Precision vs. Recall

When Good Enough Beats Perfect

Arne Vogt
Lead PM, Search & Discovery - OTTO

MICES 2026

Who we are

OTTO is a full-range online marketplace with fashion, electronics, furniture, toys and much more

CATALOG SIZE

>18M

ACTIVE CUSTOMERS

~13M

DAILY SEARCH QUERIES

~2M

Arne Vogt
Lead PM, Search & Discovery: responsible for what happens between a query and what you find.

OTTO

We rolled out hybrid search. The numbers said yes, users said no.

Rolled out in stages. Zero-hit fallback first, then progressively wider, then everywhere.

CONVERSION RATE

>5%

Online A/B uplift

Cumulative across multiple hybrid search experiments

> *"Hello Otto, the search for products is not working. Different products are being shown than what was searched for. This causes the purchase to be abandoned. A shame."*

> *"Why am I seeing women's products in a men's search?"*

> *"The search could be better. The search should be more precise. There are always many articles coming up that I definitely did not search for."*

We tried several precision interventions, none moved the needle

Dynamic similarity thresholds

Tighten the semantic similarity cutoff for queries where the lexical signal is strong.

Result after several iterations: no significant effect

Query-intent-based precision

Boost precision when the query looks specific — assuming precise queries reflect decided users.

Result: no significant effect

Two angles, two null results. Are we measuring the right thing or measuring it badly?

OTTO

So we ran the cleanest possible experiment we could imagine

Filtering out opposite-gender products for queries with explicit gender terms (men/women)

✓ Clean query segmentation ✓ Clean product taxonomy

Four signals. Four times zero.

Zero impact on both economic and efficiency metrics

CONVERSION RATE	FILTER USAGE	SCROLL DEPTH	REFORMULATION RATE
≈ 0*	**≈ 0**	**≈ 0**	**≈ 0**
HYPOTHESIS	HYPOTHESIS	HYPOTHESIS	HYPOTHESIS
Bad precision → frustration → drop-off.	*Bad precision → users filter the noise out.*	*Bad precision → users scroll past to find it.*	*Bad precision → users re-type. Soft drop-off.*
**negative tendencies*			

That's when we started to doubt what we were measuring.

OTTO

We can prove the chain for Recall. We can’t for Precision.

	OUTPUT		OUTCOME	
	things you build		*things you achieve*	
	System change	**Result change**	**Behavior change**	**Business impact**
Recall	✓ Added semantic retrieval	✓ More products surfaced	✓ More purchases	✓ More revenue (€)
Precision	✓ Thresholds, intent, filters	✓ Tighter result sets	✗ Couldn’t prove	✗ Couldn’t prove

Walmart (Rossi et al., CIKM 2024) report a +5% precision lift on hybrid retrieval and zero movement in Orders or GMV

OTTO

Users click and buy ‘irrelevant’ results: in gender searches and across other categories

What users actually click and buy under the query “macbook pro” on otto.de.

26% clicks / 20% orders

17% clicks / 9% orders

10% clicks / 8% orders

Irrelevant?

7% clicks / 6% orders

Irrelevant?

This is what we landed on: ESCI

If S and C are valid results, then “doesn’t match query” ≠ “irrelevant”.

E

Exact

Product is what the user searched for.

S

Substitute

Different product, same job to be done.

C

Complement

Goes with what the user searched for.

I

Irrelevant

No meaningful relationship to the query.

Amazon Shopping Queries Dataset (Reddy et al., 2022) - four classes of query-product relationships

OTTO

Two ways of defining relevance, they don’t always agree

	implicit: relevant	implicit: irrelevant
explicit: relevant	Matches the query and is purchased.	Matches the query but unattractive (price, quality, shipping).
explicit: irrelevant	Doesn’t match the query, but users click and buy it.	High probability of true irrelevance.

Both the Gender case and the MacBook Air case live bottom-left: explicitly irrelevant, implicitly relevant.

OTTO

Query ≠ User Intent

Same query. Three users. Three different points in the decision. The query doesn't always match the intent.

Query: "MacBook Pro"

D E C I D E D

"I've chosen Pro. Just show me configurations."

Wants Exact matches. Anything else is noise.

C O M P A R I N G

"Pro vs Air vs Dell — show me the field."

Wants Substitutes side-by-side to decide.

E X P L O R I N G

"Need a new notebook. 'Pro' was the first thing I thought of."

Wants the category, not the query string.

The complaints come from the Decided users.

DECIDED

Sees Substitutes they didn't ask for.

Writes a ticket.

We hear them. Loudly.

COMPARING

Sees the field. Decides faster.

Says nothing — they're happy.

We don't hear them.

EXPLORING

Discovers options they wouldn't have typed.

Says nothing — they're happy.

We don't hear them.

Conversion says the system wins. Tickets say it loses. Both are right, but for different users.

OTTO

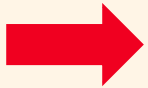
We didn’t build an imprecise retrieval system. Hybrid Search enabled discovery use cases.

before

Retrieval

“Did the result match the query?”

Optimization target: query–product fit.



after

Discovery

“Did the user find what they needed?”

Optimization target: user–need fit.

OTTO

Four open questions we're still working through

Q1 · DETECTION

Which queries are Discovery-heavy and how do we identify them in real time?

We tried reading intent from the query string. The query doesn't tell us. So where does intent live?

Q2 · COMMUNICATION

How do we communicate the Discovery mix so it doesn't feel like noise?

Can we reduce the user irritation and keep the benefits?

Q3 · Metrics

Which metrics can we use to measure the impact of precision optimization?

Economic metrics didn't work for us. Efficiency metrics didn't work for us either. What else is there?

Q4 · RISK

Is the short-term conversion win masking long-term trust we can't see?

A/B tests run for weeks; trust erodes over months. We have not falsified this risk.

Come discuss with us this afternoon.

Barcamp session - bring your hypothesis, your data, your experiences.

Arne Vogt
Lead PM, Search & Discovery - OTTO

linkedin.com/in/arne-vogt-16201387