3. The *value* you expect the measurement to take.

Each of the parts above is represented by a set of cards. Go through the objectives, and combine cards in a way that they reflect what you wrote down. Use the provided **Metric Builder** in your discussions. When you are satisfied with a metric, give it a name and write it down.

WHAT DO YOU WANT TO MEASURE?

This card expresses what aspect of the recommendation you want to measure, such as article topic, a song's artist, or a book's popularity. It must be something you can count or otherwise express as a number. Sometimes an item has just one score for this part (for example, how complex it is on a scale from 0 to 100). Other times, an item can have several values (for example, it can cover several topics, or mention people from different backgrounds).

There are three types of aspects: **Item**, **Human**, and **World**.

Item aspects are about the things being recommended themselves. For example, publication date, topic or format.

World aspects are about what is happening outside the system. For example, current events or how society is organized.

Human aspects are about the people involved. For example, the users, the people who created the items, or the people who appear in or are described by the items.

When you use a Human aspect, combine an **aspect card** (yellow and white) with a matching **role card** (white ande yellow).

WHERE DO YOU WANT TO MEASURE?

Where are you looking for the aspect you want to measure? Does it only matter what the user is seeing in the current recommendation, or do the recommendations they have seen in the past also factor in? Or are you, alternatively, only monitoring trends for all users at the aggregated level? For example:

- Making sure that over time a user sees articles from all different political parties.
- Monitoring for increases in clicks among all the users of the system.

Other times, you want to compare the recommendation to something outside the recommendation. For example:

- Comparing the user's past activity (their history) to what they are now being recommended.
- Comparing the recommendation to what other users are being recommended, to ensure that there is still overlap between them.

Also discuss whether items that are positioned higher in the recommendations should be counted more strongly than those at the bottom.

WHAT VALUE DO YOU EXPECT THE MEASUREMENT TO TAKE?

This card is about deciding what a 'good' value looks like for your measurement. When is the recommendation doing well? When is it doing poorly?

There are two types of value cards: **relational** and **internal**. **Relational values** compare the recommendation to something else, e.g. it must be *different* from what a user has seen before or *similar* to the day's mix of topics published by the organization.

Internal values are measured within the recommendation itself, e.g. it must contain 20% sports news or include as many different topics as possible.

EXAMPLE OF A FULL METRIC

Measure whether a recommendation reflects the full breadth of content produced by the organization.

- What to measure: "category/genre/topic".
- Where to measure: Every single recommendation needs to consist of a good mix, so *One single recommendation.*
- What value: the recommendation needs to reflect what the organization has produced, so *Similar to* combined with the *Organization* context.

STEP 4

NOW, HOW, WOW (10 MIN)

Use the **Now/How/Wow matrix** to sort ideas by how easy they are and how potentially impactful they are.

NOW

Easy and clear to implement, data is already available, could be done relatively quickly

HOW

Good ideas but not ready yet; need more research, tools, money, or skills.

WOW

Great ideas with a lot of value that should be picked up as soon as possible

LATER/DREAM

If something feels almost impossible with today's resources, park it in a 'Later/Dream' corner.

STEP 5

WHO (10 MIN)

Look at the **Now/How/Wow matrix** you created. Who is responsible for taking the next steps towards implementation? Write down the name of the objective, the name of the person or team responsible, and the specific action they will take.

STEP 6

CLOSE (15 MIN)

Finally, reflect on the process as a whole. What went well, and what would you do differently if you were to start again? Has everyone's perspective been sufficiently heard? Were any perspectives structurally missed? When will the actions identified in the previous step be completed?

ADDITIONAL INSIGHTS

During the workshop, topics may arise that fall outside the immediate scope of the session but may still be relevant or valuable. These can be collected in an action list. At the end of the workshop, this list can be reviewed together and followed up where needed.

ACKNOWLEDGEMENTS

This toolkit is sponsored by the Impact Fund at University of Amsterdam and part of the AI, Media & Democracy ELSA Lab (Dutch Research Council project number: NWA.1332.20.009). For more information about the lab and its further activities, visit **aim4dem.nl**

BUILDING A METRIC

BUILDING A METRIC

Build evaluation metrics that reflect the objectives you defined in the previous step. What would you need to evaluate or optimize in order to determine whether the recommender system behaves as you want?

A METRIC CONSISTS OF THREE PARTS:

- *What* needs to be measured.
- *Where* remove you want to measure.
- The *value* you expect the measurement to take.

NAME THE METRIC

When you are satisfied with a metric, give the metric a name and write it down.

VALUES TO METRICS TOOLKIT

AI, MEDIA & DEMOCRACY LAB
UNIVERSITEIT VAN AMSTERDAM

VALUES TO METRICS

A card-based workshop tool for news organizations to reflect on and improve their recommendation systems. The steps guide participants from defining objectives to designing metrics and identifying next steps.

CONCEPT & RESEARCH

Developed within the AI, Media & Democracy Lab. Idea and primary research by **Sanne Vrijenhoek** and **Sara Spaargaren**.

CONTRIBUTORS

Kornelija Gruodyte, Lien Michiels, Savvina Daniil, Pascal Wiggers, Maaike Harbers, Natali Helberger.

VISUAL CONCEPT & GRAPHIC DESIGN

Daphne de Vries / Bureau Merkwaardig

This toolkit is part of the AI, Media & Democracy Lab. For more information about the lab and its further activities, visit **aim4dem.nl**