FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

User-Driven Re-Ranking for Adapting the Variety in Search Results

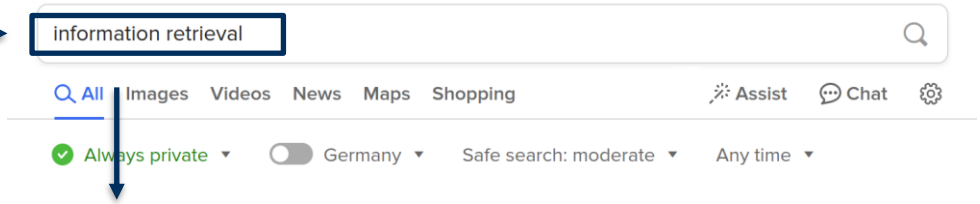
Daphne Auer, Divyasha Sunil Naik, Birgitta König-Ries

Open Search Symposium 2024, 11.10.2024

Information Need

What is information retrieval about?

Query



Document Ranking

https://en.wikipedia.org › wiki › Information_retrieval

Information retrieval - Wikipedia

Information retrieval (IR) in computing and **information** science is the task of identifying and retrieving **information** system resources that are relevant to an **information** need.The **information** need can be specified in the form of a search query. In the case of document **retrieval**, queries can be based on ful...

https://www.geeksforgeeks.org › what-is-information-retrieval

What is Information Retrieval? - GeeksforGeeks

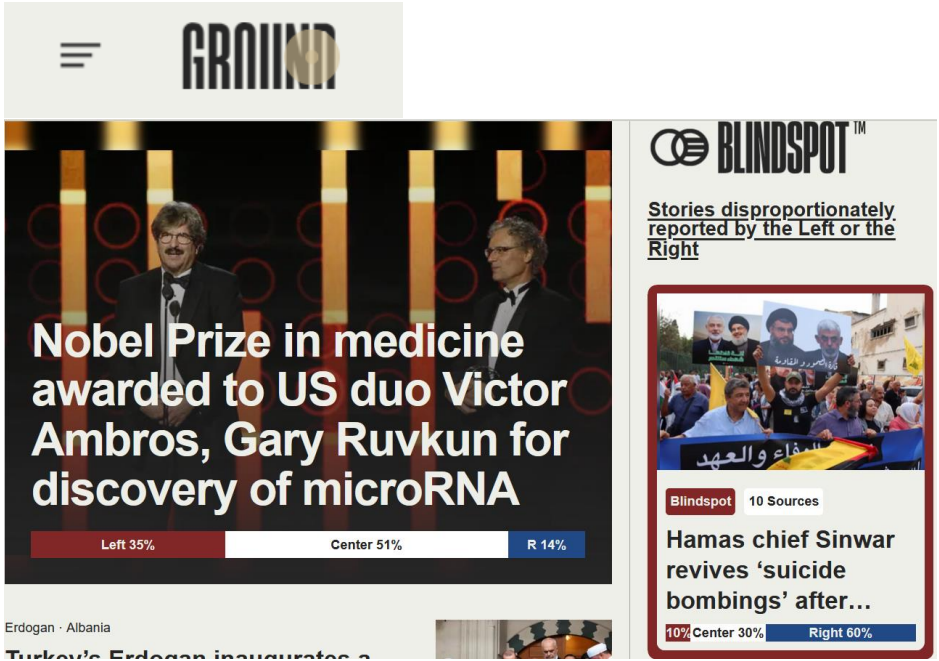
Sep 19, 2023 · **Information Retrieval** (IR) can be defined as a software program that deals with the organization, storage, **retrieval**, and evaluation of **information** from document repositories, particularly textual **information**. **Information Retrieval** is the activity of obtaining material that can usually be...

https://www.britannica.com › technology › information-retrieval

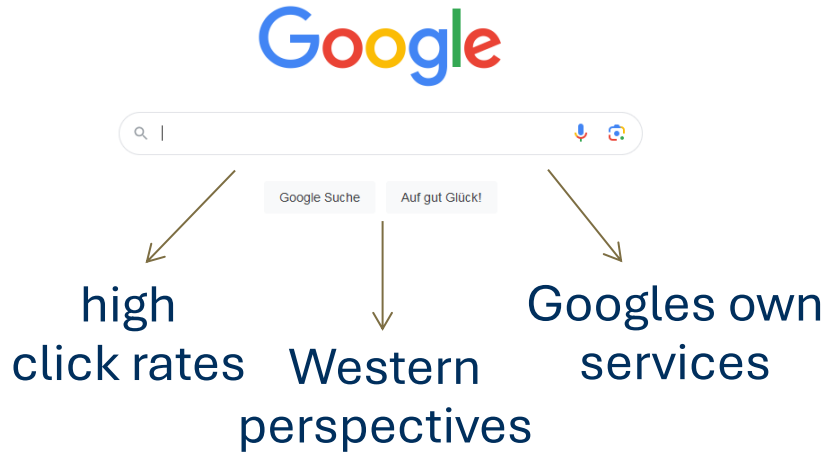
Information retrieval | Definition, Methods, & Facts | Britannica

information retrieval, recovery of **information**, especially in a database stored in a computer.Two main approaches are matching words in the query against the database index (keyword searching) and traversing the database using hypertext or hypermedia links. Evolving **information-retrieval** technique...

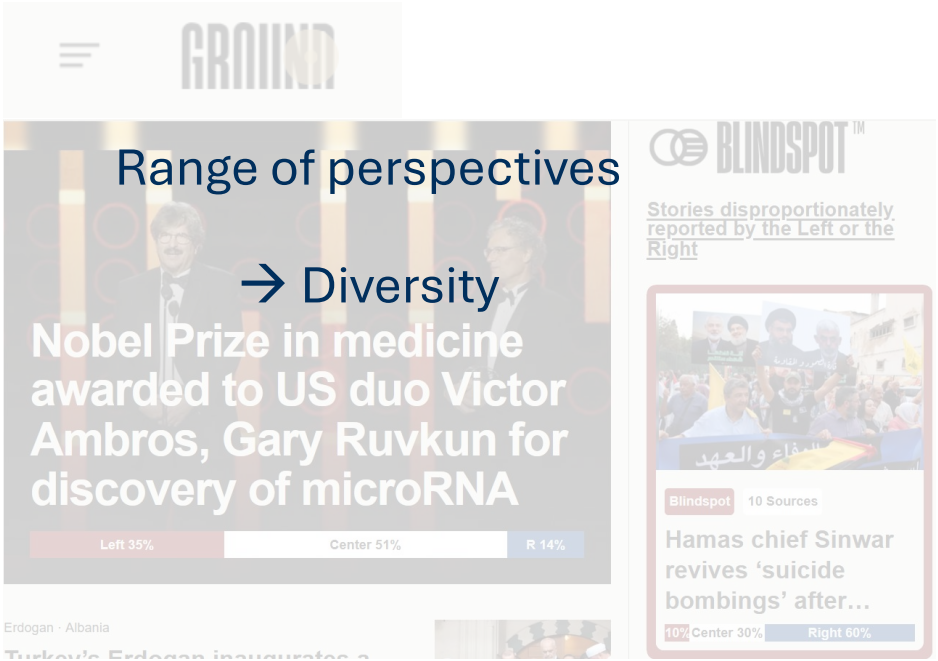
Why should we care about variety?



ground.news



Why should we care about variety?



Range of perspectives

→ Diversity

Characterization using 'protected attributes'

→ Fairness

high click rates

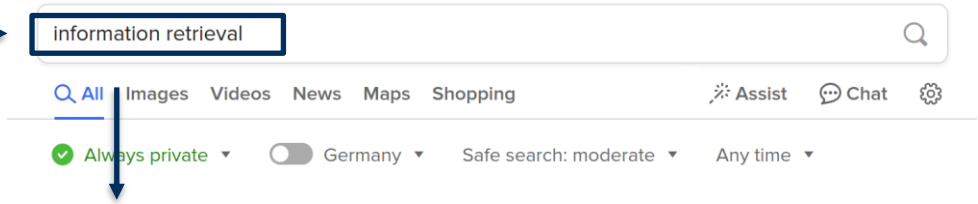
Western perspectives

Googles own services

Information Need

What is information retrieval about?

Query



Document Ranking

https://en.wikipedia.org › wiki › Information_retrieval

Information retrieval - Wikipedia

Information retrieval (IR) in computing and **information** science is the task of identifying and retrieving **information** system resources that are relevant to an **information** need.The **information** need can be specified in the form of a search query. In the case of document **retrieval**, queries can be based on ful...

https://www.geeksforgeeks.org › what-is-information-retrieval

What is Information Retrieval? - GeeksforGeeks

Sep 19, 2023 · **Information Retrieval** (IR) can be defined as a software program that deals with the organization, storage, **retrieval**, and evaluation of **information** from document repositories, particularly textual **information**. **Information Retrieval** is the activity of obtaining material that can usually be...

https://www.britannica.com › technology › information-retrieval

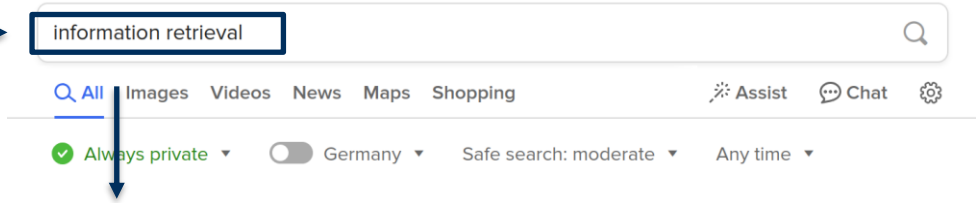
Information retrieval | Definition, Methods, & Facts | Britannica

information retrieval, recovery of **information**, especially in a database stored in a computer.Two main approaches are matching words in the query against the database index (keyword searching) and traversing the database using hypertext or hypermedia links. Evolving **information-retrieval** technique...

Information Need

What is information retrieval about?

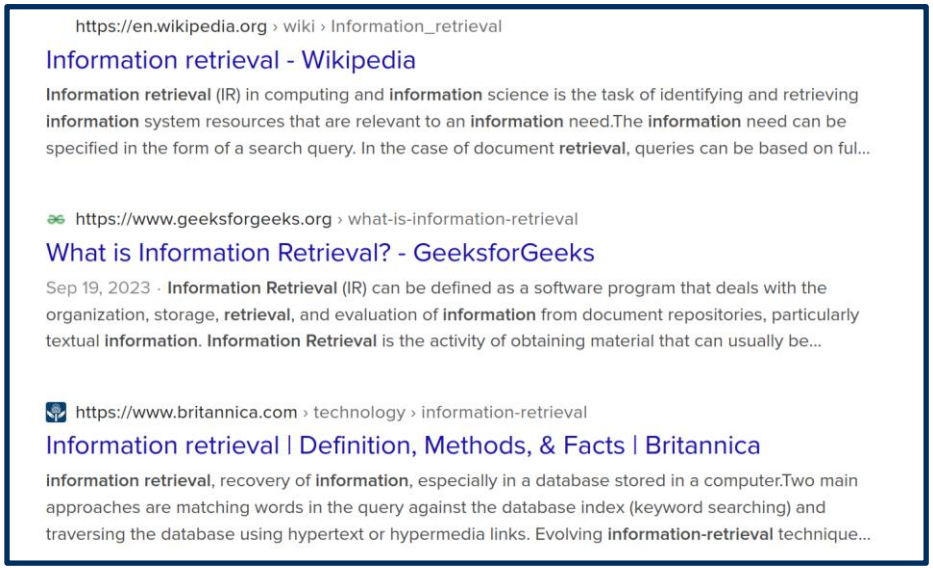
Query



Variety Setting

Diversity Degree δ

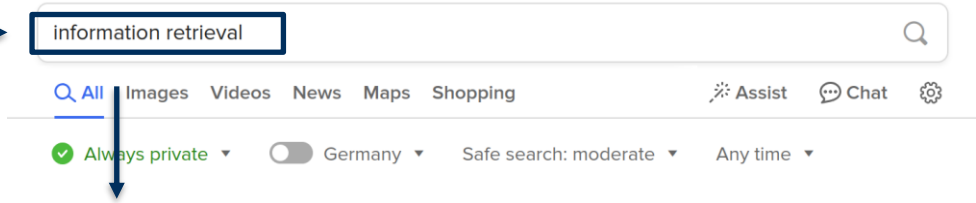
Document Ranking



Information Need

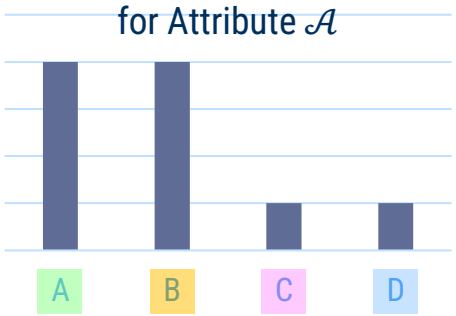
What is information retrieval about?

Query



Variety Setting

Target Distribution for Attribute $\mathcal{A}$



Document Ranking

https://en.wikipedia.org › wiki › Information_retrieval

Information retrieval - Wikipedia

Information retrieval (IR) in computing and **information** science is the task of identifying and retrieving **information** system resources that are relevant to an **information** need.The **information** need can be specified in the form of a search query. In the case of document **retrieval**, queries can be based on ful...

https://www.geeksforgeeks.org › what-is-information-retrieval

What is Information Retrieval? - GeeksforGeeks

Sep 19, 2023 · **Information Retrieval** (IR) can be defined as a software program that deals with the organization, storage, **retrieval**, and evaluation of **information** from document repositories, particularly textual **information**. **Information Retrieval** is the activity of obtaining material that can usually be...

https://www.britannica.com › technology › information-retrieval

Information retrieval | Definition, Methods, & Facts | Britannica

information retrieval, recovery of **information**, especially in a database stored in a computer.Two main approaches are matching words in the query against the database index (keyword searching) and traversing the database using hypertext or hypermedia links. Evolving **information-retrieval** technique...

Contributions

1. Include demand of a greater **range of perspectives** in search results using a topic hierarchy

2. Include demand of a more **fair representation** in search results using a protected attribute

3. Evaluate the fairness and diversity in search results

Data

Water Innovation Cluster

The **supply of clean water** is one of the biggest social challenges of the future, on a global scale and in Germany alike.

To **prevent water scarcity** and the potential for social conflict that is linked to it, **new alliances** across research, the economy, the public sector and civil society are necessary.

https://www.thwic.uni-jena.de/en; 19.08.2024

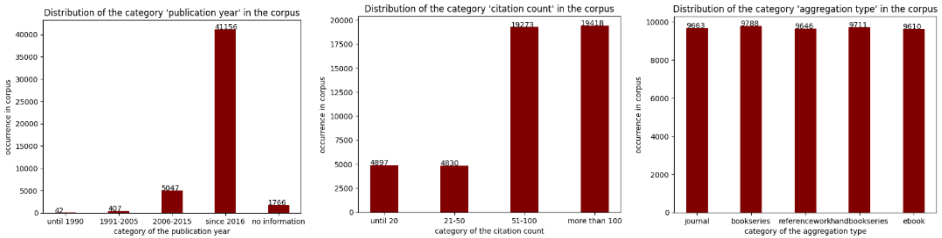
Literature Search

- crawled 48’000 publications from Elsevier/ScienceDirect
- 13 queries from chemistry
- Relevance judgments by one expert

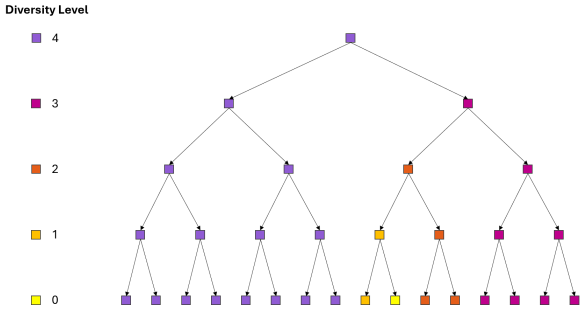
Literature Search

- crawled 48’000 publications from Elsevier/ScienceDirect
- 13 queries from chemistry
- Relevance judgments by one expert

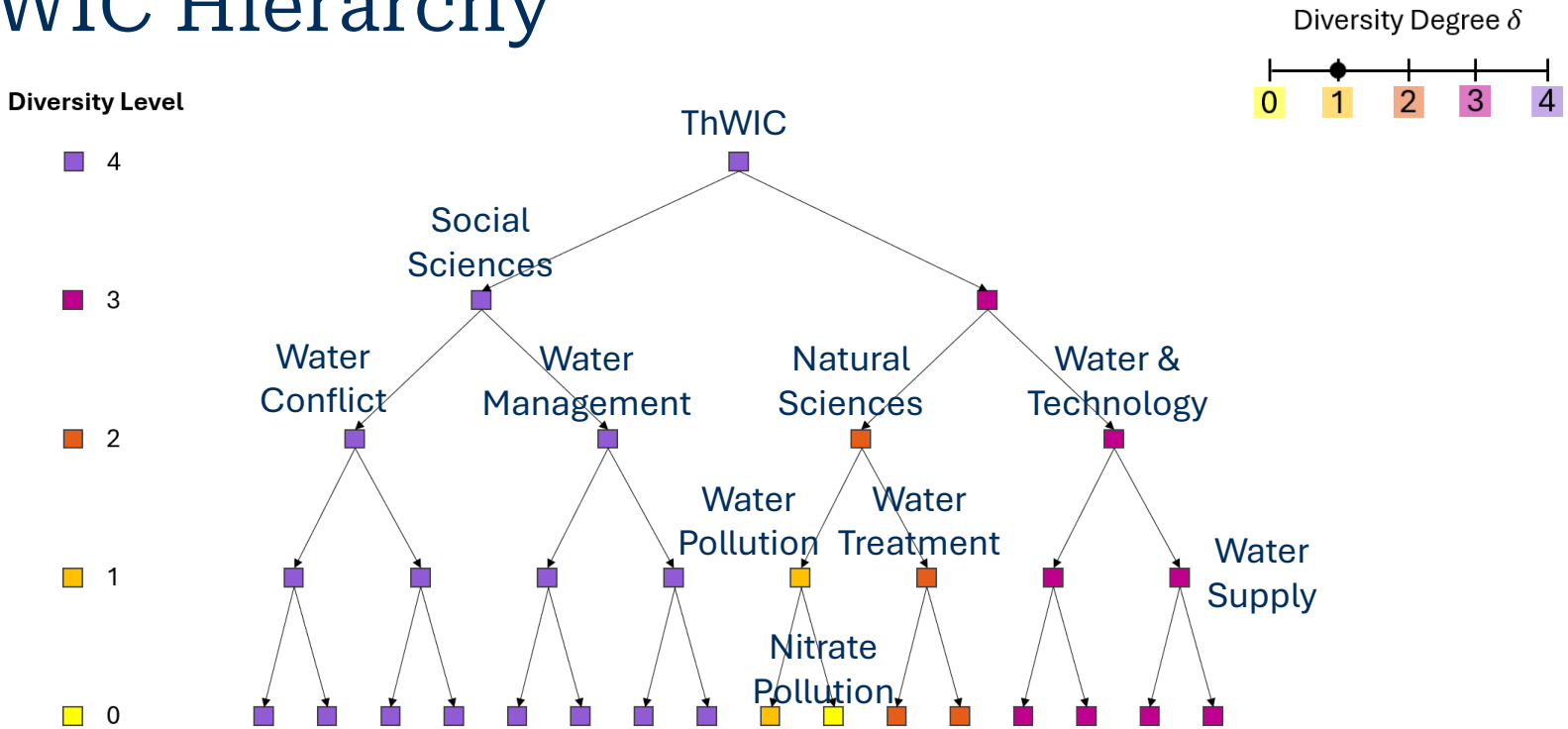
Attributes and Categories



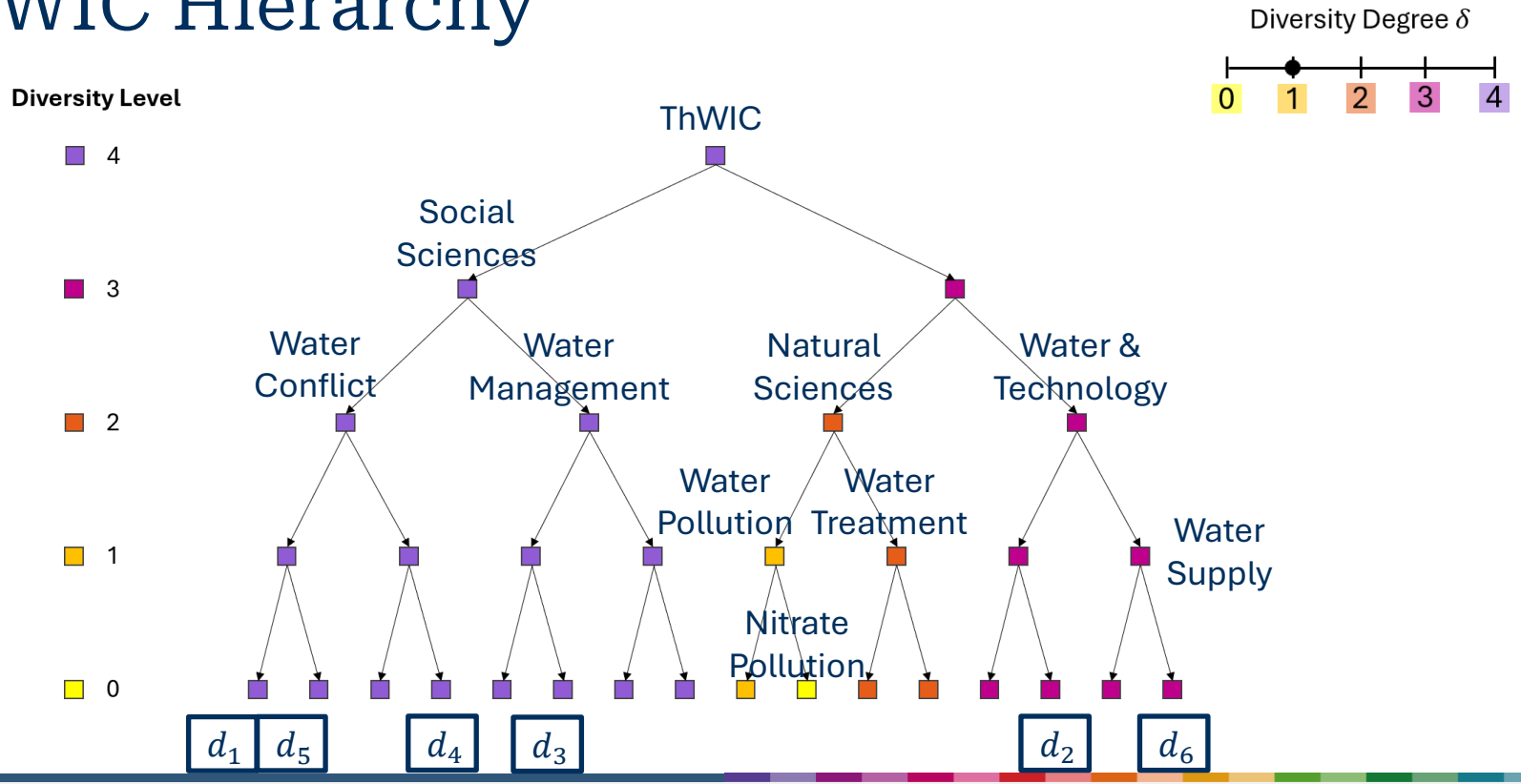
Topic Hierarchy



ThWIC Hierarchy



ThWIC Hierarchy



Diversity-Based Ranking

Diversity

The aim of diversity in search results is to present various aspects of a query.

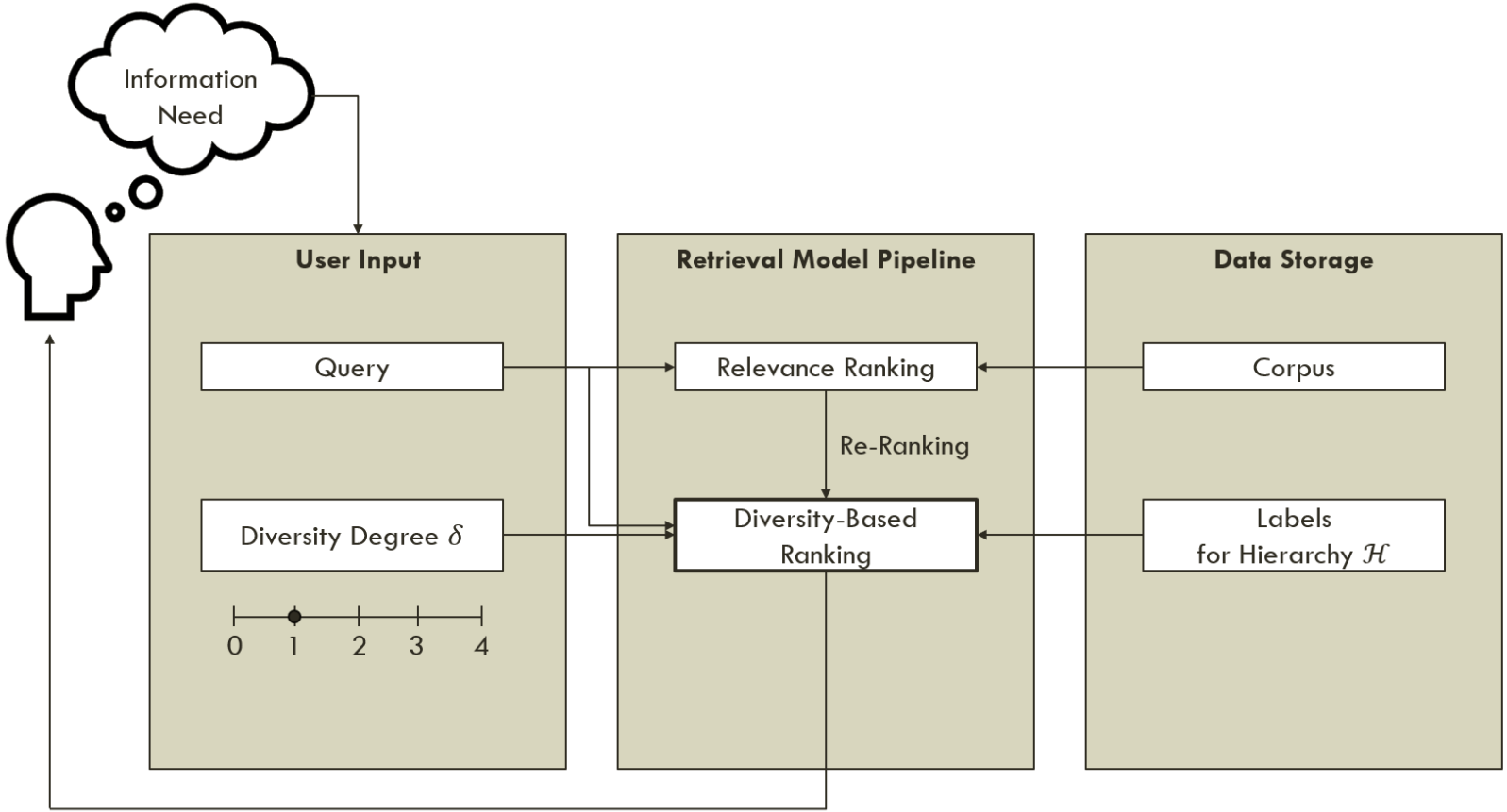
Economic perspective

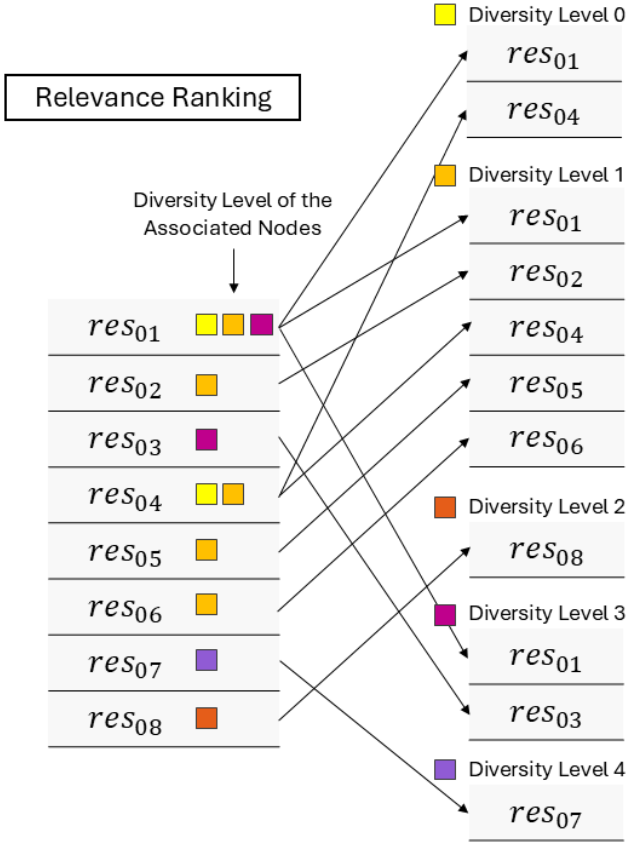
Chemical perspective

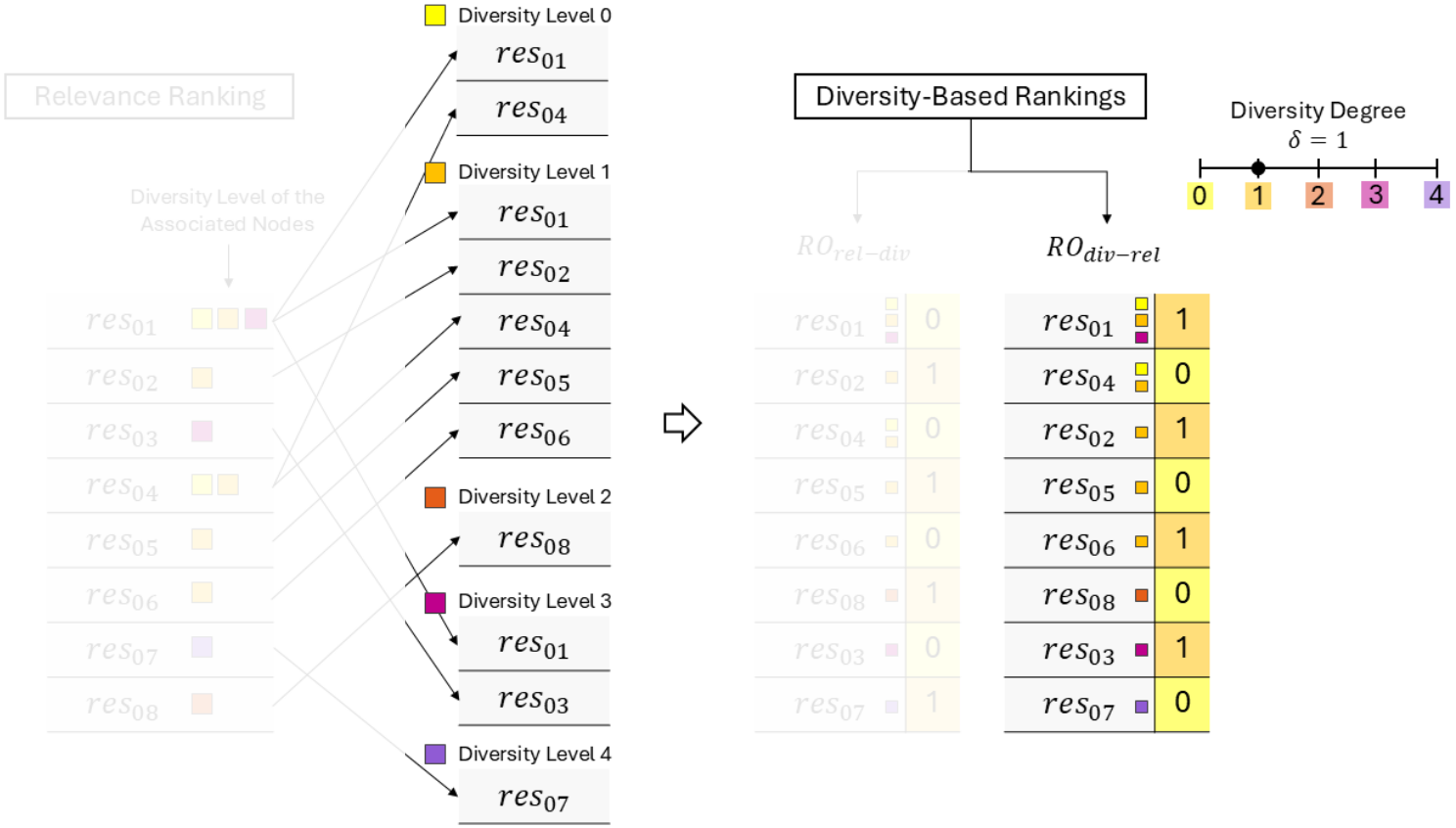
Social perspective

Geological perspective

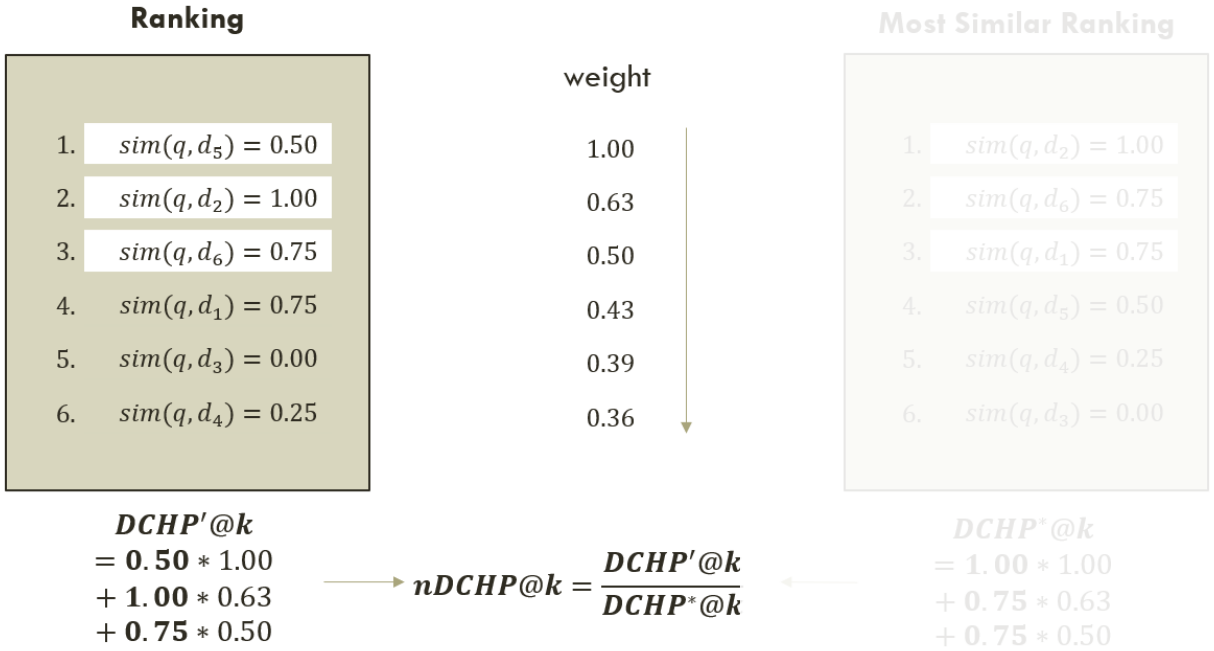
H Wu et al.: Result Diversification in Search and Recommendation: A Survey. arXiv, 2024


Information Need
User Input
Query
Diversity Degree δ
0 1 2 3 4
Retrieval Model Pipeline
Relevance Ranking
Re-Ranking
Diversity-Based Ranking
Data Storage
Corpus
Labels for Hierarchy $\mathcal{H}$


Relevance Ranking
Diversity Level of the Associated Nodes
res_{01}
res_{02}
res_{03}
res_{04}
res_{05}
res_{06}
res_{07}
res_{08}
Diversity Level 0
res_{01}
res_{04}
Diversity Level 1
res_{01}
res_{02}
res_{04}
res_{05}
res_{06}
Diversity Level 2
res_{08}
Diversity Level 3
res_{01}
res_{03}
Diversity Level 4
res_{07}


Relevance Ranking
Diversity Level of the Associated Nodes
res_{01}
res_{02}
res_{03}
res_{04}
res_{05}
res_{06}
res_{07}
res_{08}
Diversity Level 0
res_{01}
res_{04}
Diversity Level 1
res_{01}
res_{02}
res_{04}
res_{05}
res_{06}
Diversity Level 2
res_{08}
Diversity Level 3
res_{01}
res_{03}
Diversity Level 4
res_{07}
Diversity-Based Rankings
Diversity Degree
$\delta = 1$
0 1 2 3 4
$RO_{rel-div}$
$RO_{div-rel}$
res_{01} 0
res_{02} 1
res_{04} 0
res_{05} 1
res_{06} 0
res_{08} 1
res_{03} 0
res_{07} 1
res_{01} 1
res_{04} 0
res_{02} 1
res_{05} 0
res_{06} 1
res_{08} 0
res_{03} 1
res_{07} 0

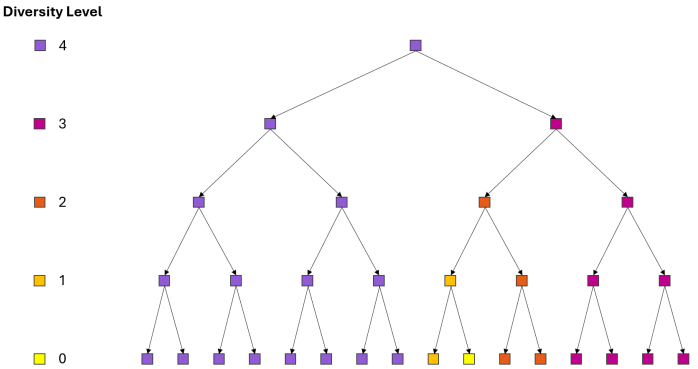
nDCHierarchical Precision



nDCHierarchical Precision

Ranking

		weight
1.	$sim(q, d_5) = 0.50$	1.00
2.	$sim(q, d_2) = 1.00$	0.63
3.	$sim(q, d_6) = 0.75$	0.50
4.	$sim(q, d_1) = 0.75$	0.43
5.	$sim(q, d_3) = 0.00$	0.39
6.	$sim(q, d_4) = 0.25$	0.36

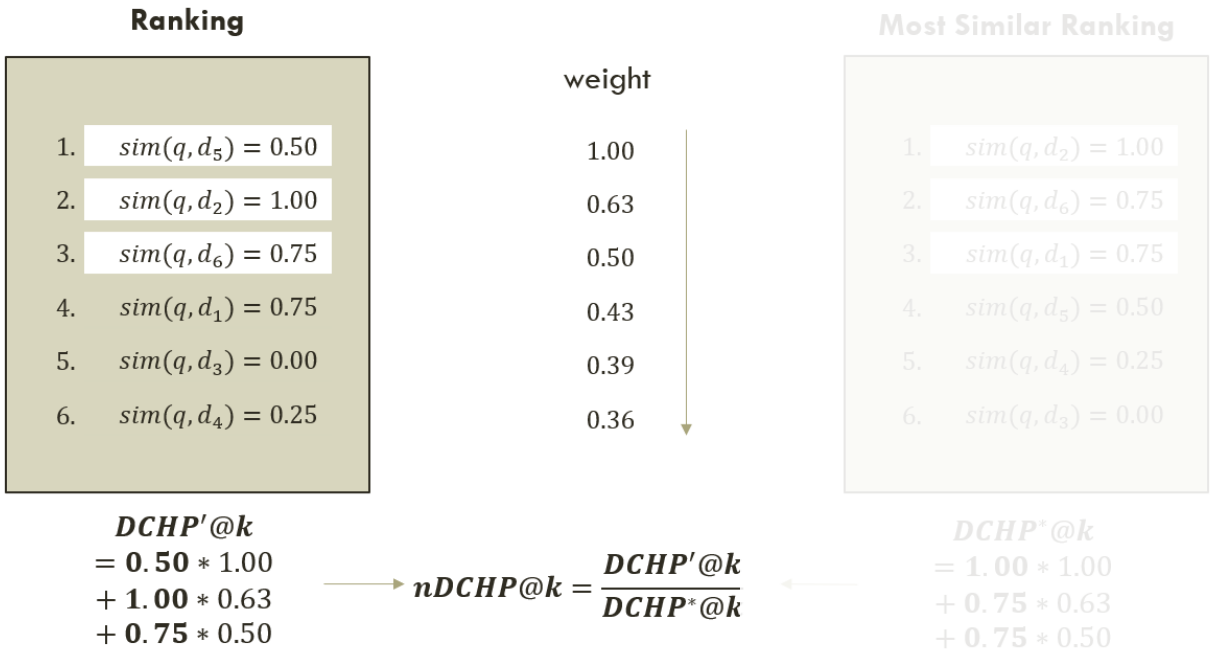


$$DCHP'@k = \mathbf{0.50} * 1.00 + \mathbf{1.00} * 0.63 + \mathbf{0.75} * 0.50$$

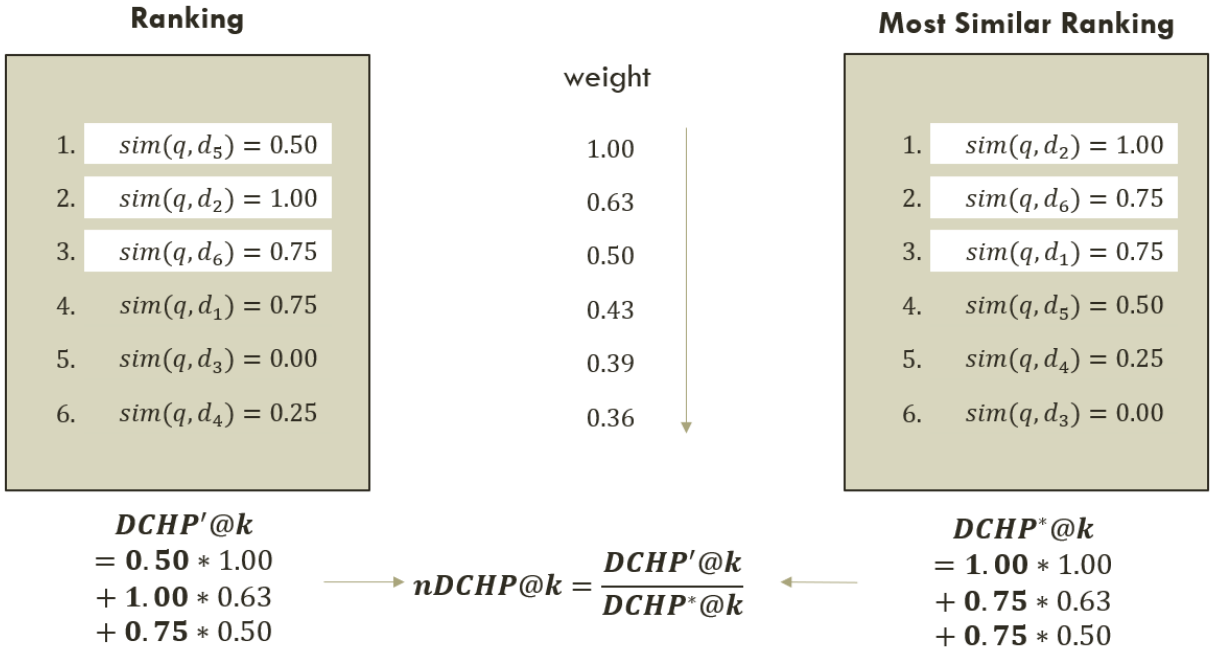
$$nDCHP@k = \frac{DCHP'@k}{DCHP^*@k}$$

$$DCHP^*@k = \mathbf{1.00} * 1.00 + \mathbf{0.75} * 0.63 + \mathbf{0.75} * 0.50$$

nDCHierarchical Precision



nDCHierarchical Precision



Results

δ		$nDCHP\downarrow$	
	rel-ranking	0.51987	±0.14979
0	O_{spec}	0.55740	±0.15201
	O_{div}	0.55740	±0.15201
1	O_{spec}	0.55662	±0.14983
	O_{div}	0.55690	±0.15007
2	O_{spec}	0.55212	±0.14432
	O_{div}	0.55119	±0.14443
3	O_{spec}	0.53921	±0.13909
	O_{div}	0.53439	±0.13946
4	O_{spec}	0.53277	±0.14848
	O_{div}	0.52557	±0.15012

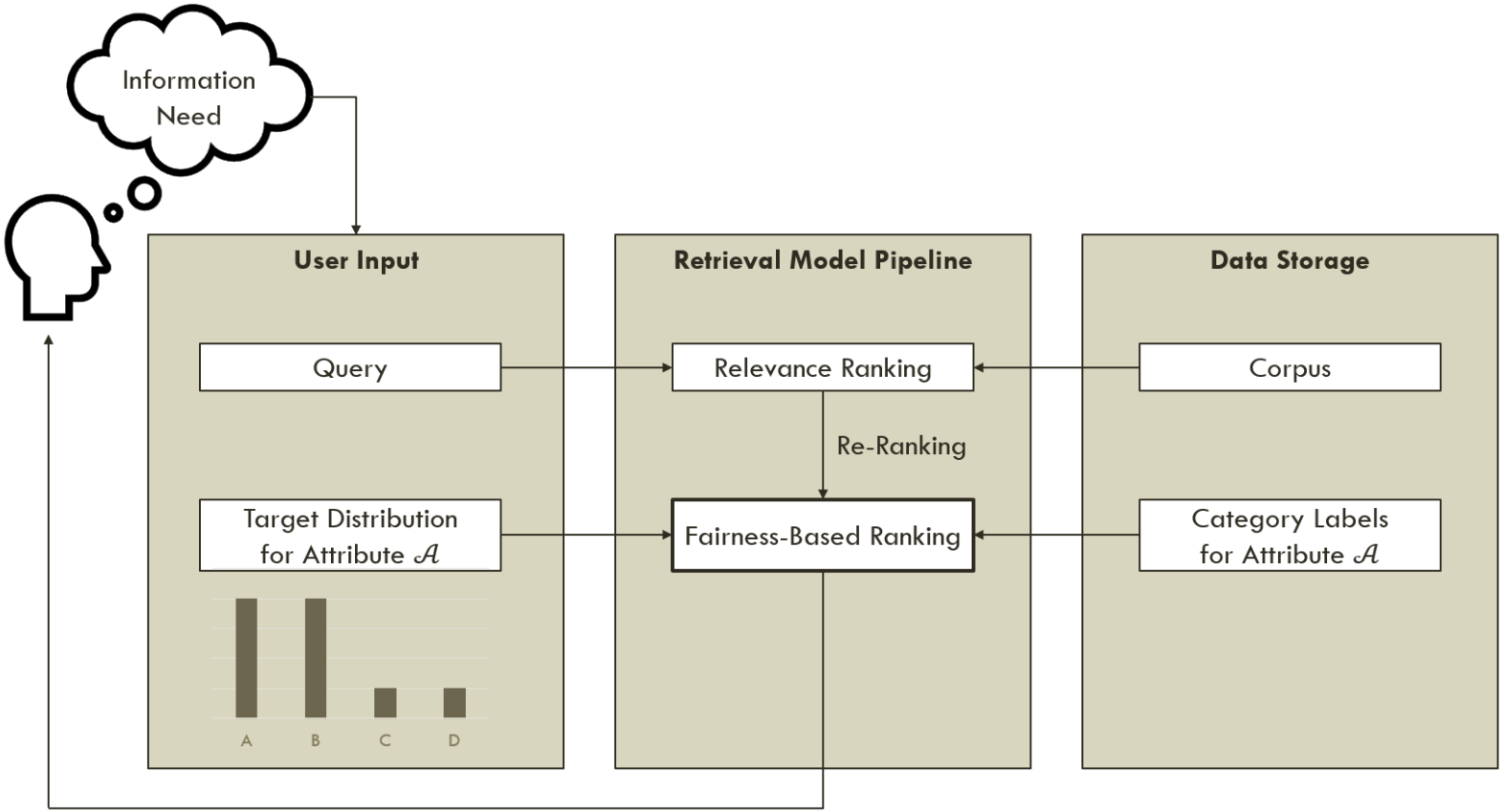
Fairness-Based Ranking

Fairness

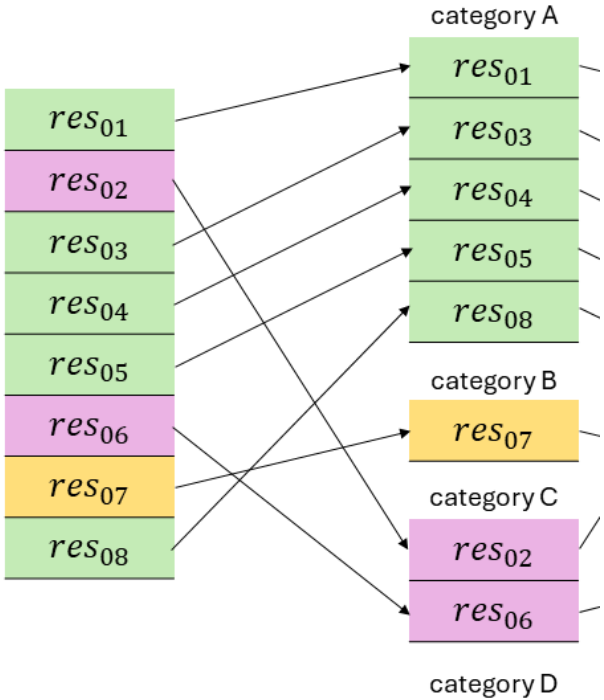
The aim of investigating algorithmic fairness is to identify, measure, and mitigate aspects that make a system unfair in a particular way.

However, “there is not one particular definition of what constitutes fairness.”

MD Ekstrand et al.: Fairness in Information Access Systems. In: Foundations and Trends in Information Retrieval 2022

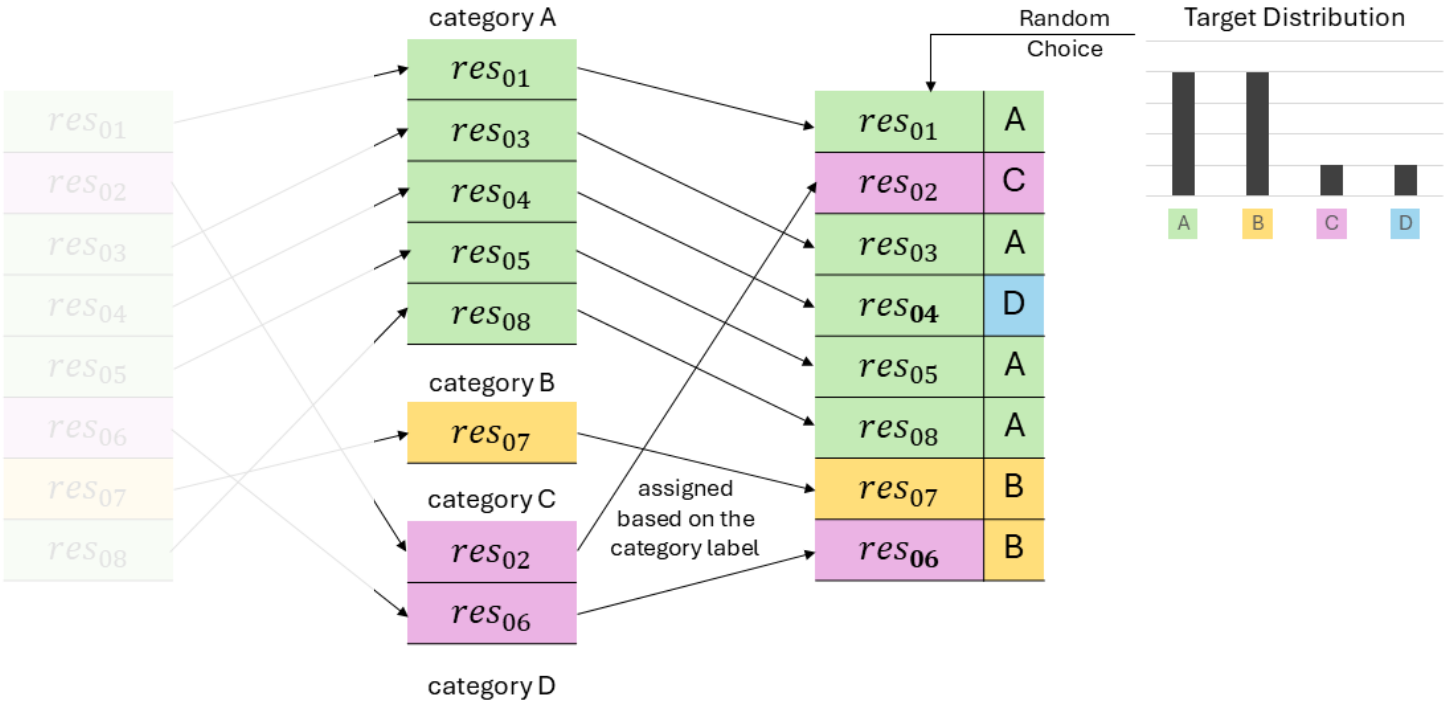

Information Need
User Input
Retrieval Model Pipeline
Data Storage
Query
Relevance Ranking
Corpus
Re-Ranking
Target Distribution for Attribute $\mathcal{A}$
Fairness-Based Ranking
Category Labels for Attribute $\mathcal{A}$
A
B
C
D

Relevance Ranking



Relevance Ranking

Fairness-Based Ranking



Distribution Shift

	until 1990	1991-2005	2006-2015	since 2016	unknown
document corpus	**0.001**	**0.008**	**0.103**	**0.851**	**0.037**
rel-ranking@500	0.000	0.006	0.094	0.874	0.026
rel-ranking@100	0.00	0.00	0.09	0.86	0.05
rel-ranking@20	0.00	0.00	0.10	0.85	0.05
rel-ranking@10	0.00	0.00	0.00	0.90	0.10
cat-pub@100	0.00	0.03	0.30	0.64	0.03
cat-pub@20	0.00	0.15	0.35	0.45	0.05
cat-pub@10	0.00	0.30	0.30	0.40	0.00
target distribution	**0.10**	**0.20**	**0.30**	**0.40**	**0.00**

Table 6.5: Distribution of the categories of attribute *publication year* when re-ranking the relevance-ranking of query 4. The target distribution aims to mitigate the strong bias towards category *since 2016*.

Conclusion

User-Driven Re-Ranking for Adapting the Variety in Search Results

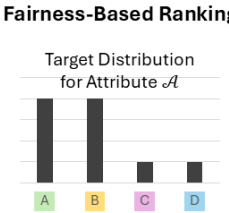
1. Diversity-based re-ranking
   with a topic hierarchy

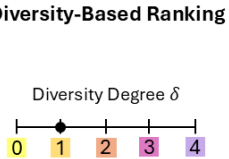
2. Fairness-based re-ranking
   with a protected attribute

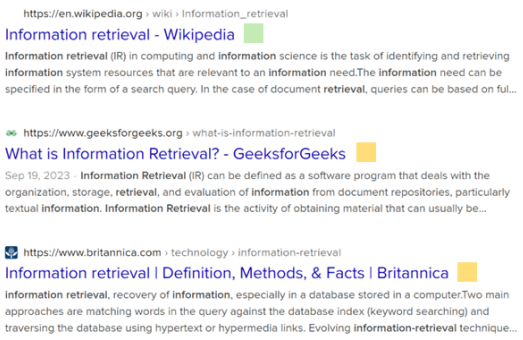
3. Extension of an
   evaluation metric
   for hierarchies in retrieval

User-Driven Re-Ranking for Adapting the Variety in Search Results

1. Diversity-based re-ranking
   with a topic hierarchy
2. Fairness-based re-ranking
   with a protected attribute
3. Extension of an
   evaluation metric
   for hierarchies in retrieval







The Flute Parable

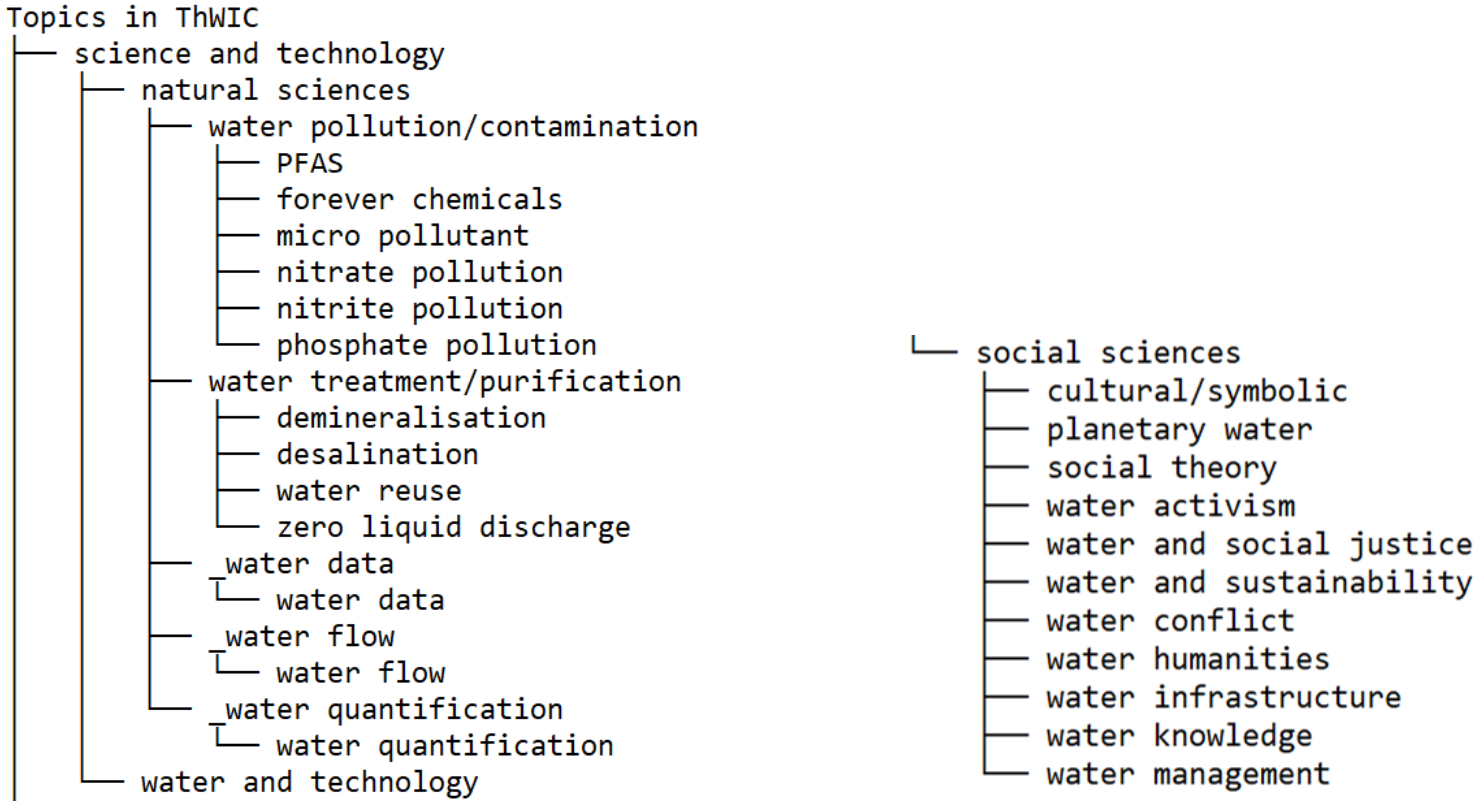
Three children want to own a flute, but only one is allowed to have it.

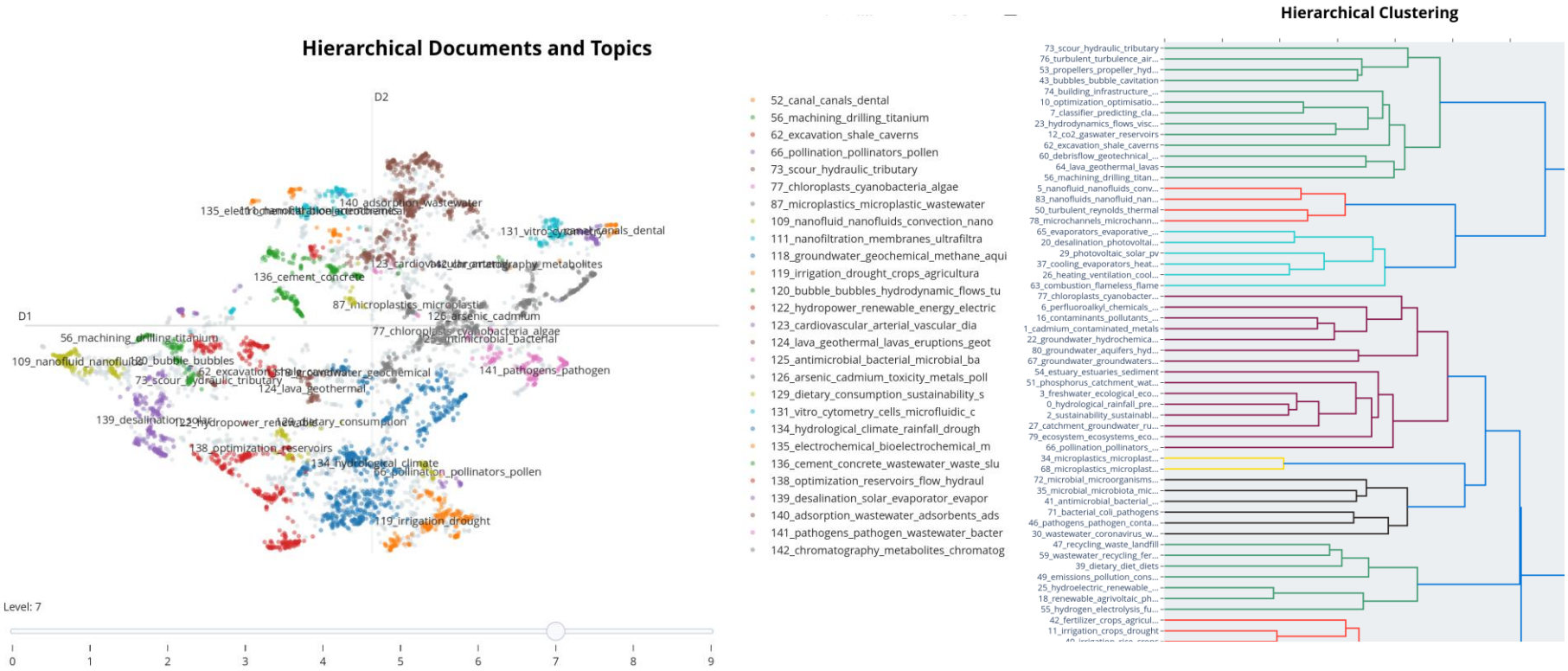
Would you give it:

A) to the child that produced the flute,

B) to the child that has no other toys, or

C) to the child that can play the flute magically?


Topics in ThWIC
science and technology
natural sciences
water pollution/contamination
PFAS
forever chemicals
micro pollutant
nitrate pollution
nitrite pollution
phosphate pollution
water treatment/purification
demineralisation
desalination
water reuse
zero liquid discharge
_water data
water data
_water flow
water flow
_water quantification
water quantification
water and technology
social sciences
cultural/symbolic
planetary water
social theory
water activism
water and social justice
water and sustainability
water conflict
water humanities
water infrastructure
water knowledge
water management


Hierarchical Documents and Topics
D2
D1
140_adsorption_wastewater
131_vitro_cytometry_cells_microfluidic_c
136_cement_concrete
87_microplastics_microplastic
125_arsenic_cadmium
77_chloroplasts_cyanobacteria_algae
25_antimicrobial_bacterial
56_machining_drilling_titanium
109_nanofluid_nanofluids
120_bubble_bubbles
62_excavation_shale_caverns
73_scour_hydraulic_tributary
124_lava_geothermal
141_pathogens_pathogen
139_desalination_solar
122_hydropower_renewable_energy_electric
138_optimization_reservoirs
134_hydrological_climate
66_pollination_pollinators_pollen
119_irrigation_drought
Level: 7
0
1
2
3
4
5
6
7
8
9
52_canal_canals_dental
56_machining_drilling_titanium
62_excavation_shale_caverns
66_pollination_pollinators_pollen
73_scour_hydraulic_tributary
77_chloroplasts_cyanobacteria_algae
87_microplastics_microplastic_wastewater
109_nanofluid_nanofluids_convection_nano
111_nanofiltration_membranes_ultrafiltra
118_groundwater_geochemical_methane_aqui
119_irrigation_drought_crops_agricultura
120_bubble_bubbles_hydrodynamic_flows_tu
122_hydropower_renewable_energy_electric
123_cardiovascular_arterial_vascular_dia
124_lava_geothermal_lavas_eruptions_geot
125_antimicrobial_bacterial_microbial_ba
126_arsenic_cadmium_toxicity_metals_poll
129_dietary_consumption_sustainability_s
131_vitro_cytometry_cells_microfluidic_c
134_hydrological_climate_rainfall_drough
135_electrochemical_bioelectrochemical_m
136_cement_concrete_wastewater_waste_slu
138_optimization_reservoirs_flow_hydraul
139_desalination_solar_evaporator_evapor
140_adsorption_wastewater_adsorbents_ads
141_pathogens_pathogen_wastewater_bacter
142_chromatography_metabolites_chromatog
Hierarchical Clustering
73_scour_hydraulic_tributary
76_turbulent_turbulence_air...
53_propellers_propeller_hyd...
43_bubbles_bubble_cavitation
74_building_infrastructure_...
10_optimization_optimisatio...
7_classifier_predicting_cla...
23_hydrodynamics_flows_visc...
12_co2_gaswater_reservoirs
62_excavation_shale_caverns
60_debrisflow_geotechnical_...
64_lava_geothermal_lavas
56_machining_drilling_titan...
5_nanofluid_nanofluids_conv...
83_nanofluids_nanofluid_nan...
50_turbulent_reynolds_thermal
78_microchannels_microchann...
65_evaporators_evaporative_...
20_desalination_photovoltai...
29_photovoltaic_solar_pv
37_cooling_evaporators_heat...
26_heating_ventilation_cool...
63_combustion_flameless_flame
77_chloroplasts_cyanobacter...
6_perfluoroalkyl_chemicals_...
16_contaminants_pollutants_...
1_cadmium_contaminated_metals
22_groundwater_hydrochemica...
80_groundwater_aquifers_hyd...
67_groundwater_groundwaters...
54_estuary_estuaries_sediment
51_phosphorus_catchment_wat...
3_freshwater_ecological_eco...
0_hydrological_rainfall_pre...
2_sustainability_sustainabl...
27_catchment_groundwater_ru...
79_ecosystem_ecosystems_eco...
66_pollination_pollinators_...
34_microplastics_microplast...
68_microplastics_microplast...
72_microbial_microorganisms...
35_microbial_microbiota_mic...
41_antimicrobial_bacterial_...
71_bacterial_coli_pathogens
46_pathogens_pathogen_conta...
30_wastewater_coronavirus_w...
47_recycling_waste_landfill
59_wastewater_recycling_fer...
39_dietary_diet_diets
49_emissions_pollution_cons...
25_hydroelectric_renewable_...
18_renewable_agrivoltaic_ph...
55_hydrogen_electrolysis_fu...
42_fertilizer_crops_agricul...
11_irrigation_crops_drought

	journal	bookseries	referencework	handbookseries	ebook
document corpus	**0.199**	**0.200**	**0.200**	**0.203**	**0.198**
rel-ranking@500	0.190	0.206	0.174	0.250	0.180
rel-ranking@100	0.16	0.23	0.14	0.23	0.24
rel-ranking@20	0.30	0.25	0.20	0.10	0.15
rel-ranking@10	0.50	0.30	0.00	0.10	0.10
cat-agg@100	0.62	0.07	0.09	0.00	0.22
cat-agg@20	0.55	0.05	0.15	0.00	0.25
cat-agg@10	0.60	0.00	0.10	0.00	0.30
target distribution	**0.60**	**0.10**	**0.10**	**0.00**	**0.20**

Table 6.9: Distribution of the categories of attribute *aggregation type* when re-ranking the relevance-ranking of query 1. The target distribution aims to introduce a strong bias towards category *journal*.

	until 20	21-50	51-100	more than 100
document corpus	**0.103**	**0.097**	**0.401**	**0.399**
rel-ranking@500	0.098	0.088	0.408	0.406
rel-ranking@100	0.09	0.12	0.39	0.40
rel-ranking@20	0.05	0.25	0.30	0.40
rel-ranking@10	0.10	0.20	0.30	0.40
cat-cit@100	0.21	0.23	0.30	0.26
cat-cit@20	0.20	0.25	0.30	0.25
cat-cit@10	0.20	0.30	0.30	0.20
target distribution	**0.25**	**0.25**	**0.25**	**0.25**

Table 6.7: Distribution of the categories of attribute *citation count* when re-ranking the relevance-ranking of query 12. The target distribution aims at a uniform distribution of all categories.

	$AWRF_{uniform} \uparrow$	
rel-ranking	0.63905	$\pm$ 0.01181
cat-pub	0.65660	$\pm$ 0.01216
cat-cit	0.64513	$\pm$ 0.01180
cat-agg	0.63736	$\pm$ 0.01128

Chemistry data

Pipeline	nDCG @10	Success @10	Success @5	Precision @10	Precision @5
BM25→$\mathcal{C}$					
cutoff@100	**.810**	**1.00**	**1.00**	**.586**	**.729**
cutoff@500	.763	**1.00**	**1.00**	.529	.714
DFIC→$\mathcal{B}$					
cutoff@100	.341	.714	.571	.357	.400
cutoff@500	.443	.786	.786	.393	.471

Data from natural and social scienes

Pipeline	nDCG @10	Success @10	Success @5	Precision @10	Precision @5
BM25→$\mathcal{C}$					
cutoff@500	.338	.462	.385	.215	.277