Robust Information Retrieval

WSDM 2025 tutorial

Yu-An Liu[a,b], Ruqing Zhang[a,b], Jiafeng Guo[a,b] and **Maarten de Rijke**[c]

https://wsdm2025-robust-information-retrieval.github.io/

March 10, 2025

01:30 – 05:00 PM

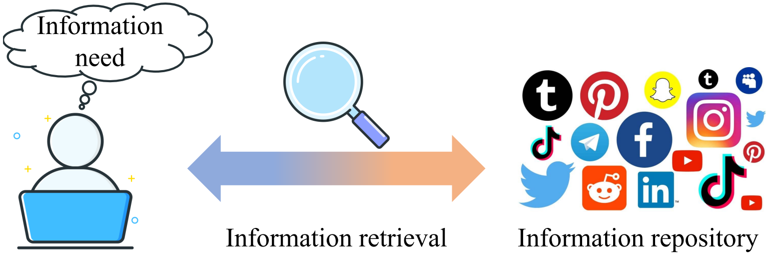
[a] Institute of Computing Technology, Chinese Academy of Sciences
[b] University of Chinese Academy of Sciences
[c] University of Amsterdam

About the presenters

			
Yu-An Liu	Ruqing Zhang	Jiafeng Guo	**Maarten de Rijke**
Phd student	Faculty	Faculty	Faculty
@ICT, CAS	@ICT, CAS	@ICT, CAS	@UvA

SCAN
ME

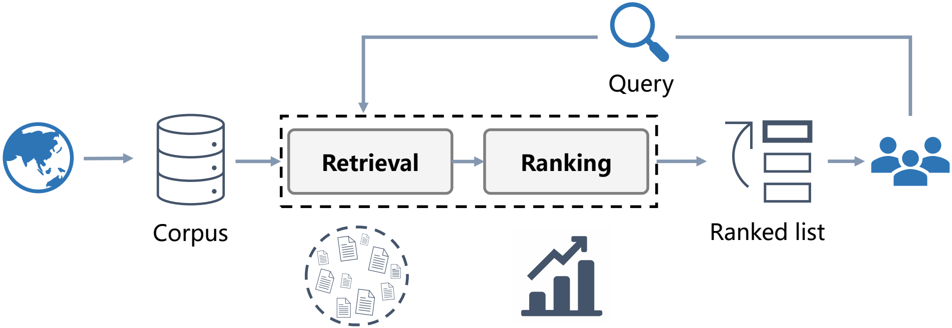
Information retrieval (IR) is the activity of obtaining information resources that are relevant to an information need from a collection of those resources.



Given: User query (keywords, question, image, . . .)
Rank: Information objects (passages, documents, images, products, . . .)
Ordered by: Relevance scores

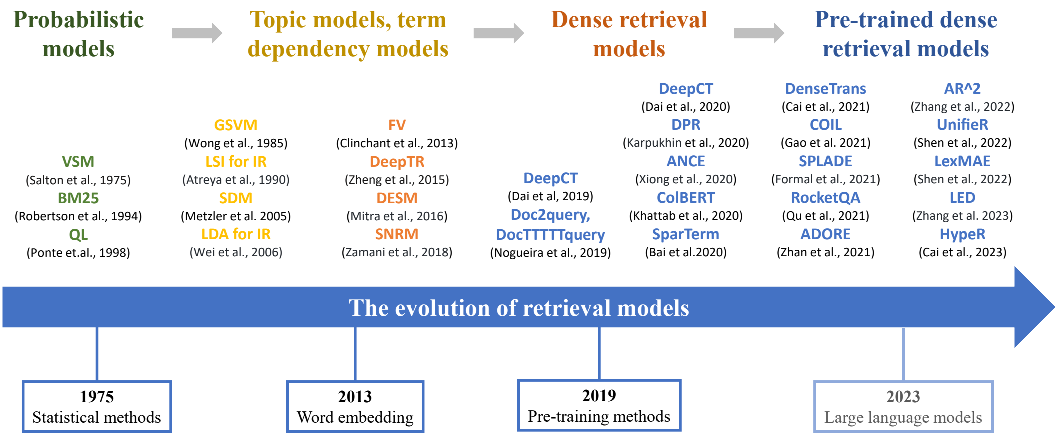
Application of information retrieval systems

Core pipelined paradigm: Retrieval-Ranking



- **Retrieval**: Find an initial set of candidate documents for a query
- **Ranking**: Determine the relevance degree of each candidate

Evolution of retrieval models



Evolution of ranking models

Probabilistic models → **Learning to rank models** → **Neural ranking models** → **Pre-trained neural ranking models**

Probabilistic models
- VSM (Salton et al., 1975)
- BM25 (Robertson et al., 1994)
- QL (Ponte et.al., 1998)

Learning to rank models
- RankSVM (Herbrich et al., 1999)
- Prank (Crammer et al., 2001)
- RankNet (Burges et al. 2005)
- ListNet (Cao et al., 2007)
- LambdaMart (Burges et al. 2010)

Neural ranking models
- DSSM (Huang et al., 2013)
- DRMM (Guo et al., 2016)
- Duet (Mitra et al., 2017)
- Conv-KNRM (Dai et al., 2018)

Pre-trained neural ranking models
- monoBERT (Nogueira et al, 2019)
- Expando-Mono-Duo (Nogueira et al., 2019)
- CEDR (MacAvaney et al., 2020)
- BERT-MaxP (Dai et al., 2020)
- PARADE (Li et al., 2020)
- BERT-QE (Zheng et al., 2020)
- ReInfoSelect (Zhang et al.2020)
- GDMTL (Liu et al., 2021)
- PROP, B-PROP (Ma et al. 2021)
- HARP (Ma et al., 2021)
- UED (Yan et al., 2021)
- RocketQAv2 (Ren et al., 2021)
- RankT5 (Zhuang et al., 2022)
- ARES (Chen et al., 2022)
- Webformer (Guo et al., 2022)
- RankGPT (Sun et al. 2023)
- ExaRanker (Ferraretto et al., 2023)

The evolution of ranking models

- **1975** Statistical methods
- **2013** Word embedding
- **2019** Pre-training methods
- **2023** Large language models

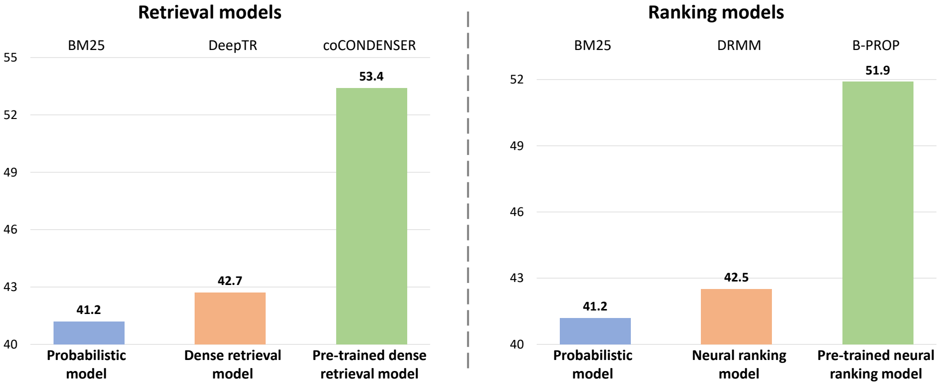
Neural IR models, including **dense retrieval models (DRMs)** and **neural ranking models (NRMs)**, have achieved promising ranking effectiveness

Data: [Dai and Callan, 2019, Lee et al., 2023, Ma et al., 2021]

Effectiveness of neural IR models

Neural IR models, including **dense retrieval models (DRMs)** and **neural ranking models (NRMs)**, have achieved promising ranking effectiveness

Let's take the NDCG@20 performance on TREC Robust04 as an example:

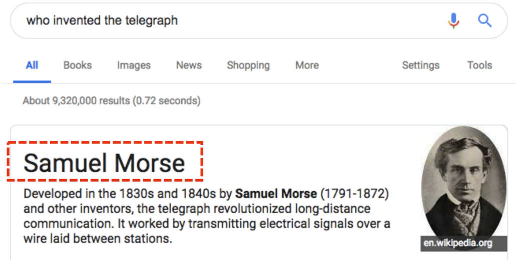


Data: [Dai and Callan, 2019, Lee et al., 2023, Ma et al., 2021]

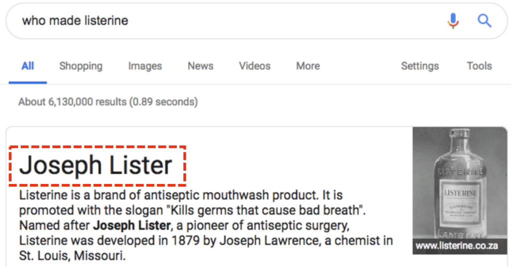
Beyond effectiveness, what are the challenges we face when applying neural IR models in the real world?

Challenges 1: Performance fluctuations between queries

Major web search engine makes over 3,200 changes to its search algorithms in a year to optimize underperforming search results for a small number of queries



(a) A correct answer for the query "*who invented the telegraph*".

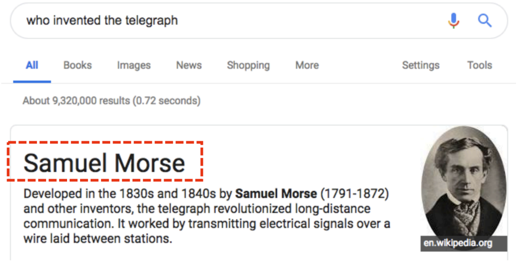


(b) A wrong answer for the query "*who made listerine*".

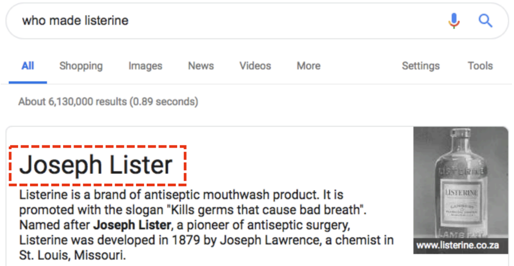
Data: How We Keep Search Relevant and Useful; Image: [Su et al., 2019]

Challenges 1: Performance fluctuations between queries

Major web search engine makes over 3,200 changes to its search algorithms in a year to optimize underperforming search results for a small number of queries



(a) A **correct answer** for the query "*who invented the telegraph*".



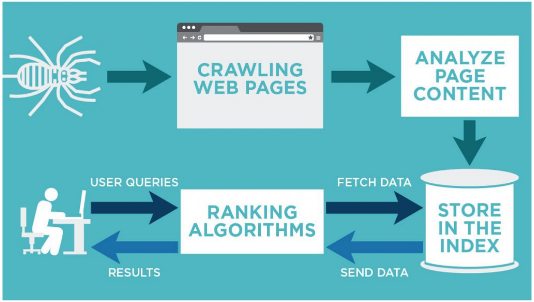
(b) A **wrong answer** for the query "*who made listerine*".

Neural IR models need to **avoid performance fluctuations** between queries

Data: How We Keep Search Relevant and Useful; Image: [Su et al., 2019]

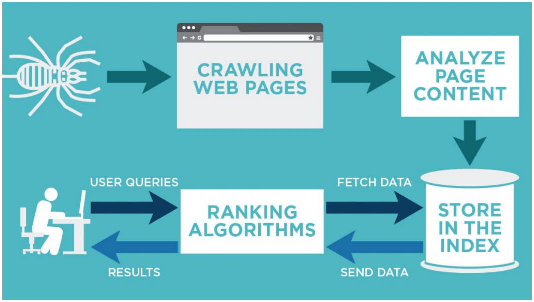
Challenges 2: A dynamic flow of new data

Every day, billions of new web pages emerge and 15% of search queries are brand new



Challenges 2: A dynamic flow of new data

Every day, billions of new web pages emerge and 15% of search queries are brand new



Neural IR models need to continuously **adapt to new queries and documents**

Data: Forbes: Top Website Statistics For 2024; Image: How Do. . .

Challenges 3: Search engine optimization (SEO)

About 60% of marketers get quality leads by SEO, and it can drive over 1,000% more traffic than before, with a 14.6% conversion rate

Data: SEO Statistics - 2024; Image: Negative SEO

Challenges 3: Search engine optimization (SEO)

About 60% of marketers get quality leads by SEO, and it can drive over 1,000% more traffic than before, with a 14.6% conversion rate

Neural IR models need to be able to **withstand potential SEO attacks**

Data: SEO Statistics - 2024; Image: Negative SEO

Distinct from effectiveness, these challenges can be characterized as robustness

Robustness refers to the ability of a system to withstand disturbances or external factors that may cause it to malfunction or provide inaccurate results.

Effectiveness
The average performance under normal purpose

→

Robustness
The performance in abnormal situations

"Theoretically principled trade-off between robustness and accuracy" [Zhang et al., 2019]

What is the robustness in IR?

There is a large volume of work that covers many aspects of IR robustness, e.g.,

"Robust Neural Information Retrieval" [Liu et al., 2024]; "Are Neural Ranking Model Robust?" [Wu et al., 2022]

What is the robustness in IR?

There is a large volume of work that covers many aspects of IR robustness, e.g.,

- **Performance variance** emphasizes the worst-case performance across different individual queries under the independent and identically distributed (IID) data

"Robust Neural Information Retrieval" [Liu et al., 2024]; "Are Neural Ranking Model Robust?" [Wu et al., 2022]

What is the robustness in IR?

There is a large volume of work that covers many aspects of IR robustness, e.g.,

- **Performance variance** emphasizes the worst-case performance across different individual queries under the independent and identically distributed (IID) data
- **Out-of-distribution (OOD) robustness** measures the performance on unseen queries and documents from different distributions of the training dataset

"Robust Neural Information Retrieval" [Liu et al., 2024]; "Are Neural Ranking Model Robust?" [Wu et al., 2022]

What is the robustness in IR?

There is a large volume of work that covers many aspects of IR robustness, e.g.,

- **Performance variance** emphasizes the worst-case performance across different individual queries under the independent and identically distributed (IID) data
- **Out-of-distribution (OOD) robustness** measures the performance on unseen queries and documents from different distributions of the training dataset
- **Adversarial robustness** focuses on the ability to defend against malicious adversarial attacks aimed at manipulating rankings

"Robust Neural Information Retrieval" [Liu et al., 2024]; "Are Neural Ranking Model Robust?" [Wu et al., 2022]

Impact of poor robustness on IR systems

If we only focus on effectiveness while ignoring robustness . . .

Impact of poor robustness on IR systems

If we only focus on effectiveness while ignoring robustness . . .

- Search engine results pages may be flooded with commercial websites that manipulate rankings

Impact of poor robustness on IR systems

If we only focus on effectiveness while ignoring robustness . . .

- Search engine results pages may be flooded with commercial websites that manipulate rankings
- When we want to explore a new topic, it's difficult to find relevant results

Impact of poor robustness on IR systems

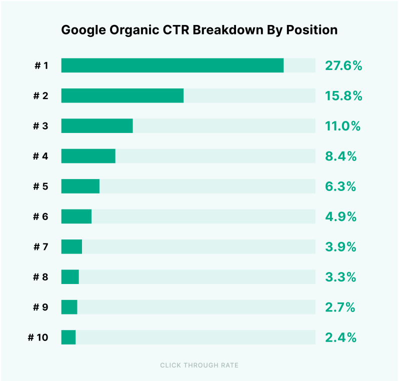
If we only focus on effectiveness while ignoring robustness . . .

- Search engine results pages may be flooded with commercial websites that manipulate rankings
- When we want to explore a new topic, it's difficult to find relevant results

If these robustness issues are unresolved, they can directly impact user satisfaction, which in turn hinder the widespread adoption of neural IR models

Can we follow the experience of other fields to solve the robustness issues in IR?

User attention mainly focuses on the Top-K results and increases with higher rankings



The core of robust IR is to protect the stability of the Top-K results

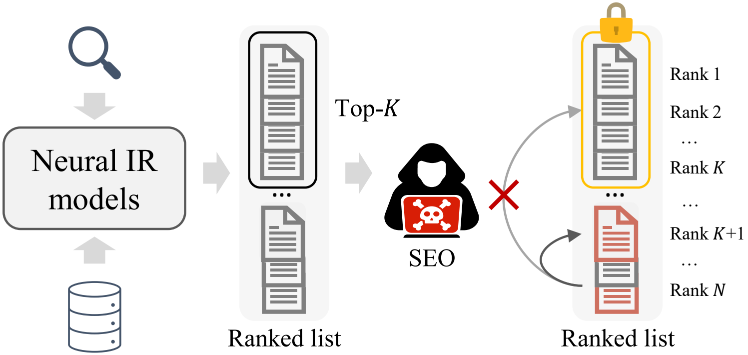


Image: How Do You Google? and What We Learned About . . .

Comparison with CV and NLP

	CV	NLP	IR
Representative task	Image classification	Text classification	Document ranking
Input format	Single image	Single text	Paired text
Input space	Continuous	Discrete	Discrete
Robustness requirement	Stability of classification (dog or cat)	Stability of classification (pos or neg)	Stability of top-K result (permutation maintenance)

normal challenging hard

Comparison with CV and NLP

	CV	NLP	IR
Representative task	Image classification	Text classification	Document ranking
Input format	Single image 🙂	Single text 🙂	Paired text 🤔
Input space	Continuous 🙂	Discrete 🤔	Discrete 🤔
Robustness requirement	Stability of classification 🤔 (dog or cat)	Stability of classification 🤔 (pos or neg)	Stability of top-K result 🤯 (permutation maintenance)

🙂 normal 🤔 challenging 🤯 hard

Experiences from other fields may not be as effective in IR 😦

Comparison with CV and NLP

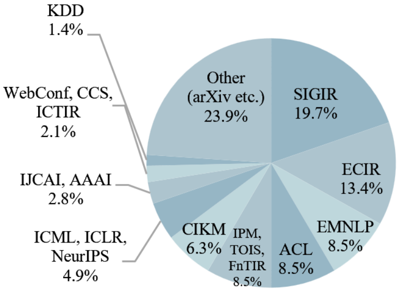
	CV	NLP	IR
Representative task	Image classification	Text classification	Document ranking
Input format	Single image 🙂	Single text 🙂	Paired text 🤔
Input space	Continuous 🙂	Discrete 🤔	Discrete 🤔
Robustness requirement	Stability of classification 🤔 (dog or cat)	Stability of classification 🤔 (pos or neg)	Stability of top-K result 🤯 (permutation maintenance)

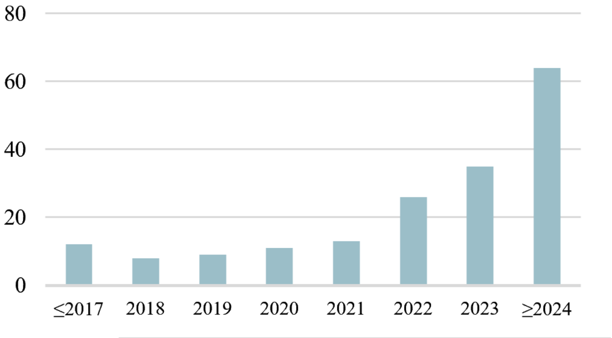
🙂 normal 🤔 challenging 🤯 hard

Experiences from other fields may not be as effective in IR 😦

How can we tailor solutions for robustness issues in IR?

Publications dedicated to addressing robustness issues in IR





The data statistics cover up to February 20, 2025.

Scan them!

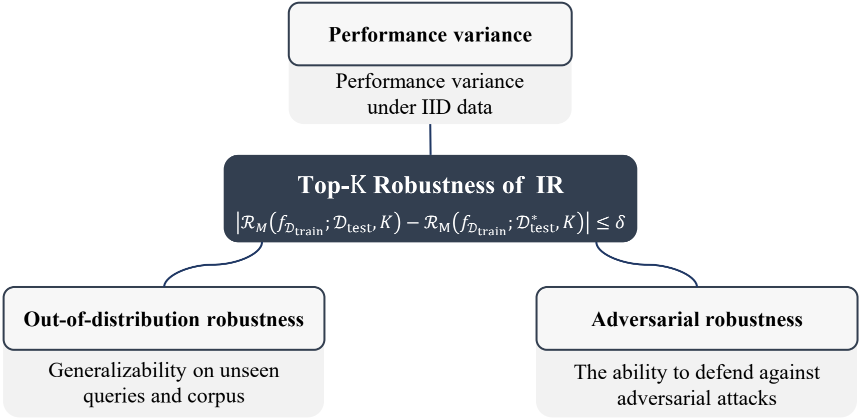
All about robust information retrieval

Our survey

Paper list

Benchmark

Our survey on robust neural information retrieval [Liu et al., 2024], is now available!



"Robust Neural Information Retrieval: An Adversarial and Out-of-distribution Perspective". [Liu et al., 2024]

In this tutorial, we pay special attention to two frequently studied types of robustness, i.e., adversarial robustness and OOD robustness

- We will cover key developments in robust information retrieval (mostly 2020–2025)
  - **Definition and taxonomy of robustness in IR**
  - **Adversarial robustness**
  - **Out-of-distribution robustness**
  - **Robust IR in the age of LLMs**

Goals of the tutorial

- We will cover key developments in robust information retrieval (mostly 2020–2025)
  - **Definition and taxonomy of robustness in IR**
  - **Adversarial robustness**
  - **Out-of-distribution robustness**
  - **Robust IR in the age of LLMs**
- Through this tutorial, we hope to . . .
  - Draw attention to the important topic of robustness in IR
  - Help interested beginners to get started and more experienced researchers to gain a systematic understanding of this field
  - Share our perspectives on **future directions**

Schedule

Time	Section	Presenter
01:30-01:50 PM	Section 1: Introduction	Maarten
01:50-02:10 PM	Section 2: Preliminaries	Yu-An
02:10-03:00 PM	Section 3: Adversarial robustness	Yu-An

30min coffee break

Time	Section	Presenter
03:30-04:20 PM	Section 4: Out-of-distribution robustness	Yu-An
04:20-04:30 PM	Section 5: Robust IR in the age of LLMs	Yu-An
04:30-04:50 PM	Section 6: Conclusions and future directions	Yu-An
04:50-05:00 PM	Q & A	All

References

Z. Dai and J. Callan. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv preprint arXiv:1910.10687*, 2019.

D. Lee, S.-w. Hwang, K. Lee, S. Choi, and S. Park. On complementarity objectives for hybrid retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13357–13368, 2023.

Y.-A. Liu, R. Zhang, J. Guo, M. de Rijke, Y. Fan, and X. Cheng. Robust neural information retrieval: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2407.06992*, 2024.

X. Ma, J. Guo, R. Zhang, Y. Fan, Y. Li, and X. Cheng. B-prop: Bootstrapped pre-training with representative words prediction for ad-hoc retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1513–1522, 2021.

L. Su, J. Guo, Y. Fan, Y. Lan, and X. Cheng. Controlling risk of web question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124, 2019.

C. Wu, R. Zhang, J. Guo, Y. Fan, and X. Cheng. Are neural ranking models robust? *ACM Transactions on Information Systems*, 41(2):1–36, 2022.

H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.