

DAN LI

EMAIL: d.li@elsevier.com GOOGLE SCHOLAR: [Dan Li](#) LINKEDIN: [Dan Li](#)

RESEARCH INTERESTS

My major research field is Information Retrieval (IR) and Text Mining (TM). I am also interested in Natural Language Processing (NLP) and general Artificial Intelligence (AI). The following are topics I have been working on:

- **TM:** extreme multi-label classification, and its application in scientific texts and patent texts
- **Semantic IR:** dense retrieval, conversational search, question answering
- **IR evaluation:** test collection construction, high recall, technology-assisted review, crowdsourcing label denoising
- **IR&NLP applications:** painting generation for classical Chinese poems
- **ML theories:** machine learning, deep learning, language models, probabilistic graphical models, Gaussian process models, Bayesian optimization

EMPLOYMENT

Data Science, Research Content Operations, Elsevier <i>Data scientist</i>	Amsterdam <i>March 2022 – now</i>
<ul style="list-style-type: none">• Research: conducting research on IR and NLP, currently focusing on dense retrieval and extreme multi-label classification• Modelling: applying state-of-the-art models and developing novel models in IR and NLP to support Elsevier's information services• Application: work with product teams to apply IR and NLP models to Elsevier's information services such as Topic Pages and Engineering Village	
IRLab, University of Amsterdam <i>ELLIS Postdoc</i>	Amsterdam <i>January 2021 – January 2022</i>
<ul style="list-style-type: none">• Research: working with Prof. Dr. Maarten de Rijke on information retrieval• Management: working with Prof. Dr. Max Welling on the scientific management of ELLS Amsterdam	
IRLab, University of Amsterdam <i>PhD student</i>	Amsterdam <i>October 2016 – October 2020</i>
<ul style="list-style-type: none">• Supervisor & Promoter: Prof. Dr. Evangelos Kanoulas• Thesis: Effective Collection Construction for Information Retrieval Evaluation and Optimization	
Huawei Technologies Co., Ltd <i>Assistant software engineer</i>	Beijing <i>July 2011 – July 2013</i>
<ul style="list-style-type: none">• Role: developing driver software for industry-level routers	

EDUCATION

University of Amsterdam (55-th in QS 2022 universities) <i>PhD, Computer Science. Supervisor: Prof. Dr. Evangelos Kanoulas</i>	Amsterdam <i>October 2016 – October 2020</i>
Tsinghua University (17-th in QS 2022 universities) <i>Research exchange. Mentor: Prof. Dr. Yiqun Liu</i>	Beijing <i>February 2018 – March 2018</i>
Dalian University of Technology (27-th in Shang Ranking 2022 universities) <i>M.A., Linguistics and Applied Linguistics. Supervisor: Prof. Dr. Jingxiang Cao</i>	Dalian, China <i>September 2013 – July 2016</i>
Dalian University of Technology <i>B.Sc., Mathematics and Applied Mathematics</i>	Dalian, China <i>September 2007 – July 2011</i>

SELECTED PUBLICATIONS

Preprint	
<ul style="list-style-type: none">• Li D., Wang S., Zou J., Tian C., Nieuwburg E., Sun F., Kanoulas E. (2021). Paint4Poem: A dataset for artistic visualization of classical Chinese poems. <i>ArXiv preprint</i>.	
Conferences/Journals	

- Li D., Yadav V., Afzal Z., Tsatsaronis G. (2022). Unsupervised Dense Retrieval for Scientific Articles. Industry track of *EMNLP 2022*.
- Li D., Ren Z., Kanoulas E. (2021). [CrowdGP: A Gaussian Process model for inferring relevance from crowd annotations](#). *WWW 2021*.
- Li D. (2021). [Effective collection construction for information retrieval evaluation and optimization](#). *ACM SIGIR Forum*. **PhD dissertation**.
- Voskarides N., Li D., Ren P., Kanoulas E., de Rijke M. (2020). [Query resolution for conversational search with limited supervision](#). *SIGIR 2020*.
- Li D., Zafeiriadis P., Kanoulas E. (2020). [APS: An active PubMed search system for technology assisted reviews](#). *SIGIR 2020*.
- Li D., Kanoulas E. (2020). [When to stop reviewing in technology-assisted reviews](#). *ACM Transactions on Information Systems (TOIS)*.
- Zou J., Li D., Kanoulas E. (2018). [Technology assisted reviews: Finding the last few relevant documents by asking yes/no questions to reviewers](#). *SIGIR 2018*.
- Inel O., Haralabopoulos G., Li D., Van Gysel C., Szilávik Z., Simperl E., Aroyo L. (2018). [Studying topical relevance with evidence-based crowdsourcing](#). *CIKM 2018*.
- Li D., Kanoulas E. (2018). [Bayesian optimization for optimizing retrieval systems](#). *WSDM 2018*.
- Zheng Y., Li D., Fan Z., Liu Y., Zhang M., Ma S. T-Reader: A multi-task deep reading comprehension model with [self-attention mechanism](#). *Journal of Chinese Information Processing*.
- Li D., Kanoulas E. (2017). [Active sampling for large-scale information retrieval evaluation](#). *CIKM 2017*.

Evaluation forums

- Kanoulas E., Li D., Azzopardi L., Spijker R. (2019). [CLEF 2019 technology assisted reviews in empirical medicine overview](#). *CLEF (Working Notes) 2019*.
- Kanoulas E., Li D., Azzopardi L., Spijker R. (2018). [CLEF 2018 technology assisted reviews in empirical medicine overview](#). *CLEF (Working Notes) 2018*.
- Kanoulas E., Li D., Azzopardi L., Spijker R. (2017). [CLEF 2017 technology assisted reviews in empirical medicine overview](#). *CLEF (Working Notes) 2017*.
- Allan J., Harman D., Kanoulas E., Li D., Van Gysel C., Voorhees E. M. (2017). [TREC 2017 common core track overview](#). *TREC 2017*.

TEACHING EXPERIENCE

Master thesis supervision

- [A comparative study of text to image generation methods for visualizing classical Chinese poems](#) 2022
Zeyou Niu, Msc Artificial Intelligence
- [Automatic optimization techniques in machine learning pipelines](#) 2021
Simon Appelt, Msc Artificial Intelligence
- [Modelling task and worker correlation for crowdsourcing label aggregation](#) 2020
Ioanna Sanida, Msc Artificial Intelligence
- [Statistical question classification](#) 2019
Ruben Halfhide, Msc Data Science

Bachelor thesis supervision

- [Building a dataset for the visualization of classical Chinese poems](#) 2020
Elisha A. Nieuwburg, Bsc Artificial Intelligence
- [De-noise large-scale poem-image pairs for poem-to-image generation](#) 2020
Fengyuan Sun, Bsc Artificial Intelligence **Cum laude (outstanding) bachelor thesis**
- [A representation of classical Chinese poetry for poem based image generation](#) 2020
River Vaudrin, Bsc Artificial Intelligence
- [Image generation for classical Chinese poems](#) 2020
Nina M. van Liebergen, Bsc Artificial Intelligence
- [Semantic visualization of classical Chinese poetry](#) 2020
Silvan Murre, Bsc Artificial Intelligence

Teaching assistant

- [AI Master Thesis Coaching](#) 2019
Master course
- [Text Retrieval and Mining](#) 2018
Master course
- [Data Mining](#) 2017
bachelor course

ACADEMIC ACTIVITIES

- Talks
- [When to Stop Reviewing in Technology-assisted Reviews](#) Online
In SIGIR 2021
 - [CrowdGP: a Gaussian Process Model for Inferring Relevance from Crowd Annotations](#) Online
In WWW 2021
 - [APS: An Active PubMed Search System for Technology Assisted Review](#) Online
In SIGIR 2020
 - [Bayesian Optimization for Optimizing Retrieval Systems](#) Marina Del Rey
In WSDM 2018
 - [Active Sampling for Large-scale Information Retrieval Evaluation](#) Singapore
In CIKM 2017
- Organisation
- [Technologically Assisted Reviews in Empirical Medicine 2017, 2018, 2019 \(CLEF TAR\)](#) Dublin, Avignon, Lugano
Co-organisation with Evangelos Kanoulas, Rene Spijker, and Leif Azzopardi
 - Goal: CLEF TAR aims to evaluate high recall approaches for IR in medical domain.
 - Role: Tasks include constructing the datasets, running evaluation scripts, writing part of the worknote papers.
- Participating challenges
- [TREC Conversational Assistance Track 2019 \(TREC CAST\)](#) Gaithersburg
Co-participated with Nikos Voskarides, Pengjie Ren, Andreas Panteli from UvA IRLab
 - Role: We proposed a BERT-based model to resolve questions and to improve re-ranking performance for conversational search systems. Our best model ranked 4 among 41 runs. See [the report](#).
 - [Chinese Machine Reading Comprehension Challenge 2018](#) Beijing
Co-participated in the challenge with Yukun Zheng and Zhen Fan from Tsinghua University
 - Role: We proposed a neural machine reading comprehension model and published the work as [a journal paper](#).
 - [TREC Core Track 2017](#) Gaithersburg
Co-participated with Christophe Van Gysel and Evangelos Kanoulas from UvA IRLab
 - Role: We built a retrieval model using Indri and optimized the model hyper-parameters using Bayesian Optimization. See [the report](#).
- Reviewing
- EACL'20
 - CIKM'18/'19/'20
 - WWW'19/'20
 - WSDM'20
 - SIGIR'19/'20/'21/'22
 - TOIS, IRJ
- PC member
- EACL'21
 - CIKM'21/'22
 - EMNLP'21/'22
 - SIGIR'22/'23

Conference service

- Program chair of CLEF'23
- proceeding chair of SIGIR'23

Fellowships

- Member of European Laboratory for Learning and Intelligent Systems ([ELLIS](#)): 2020 - now

AWARDS

- SIGIR Student Travel Grant, 2020
- CIKM Student Travel Grant, 2017
- Chinese National Scholarship for Graduate Students, top 1%, 2015

SKILLS

- Coding: Python, C, Java, Latex
- Machine Learning and Deep Learning tools: Scikit-learn, Pytorch, Tensorflow, Huggingface Transformer, Sentence Transformer, GPflow
- Language: Chinese (mother tongue), English (working language), Japanese (JLPT-N1 certificate), Dutch (Inburgering certificate), Thai (basic speaking and reading)