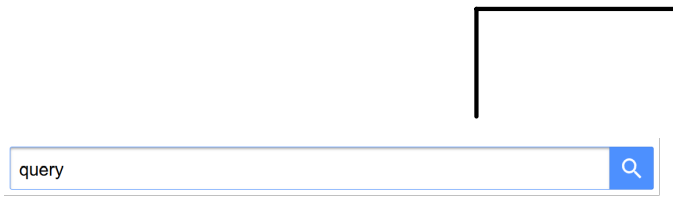
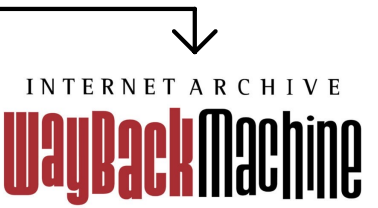
Verifying Query Logs from Unknown Sources

29.04.2025

Master's Thesis

Benjamin Schneg





Query ID	Query
1	google
2	weather
3	youtube
…	…
n	sports

?

Query ID	Query
1	video
2	finance
3	place:1028
…	…
m	query

- Perfect for research
- Suitable?

Archive Query Log

The AQL corpus

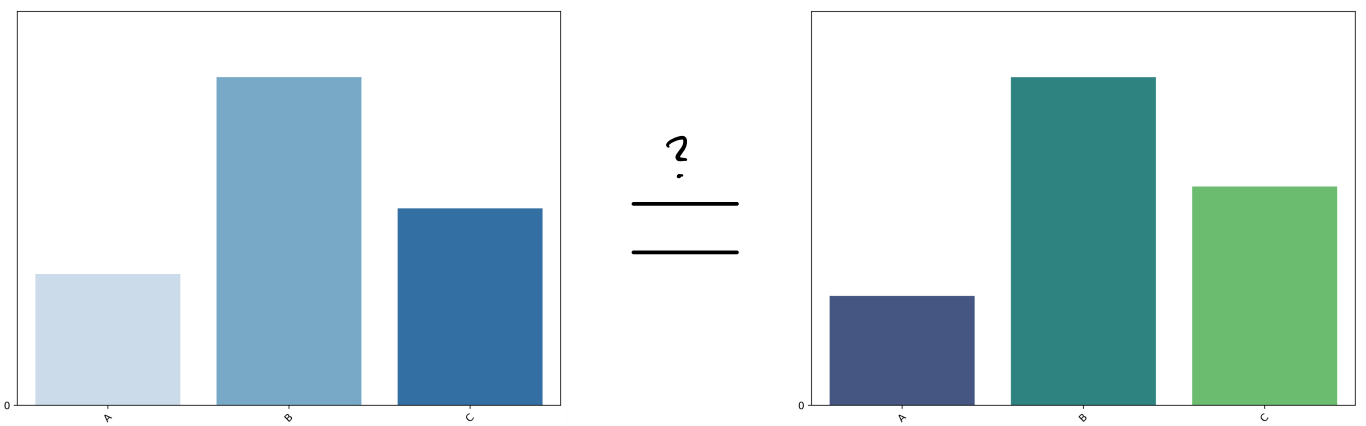
Search provider	URLs	Queries	Unique	SERPs	Results
Google	89.4 M	72.7 M	20.0 M	28.0 M	223.1 M
YouTube	41.8 M	41.4 M	11.3 M	15.9 M	339.2 M
Baidu	78.5 M	69.6 M	2.9 M	26.8 M	107.6 M
QQ	0.5 M	0.5 M	0.1 M	0.2 M	2.1 M
Facebook	3.1 M	0.2 M	0.0 M	0.1 M	0.7 M
Yahoo!	8.8 M	2.8 M	1.2 M	1.1 M	9.2 M
Amazon	66.8 M	0.8 M	0.3 M	0.3 M	7.8 M
Wikipedia	68.5 M	1.7 M	0.6 M	0.7 M	7.0 M
JD.com	4.4 M	3.9 M	0.4 M	1.5 M	16.0 M
360	1.5 M	1.1 M	0.1 M	0.4 M	3.5 M
⋮ 540 others	646.8 M	161.8 M	27.8 M	62.4 M	693.9 M
Σ **550**	**1010.2 M**	**356.5 M**	**64.5 M**	**137.3 M**	**1410.0 M**

- first large log of SERPs
- mined from the Internet Archive's Wayback Machine
- from 550 search engines across 25 years

Query log use cases

Retrieval models and user behaviour

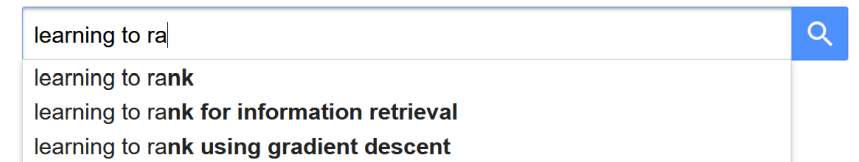
- Need for queries that encode realistic user behaviour



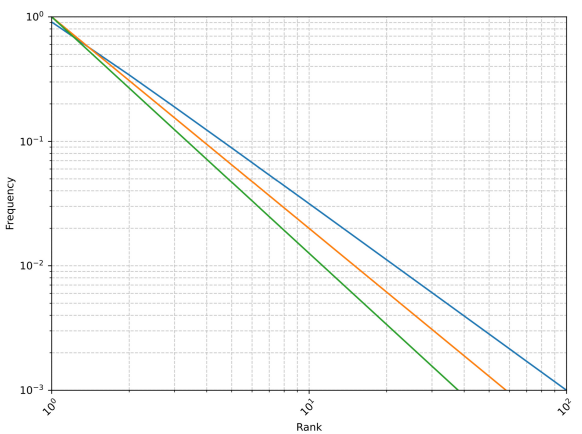
- Query intent
- Question-fraction
- Named entities

Query log use cases

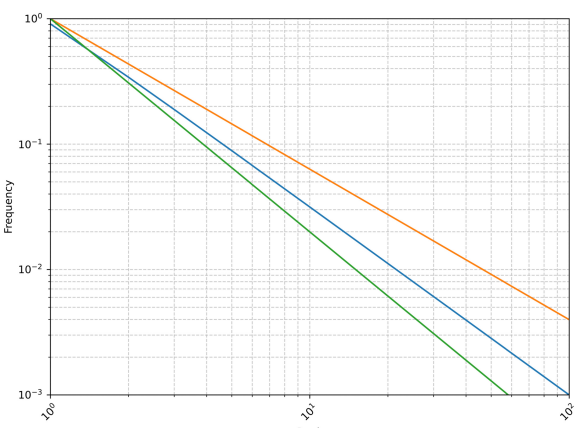
Query suggestions and reformulations



- Need for similar query structures
- Need for similar query distributions



?



- Query distribution
- Word distribution

- Query length
- Word length

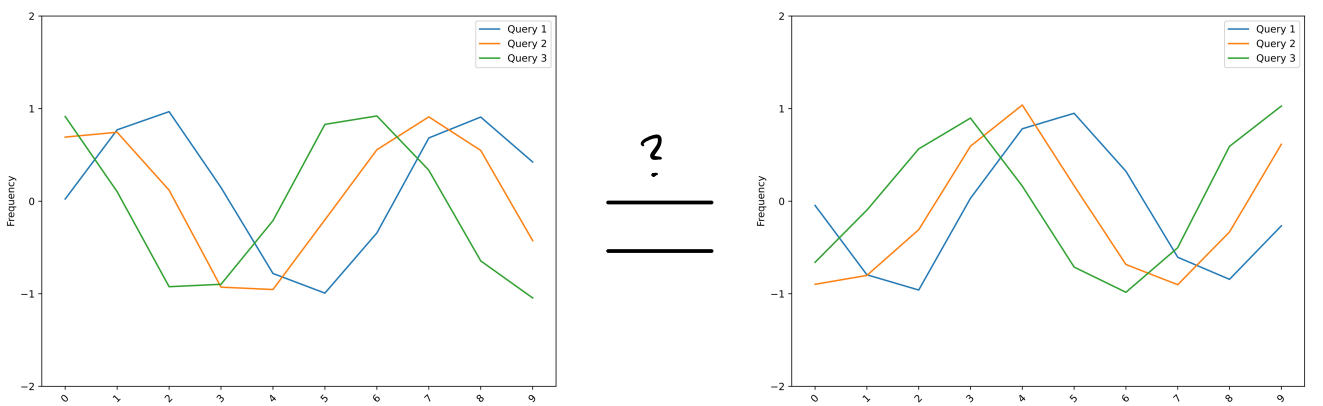
Query log use cases

Diachronic analyses

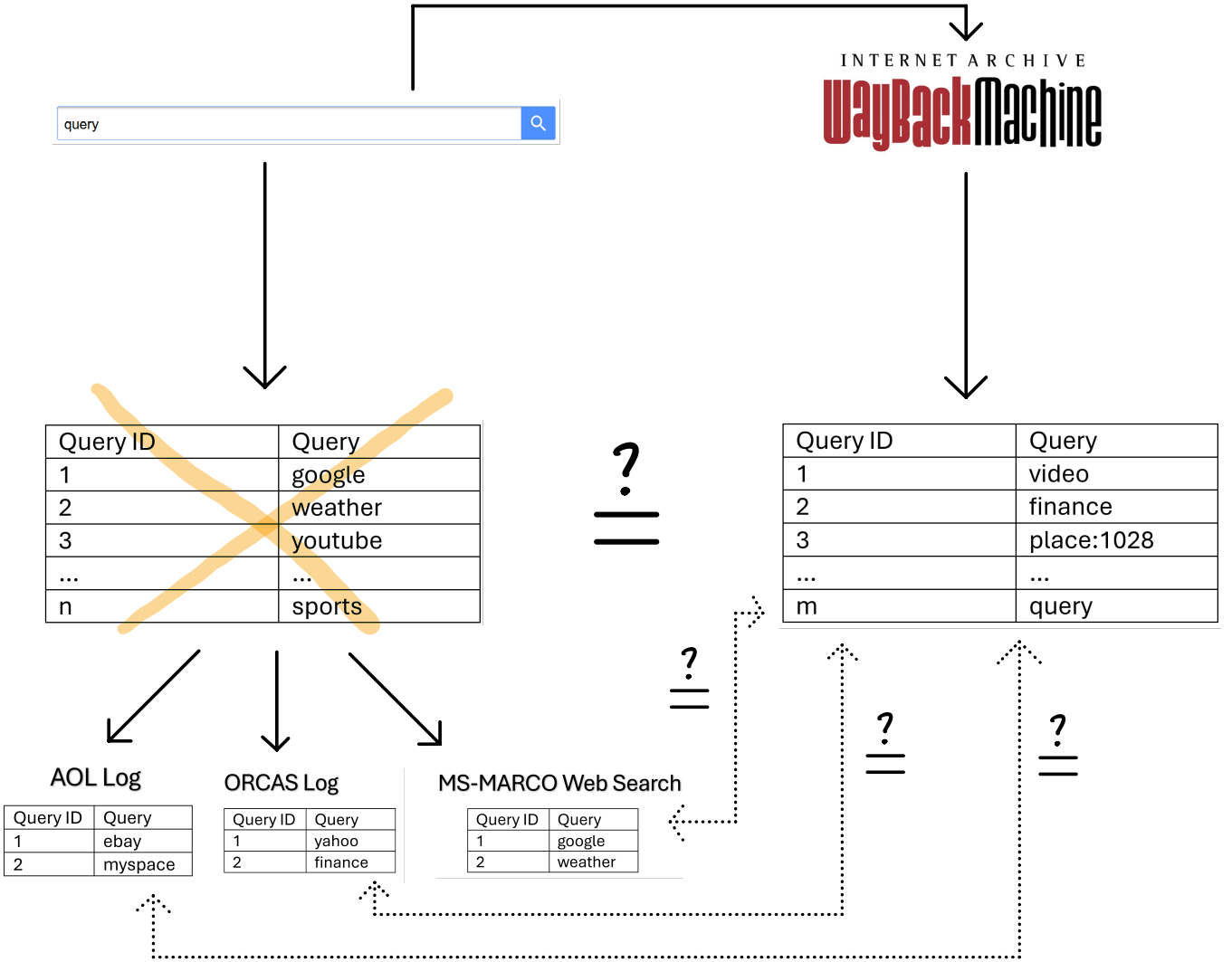
Can we derive search trends?

Can we track the change of search engine behaviour over time?

- Need of realistic temporal patterns
- But: sometimes queries from different time periods are sufficient



- Correlation of query popularities

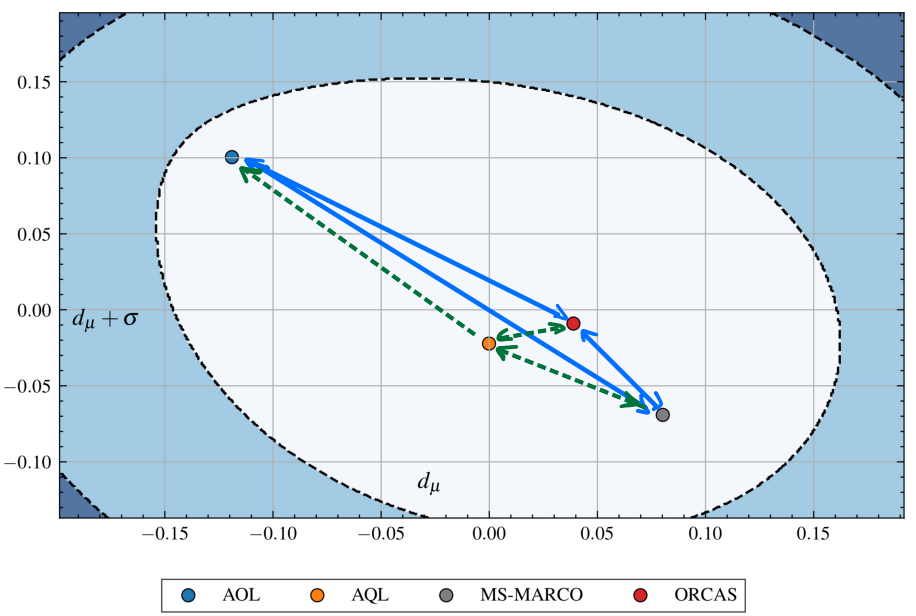

query
INTERNET ARCHIVE
WayBackMachine
Query ID
Query
1
google
2
weather
3
youtube
...
...
n
sports
?
Query ID
Query
1
video
2
finance
3
place:1028
...
...
m
query
?
?
?
AOL Log
Query ID
Query
1
ebay
2
myspace
ORCAS Log
Query ID
Query
1
yahoo
2
finance
MS-MARCO Web Search
Query ID
Query
1
google
2
weather

Comparison framework

Distances between distributions

Consider two cases:

1. Average distance within comparison group
2. Average distance between AQL and comparison group



d(AQL) = 8.1

d(CG) = 11.3

Comparison framework

Choose a suitable distance measure

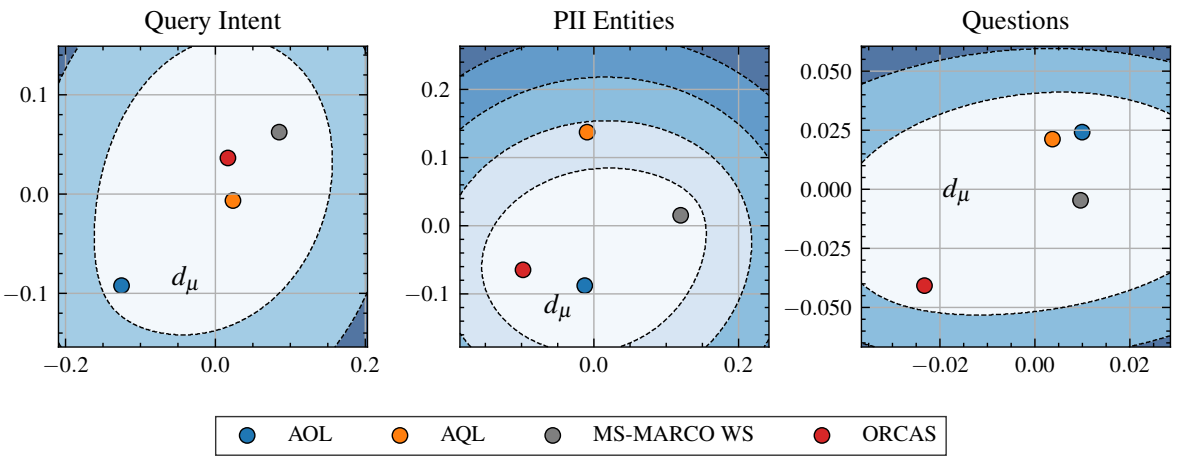
- Prerequisite: distance should be a metric
  - $d(a, b) \overset{!}{=} 0$ if and only if $a = b$
  - $d(a, b) \overset{!}{\geq} 0$
  - $d(a, b) \overset{!}{\leq} d(a, c) + d(c, b)$
  - $d(a, b) \overset{!}{=} d(b, a)$

- Kullback-Leibler divergence: does not satisfy symmetry
- Wasserstein distance: satisfies all properties
  - Optimal transport between two distributions

Evaluation

Results inference-based analyses

- Query intent, question fraction, PII entity distribution

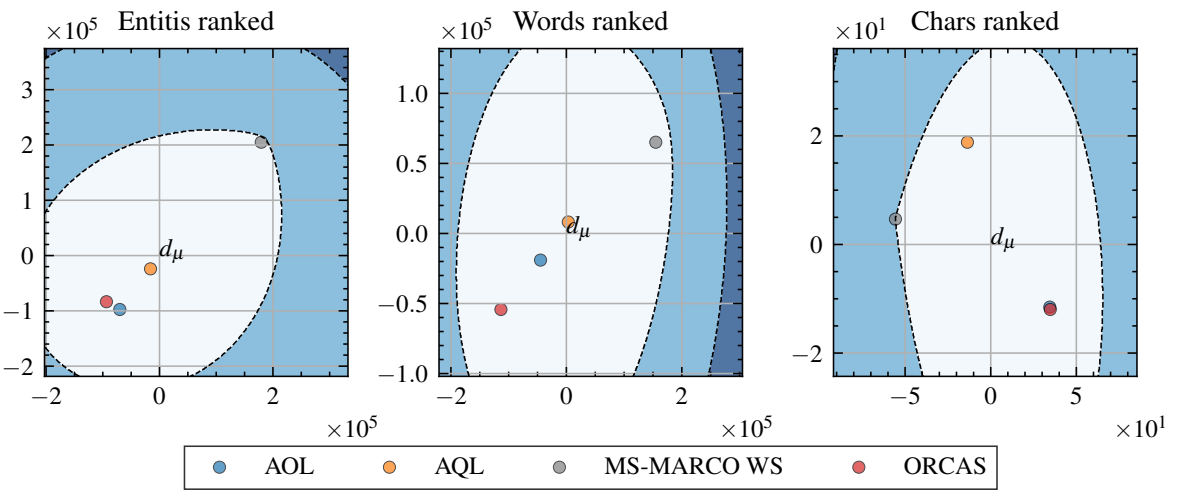


- Support for user behaviour analyses

Evaluation

Results structural similarity

- Rank-size distributions of linguistic features

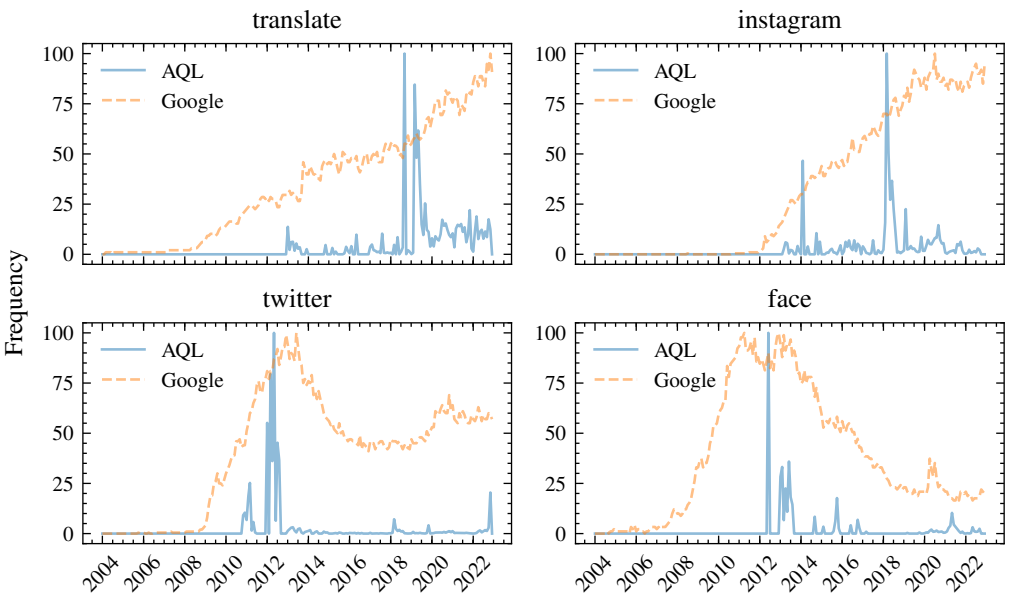


- Support for query suggestions and reformulations

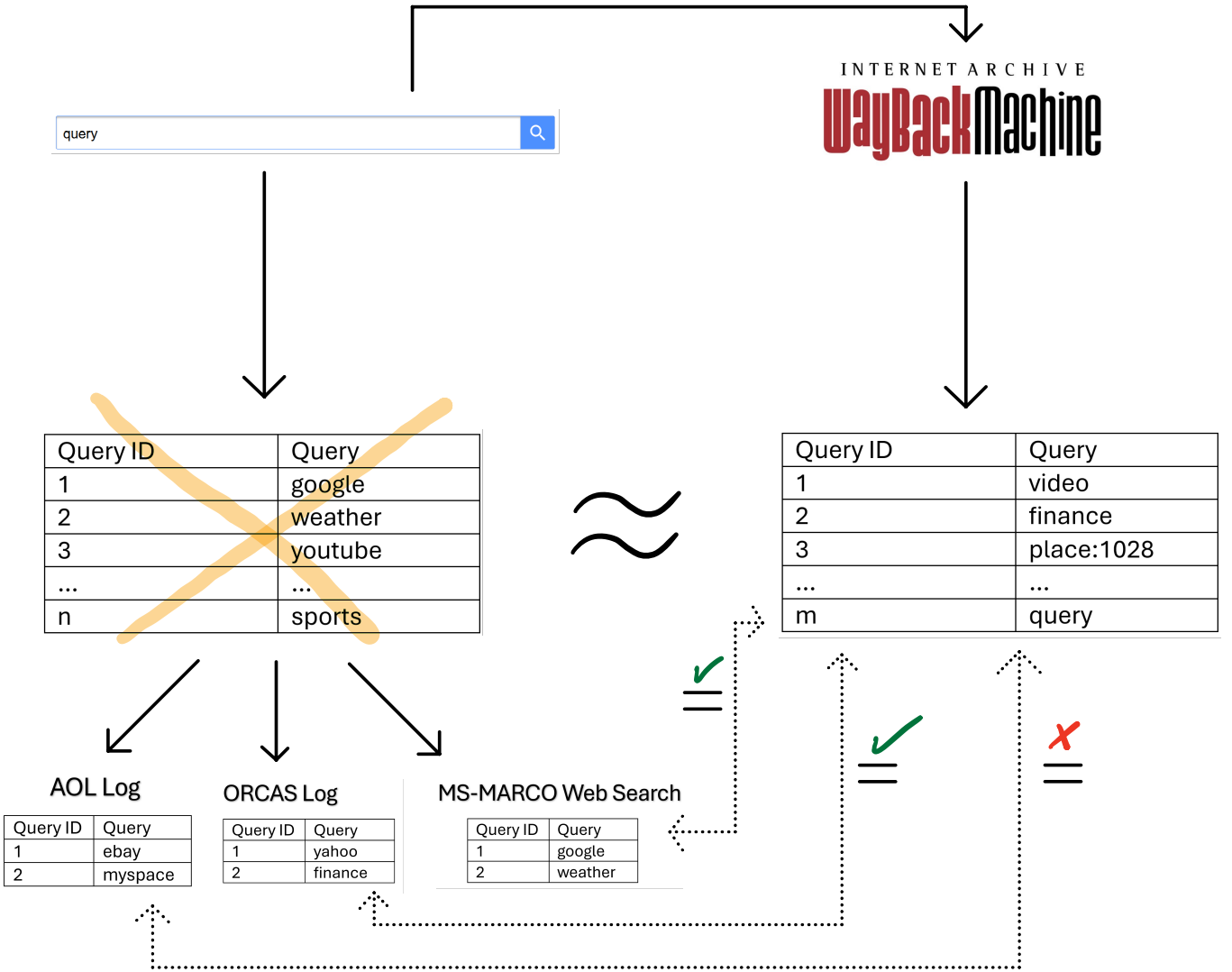
Evaluation

Results temporal patterns

- Correlation of query popularities



- Trending information needs not detectable
- But: diachronic analysis of search engine behaviour possible



Thank you for your attention!