Bridging Gaps with AI: From Web to Healthcare Information Systems at York University

JIMMY X. HUANG AND AMRAN BHUIYAN
INFORMATION RETRIEVAL AND KNOWLEDGE MANAGEMENT RESEARCH LAB
YORK UNIVERSITY, TORONTO, CANADA,
{JHUANG,AMRAN}@YORKU.CA

I. CONTEXT

The Information Retrieval and Knowledge Management Lab at York University is a top-notch research group that focuses on a range of important fields. We work on finding and organizing information, like searching for people or data, especially in big datasets. Our expertise spans various areas including health informatics, analyzing text and websites, computer vision, and understanding human languages with computers. We're committed to advancing knowledge and developing practical solutions in these vital areas of research.

II. MISSION & RESEARCH FOCUS

The Research Lab for Information Retrieval and Knowledge Management at York University is committed to shaping the future of how we interact with and understand the ever-growing sea of data. Our mission is multifaceted:

- Creating statistical models and machine learning methods that make finding information efficient and accurate.
- Tackling big data challenges to pull out useful details and patterns.
- Improving how the web gives you information based on your current situation or what you need at the moment.
- Searching through medical documents to get valuable information that can help in healthcare.
- Pulling precise information from very large data sets quickly and accurately.
- Working with different languages, including Chinese, to find and share information.
- Looking at search data to make search engines work better.
- Providing data-backed tools to help healthcare providers offer customized care.

III. PROJECTS AND ACHIEVEMENTS

A. Current Projects

1) Validating the Evaluations of Large Language Models: An Opinion Piece based on a Systematic Survey: Our lab is currently engaged in a thorough examination of Large Language Models (LLMs). Given their growing popularity and their expanding role across various sectors, there is a pressing need for a more nuanced evaluation of these models. This project aims to move beyond traditional benchmarks like MMLU, PIQA, SIQA, HumanEval, and ELO, which might not fully reflect the true performance and limitations of LLMs. We are conducting a comprehensive review that scrutinizes how LLMs function in both general and specialized scenarios, paying particular attention to the variety of examples used and the types of tasks given to ensure a robust assessment of models. We're also evaluating the correlation between standard evaluation metrics and the practical usefulness of LLMs, particularly in terms of processing speed, memory usage, and accuracy. Additionally, our research is investigating the reliability of scripts used to evaluate responses generated by LLMs and examining potential biases when LLMs are utilized as part of the evaluation process. The project is geared towards charting new courses for LLM benchmarking, calling for inventive metrics, analyses of data integrity, and relevance to real-world applications.

Expected Outcome: The anticipated outcome of this project is the publication of our findings in both peer-reviewed conferences and journal papers. Our goal is to contribute meaningful insights toward the ongoing discourse on LLMs, helping to shape the development of more effective and equitable models for future research and application

2) The Evolution of Re-identification Technologies: From Early Approaches to LLM Integration: Our present work centers on tracing the evolution of Re-identification (ReID) technologies from their inception to the present day. Starting with the foundational non-deep learning methods that used basic features and matching algorithms, our study documents the evolution to sophisticated deep learning approaches and the subsequent integration of LLMs. The initial stage of ReID technologies, which relied on handcrafted features, faced significant limitations due to variable factors like changes in lighting and pose. The emergence of deep learning propelled the field into a new era where neural networks could learn intricate and representative features from data, enhancing the precision and robustness of ReID systems. Now, by integrating LLMs, ReID technologies embark on a new phase. This combination promises to enhance the identification process by bringing in a more profound understanding of context and a flexible interaction mechanism. Our comprehensive review not only addresses key developments and challenges but also considers the potential transformative impacts of current trends on the future of ReID technologies.

Expected Outcome: The expected deliverable of this in-depth study is a detailed exposition of our findings, to be shared with the academic community through peer-reviewed conferences and journal publications. Our objective is to

offer a substantive contribution to the field of ReID technologies by providing a clear understanding of its historical progression and anticipating future advancements, which may further redefine this technological domain.

B. Past Achievement

1) Leveraging Large Language Models across Diverse Domains: An Integrated Evaluation and Application Framework: The rapid advancement of artificial intelligence (AI) and LLMs like ChatGPT has opened new frontiers in computational research, spanning from natural language processing (NLP) to biomedical text analysis and information retrieval. This comprehensive collection of studies provides an in-depth evaluation of ChatGPT and other LLMs across a spectrum of computational tasks and datasets, highlighting the models' capabilities and limitations, along with novel applications. Our first study undertakes a detailed examination of ChatGPT's performance on 140 diverse NLP tasks, marking the most extensive assessment in benchmark academic datasets. We uncover strengths, weaknesses, and emergent abilities in multi-query instruction, offering valuable insights for future LLM utilization. Subsequently, we explore LLMs' prowess in the biomedical domain, presenting the first thorough comparison in this area across 26 datasets and 6 tasks. Results suggest that zero-shot LLMs can sometimes surpass state-of-the-art models in datasets with smaller training sets, indicating pretraining's significant impact on domain specialization.

Another study investigates the potential of LLMs, specifically ChatGPT and PaLM, in correcting data annotation errors within the Debatepedia dataset. From this, the study proposes a hybrid approach of rule-based sampling and LLM query regeneration for enhanced dataset quality. Additionally, we examine ChatGPT's role in revolutionizing information retrieval (IR) by identifying the model's unique contributions in the field, and the ensuing challenges and opportunities. Finally, a zero-shot comparison of ChatGPT with fine-tuned generative transformers in biomedical tasks demonstrates ChatGPT's unexpected superiority in contexts with limited annotated data. Together, these studies not only underscore the transformative potential of LLMs across various disciplines, but also chart a course for leveraging their capabilities in addressing complex computational challenges. This will pave the way for innovative applications and methodologies in AI-driven research.

2) Advancements in BERT Applications: Bridging Information Retrieval, Semantic Analysis, and Biomedical Text Mining: The transformative impact of Bidirectional Encoder Representations from Transformers (BERT) across various Natural Language Processing (NLP) challenges has prompted a surge in innovative applications, from IR to semantic analysis, to advanced uses in biomedical text mining. This project presents a holistic examination of BERT's versatility and efficacy in addressing complex NLP problems. Our first investigation surveys the application of BERT in IR, categorizing techniques into six pivotal areas, including handling long documents and integrating semantic information. It highlights BERT's advantage in encoder-based tasks over decoder-based models like LLMs, underscoring its efficiency and effectiveness. Another study introduces a probabilistic framework that integrates sentence-level semantics via BERT into pseudo-relevance feedback for query expansion, demonstrating significant improvements in query relevance and semantic consistency. Furthermore, we explore Bert-QAnet, a novel model that leverages BERT's encoding capabilities for detecting duplicate questions in Community question-answering platforms through sophisticated cross-attention mechanisms. This approach excels in utilizing semantic information at both word and sentence levels, achieving unprecedented accuracy. Lastly, our review of BERT's applications in biomedical and clinical text mining categorizes models into pretrained and fine-tuned frameworks, with each discussing contributions, datasets, and potential research directions. This comprehensive review not only underscores BERT's transformative role across diverse domains but also sets a foundation for future explorations, aiming to harness and refine BERT's capabilities for complex NLP tasks. In turn, this will foster the development of more advanced, efficient, and contextually aware NLP models.

3) Enhancing User-System Interaction: Integrative Strategies for Conversational Search and Product Discovery: The intersection of NLP technology and conversational systems has ushered in a new era of human-machine interaction, notably through Conversational Search Systems (CSS) such as chatbots and Virtual Personal Assistants like Siri, Alexa, Cortana, and Google Assistant. These systems offer immense potential for transforming the way users search for information and products online. However, they face significant challenges, particularly in interpreting ambiguous user queries and predicting intent accurately for effective IR and product discovery. This project explores innovative strategies for improving conversational interfaces in both search and e-commerce contexts. Our first study presents a comprehensive analysis of ambiguous query clarification tasks within CSS, highlighting various approaches to enhance query understanding and document retrieval. It emphasizes the critical role of disambiguating unclear queries to meet user needs effectively. In parallel, our second study introduces ConvPS, a novel model for conversational product search that leverages representation learning to integrate user, query, item, and conversation semantics. This model adopts greedy and explore-exploit strategies to refine user interactions, aiming to clarify product preferences through targeted questions. Together, these studies contribute to the development of conversational interfaces that are more intuitive, responsive, and capable of delivering personalized search experiences and product recommendations. This work lays the foundation for future advancements in conversational technology applications by addressing the challenges of query ambiguity and preference clarification.

4) Searching Beyond Traditional Probabilistic IR: Information Retrieval (IR) systems have traditionally been built on the assumption that search terms and documents are unrelated entities and that a document's relevance is determined without considering other documents. This approach can lead to redundancy and a lack of variety in

search outcomes. In recent years, our research has focused on developing new theoretical models to better understand the connections between terms and how to diversify search results. Since 2013, we've proposed several models aimed at identifying and leveraging the relationships between search terms, known as Cross Term associations, through our novel n-gram Cross Term Retrieval (CRTER) models. These models aim to enhance search accuracy and bring a richer variety to the results presented. To further support diversity in search results, particularly for biomedical searches, we have introduced survival modeling techniques. These techniques are designed to refine the ranking process, ensuring that a broader range of relevant documents is displayed. Our efforts have led to substantial improvements in probabilistic IR models, which we continue to refine to make them more robust. Our research has yielded excellent results, outperforming traditional methods and contributing significantly to industry practices, where our methods are being adopted by major tech companies. The innovative models we've developed have made a profound impact on IR, particularly within web searches and the biomedical field, and promise to enhance how we retrieve and interact with information in these domains.

5) Modeling and Mining Real-world Knowledge for Large-Scale Unstructured Text Analysis: Text analysis and the search for relevant questions in community question and answer (cQA) platforms have become critical due to the surge of user-generated content such as reviews and blogs. These analyses often face challenges stemming from the vast number of domains this data spans, making it hard to label and train models across all of them. Bridging the lexical gap between searched questions and archived questions is another obstacle in efficiently retrieving related questions.

To overcome these issues, we developed a specialized method to extract and apply domain knowledge across various sentiment analysis domains. This method uses a collaborative technique known as joint non-negative matrix factorization. One of the challenges in sentiment classification is that a model trained on data from one domain may underperform on data from another due to differences in language use. To tackle this, we introduced a novel domain adaptation method named Bi-Transferring Deep Neural Networks (BTDNNs). BTDNNs work by mapping data from the original domain to the target domain while also ensuring that data distribution remains consistent. This consistency is maintained through a linear transformation, which is further supported by a linear data reconstruction model. In the field of cQA, our focus is the lexical gap that often hinders question retrieval. To address this, we leveraged metadata from cQA pages, particularly category information, to train two innovative models, MB-NET and ME-NET. These models are adept at understanding word distribution and significantly mitigate the issue of the lexical gap. This approach improves the ability to find previously asked questions that are semantically similar to the new queries, enhancing the relevance of search results in cQA systems.

6) Context Modeling for Boosting Traditional and Neural IR: Context-based IR has attracted much attention in both academia and industry. We first proposed a time-aware kernel density estimation method to characterize the fine-grained word-level temporal relevance for Microblog search, with this work published in TKDE'18. We also proposed a context-aware topic model that mined the query topics from the pseudo-relevant contextual snippets to satisfy topic match. To effectively and efficiently apply topical information, we proposed a novel probabilistic framework TopPRF via integrating our new concept of Topic Space into pseudo-relevance feedback (PRF). Our proposed ideas demonstrate excellent results and provide a promising avenue for constructing better Topic Space based IR systems (e.g. context-aware and topic-sensitive query representations), which are capable of searching documents beyond traditional term matching.

Moreover, we investigated two new context modeling approaches to boost neural IR, motivated by the great success of deep learning in recent years. We first proposed a context-aligned RNN model (CA-RNN) that integrated a context alignment gating to enhance question answering and paraphrase identification, with this work published in AAAI'18. It was the first work to embed a context alignment gating mechanism in RNNs, marking a unique contribution. Following this, we proposed a collaborative and adversarial network (CAN) and published it in SIGIR'18. This was the first work to extract common contextual features for sentence similarity modeling by building a generator and a discriminator to play a collaborative and adversarial game for common features extraction. It was also the first attempt to explicitly model the common features by incorporating collaborative learning into the GAN framework. In addition, we proposed a positional attention mechanism to incorporate the positional context of the question words into the answers' attentive representations, as published in SIGIR'17.

7) Modeling Feature-based Medical IR: Text-based image retrieval (TBIR) has proven highly effective for finding images that are tagged with descriptive text. Within this system, each word in both the search query and the image descriptions is given equal importance. Recognizing that specific medical terms might influence the outcomes of image searches, we aimed to refine TBIR by incorporating unique medical-dependent features (MDF). These features are a combination of image properties and medical terms, which we crafted to enhance how images are matched to queries. In our previous work, we pioneered the process of categorizing search queries to improve retrieval systems. We did this by considering not just the medical terms but also generic characteristics like how specific or vague a query is. The integration of MDF into our systems helped streamline the re-ranking process, allowing us to better gauge the impact on search results. We observed a substantial gain in efficiency, cutting down on both time and computational resources. Furthermore, we employed these MDFs within a Bayesian Network-based image retrieval framework. Although building Bayesian Networks is complex and resource-intensive, the inclusion of MDF is anticipated to bring substantial benefits. Our experiments with various medical image databases have shown that our

new models substantially outperform the established baseline models in retrieving images, marking a significant advancement in the field.

8) Machine Learning (ML) and Deep Learning (DL) for IR and Healthcare Decision Making: In the research lab, we have made significant strides in the integration of ML and DL in the advancement of IR systems, as outlined in our wide range of publications. Notably, the laboratory innovated a novel feedback strategy incorporating co-training with pseudo-relevance feedback (PRF), which has proven effective in IR, in particular scenarios lacking labeled data. This method has demonstrated its value through improved IR performance.

The IRLab also formulated a learning-to-rank model with a focus on quality-aware PRF. The approaches we have pioneered for embedding ML and DL into PRF for IR have consistently delivered superior performance.

In real-world applications, this lab has applied a composite DL model (DBN+SVM) to streamline automated diagnosis and decision-making processes in healthcare We enhanced convolutional deep belief networks and achieved notable performance gains over well-established benchmarks on extensive datasets Additionally, the laboratory has effectively employed a sparse Bayesian multi-instance multi-label model for the analysis of skin biopsy images, developing a robust ML model using deep convolutional neural networks (CNNs) to classify cell images for malaria screening These innovations signify IRLab's commitment to advancing medical diagnostics through cutting-edge IR technologies.

9) Additional Information on Contribution: The lab has established a significant track record of contributions to the field of information retrieval. Under the guidance of the laboratory's director, Professor Jimmy Huang, who serves as author and co-author, the lab has produced over 320 peer-reviewed publications in the past six years. These works are often led by graduate students and postdoc fellows of the lab, with senior researchers contributing through ideation, critical feedback, algorithmic development, and, in many cases, extensive revision of manuscripts to elevate the writing to academic standards (especially for international students).

The research endeavors at this lab are strategically directed toward the most revered academic journals and conferences. These include esteemed journals such as ACM Transactions on Information Systems (TOIS), IEEE Transactions on Knowledge and Data Engineering (TKDE), and conferences like ACM SIGIR and IJCAI. The selection of these venues is intentional, with each chosen for their rigorous peer review processes and the impact their publications have on the research community.

The efforts of this lab are accelerated towards pioneering areas of IR, including task-oriented and context-sensitive frameworks, with noticeable advancements in conversational search technologies and models like ChatGPT and BERT. These initiatives are part of the lab's NSERC Discovery Grant-sponsored research and have garnered recognition, including accolades such as best paper awards. For instance, the paper titled "Hypergraph contrastive collaborative filtering" presented at ACM SIGIR 2022 was celebrated as one of the most influential papers in the subsequent year.

Leadership within this lab extends to significant roles in organizing major conferences, evidencing the lab's commitment to fostering scholarly exchange and progress in the field. For instance, the director of this lab, Professor Jimmy Huang, was the General Conference Chair for the 19th International ACM CIKM Conference, the Program Chair for IEEE/ACM International Joint Conferences on Web Intelligence & Intelligent Agent Technology in 2010, and the General Conference Chair for the 43rd International ACM SIGIR Conference in 2020. He has been the Chair of the IEEE Technical Committee on Intelligent Informatics (TCII) since 2023.

The research from our lab has been enormously significant, reaching beyond academia to influence major companies like Google, Microsoft, Baidu, IBM, eBay, and Dapasoft, all of whom have used our findings to enhance their products. We've also formed strong bonds with organizations such as Southlake Regional Health Centre, Institute for Clinical Evaluative Sciences, Dapasoft, IBM, Google, Microsoft, Baidu, Manifold Data Mining, National Institute of Standards and Technology, and the Ontario Ministry of Health, further building a bridge between our research and real-world applications.

IV. Collaborations and Partnerships

We are always interested in collaborations with partners who have interests that intersect with ours. Current and past collaborating organizations are listed below:

- OpenText Corporation
- Institute for Clinical Evaluative Sciences (ICES)
- MRC-CRC (Medical Research Council - Clinical Research Centre)
- IBM
- Microsoft Research
- Yahoo!
- City University London
- InfoBright
- Scuola Superiore Sant'Anna

V. Funding & Acknowledgements

The lab is currently supported by the Ontario Ministry of Research & Innovation, Natural Sciences and Engineering Research Council, NSERC Research Tools and Instruments Grant, Tri-Agency (SSHRC, CIHR and NSERC) Syntheses Grant, AlphaGlobal iT, CRD Grant, IBM, ICES, CIHR, SSHRC, Petro Canada, SHARCNET, VPRI, CGA-Canada/CAAA, IRF, MITACS, Atkinson and York University.

The following agencies and organizations are gratefully acknowledged for funding our current and past research activities:

- Ontario Ministry of Research & Innovation
- AT&T
- Collaborative Research and Development (CRD)
- Institute for Clinical Evaluative Sciences (ICES)
- Certified General Accountants Association of Canada (CGA)
- IBM
- OpenText Corporation
- Petro-Canada

Fig. 1. Team Members.

- NSERC (Natural Sciences and Engineering Research Council of Canada)
- York University
- SSHRC (Social Sciences and Humanities Research Council of Canada)
- CIHR (Canadian Institutes of Health Research)

VI. LOOKING TO THE FUTURE

Our vision is to establish a center of excellence for research that transcends traditional boundaries and serves as a catalyst for transformative applications in IR and knowledge management. We are dedicated to:

- Innovation in the field of person re-identification and retrieval by enhancing the capability to recognize and distinguish individuals within data systems to improve security, customer experience, and personalization.
- Exploring search tasks to understand how users engage with IR systems, aiming to streamline the search process and deliver results that align closely with user intentions.
- Conducting groundbreaking research on LLMs that explores their potential to enhance the interpretability and responsiveness of search systems, thereby improving how machines understand and process human language.

VII. REPRESENTATIVE PUBLICATIONS

- Wang, J., **Huang, J. X.**, Tu, X., Wang, J., Huang, A.J., MD Tahmid Rahman Laskar and Amran Bhuiyan. "Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges" *ACM Computing Surveys (CSUR)*. ACM Publisher. Accepted on February 09, 2024. ISSN: 0360-0300 and EISSN:1557-7341 (ACM Computing Surveys is a quarterly peer-reviewed scientific journal published by the Association for Computing Machinery. It receives the highest rank "A*", and publishes survey articles and tutorials related to computer science and computing. I am the contact author of this paper. Jiajia Wang, Junmei Wang, Angela Huang and MD Tahmid Rahman Laskar was my PhD student working in my research lab. Amran Bhuiyan is my postdoc under my supervision in my research lab. This research was supported by my NSERC DG grant). ISI Journal Impact Factor: 10.282 (2020), ranking it 4 out of 137; 14.324 (2021); 16.6 (2022), ranking 3 out of 109 in Computer Science Theory & Methods.
- Keyvan, K. and **Huang, J. X.** "How to Approach Ambiguous Queries in Conversational Search? A Survey of Techniques, Approaches, Tools and Challenges" *ACM Computing Surveys (CSUR)*. ACM Publisher. Vol. 55, No. 6, Article 129, 1-40 pages. June 2023. ISSN: 0360-0300 and EISSN:1557-7341 (ACM Computing Surveys is a quarterly peer-reviewed scientific journal published by the Association for Computing Machinery. It receives the highest rank "A*", and publishes survey articles and tutorials related to computer science and computing. Kimiya Keyvan was a graduate student working in my research lab. This research was supported by my NSERC DG grant). ISI Journal Impact Factor: 10.282 (2020), ranking it 4 out of 137; 14.324 (2021); 16.6 (2022), ranking 3 out of 109 in Computer Science Theory & Methods.
- Laskar, M.T.R., Bari, M.S., Rahman, M., Bhuiyan, M.A.H., Joty, S. and **Huang, J.X.** "A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets" (full paper), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Findings of ACL 2023)*, Toronto, Ontario, Canada. July 9-14, 2023 (**ACL** is the top-tier conference in the fields of NLP and Computational Linguistics and I am the contact author).
- Zou, J., **Huang, J. X.**, Ren, Z. and Kanoulas, E. "Learning to Ask: Conversational Product Search via Representation Learning", *ACM Transactions on Information Systems (TOIS)*. ACM Publisher. Accepted on July 23, 2023. ISSN: 1046-8188 and EISSN:1558-2868 (ACM TOIS is a top-tier journal

in information systems. J. Zou was a visiting PhD student working in my research lab since July 2020. This research was supported by my NSERC DG grant). ISI Journal Impact Factor: 4.797 (2020).

- Zhao, J., **Huang, J. X.**, Deng, H, Chang, Y. and Xia, L. "Are Topics Interesting or Not? An LDA-based Topic-graph Probabilistic Model for Web Search Personalization" *ACM Transactions on Information Systems (TOIS)*. ACM Publisher. Vol. 40, No. 3, Article 51, 1-24 pages. July 2022. ISSN: 1046-8188 and EISSN:1558-2868 (ACM TOIS is a top-tier journal in information systems. J. Zhao was a former PhD student in my research lab and L. Xia is a postdoc fellow working in my research lab. This research was supported by my NSERC DG grant). ISI Journal Impact Factor: 4.797 (2020).
- Xia, L., Huang, C., Xu, Y., Zhao, J., Yin, D. and **Huang, J. X.** "Hypergraph Contrastive Collaborative Filtering" (full paper), *Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022)*, Madrid, Spain. July 11-15, 2022. 749-758 (20.27% acceptance rate: 161 regular research papers accepted out of 1,469 submissions which include 794 valid full paper and 675 valid short paper submissions conference in the fields of Information Retrieval and Web search. This research was supported in part by my NSERC DG grant). This paper was ranked as the **Most Influential SIGIR Papers (2023-04)** by Best Paper Digest in New York (https://www.paperdigest.org/2023/04/most-influential-sigir-papers-2023-04/).
- Tan, X. and **Huang, J. X.** "A Complexity-theoretic Analysis of Green Pickup-and-Delivery Problems" (full research paper and invited for oral presentation), *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021)*, Vancouver, Canada. February 2-9, 2021. 11990-11997 (21.0% acceptance rate: 1,692 papers accepted out of a record 9,034 submissions. **AAAI** [1] is the top-tier and premier interdisciplinary conference in the field of artificial intelligence. X. Tan was a postdoc fellow working in my research lab. This research is supported by my NSERC DG grant).
- Tal, O., Liu, Y., **Huang, J. X.**., Yu, X. and Aljbaw, B. "Neural Attention Frameworks for Explainable Recommendation", *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. IEEE Publisher. Vol 33, No 5, pp. 2137-2150, 2021. ISSN: 1041-4347 (IEEE TKDE is a top tier journal in data mining. I am the contact author of this paper. O. Tal was graduate student and Y. Liu was my former Ph.D. student. This research was supported by NSERC DG grant). ISI Journal Impact Factor: 3.438 (2017); 3.857 (2018); 4.561 (2019); 4.935 (2020).
- Huang, C., Chen, J., Xia, L., Dai, P., Chen, Y., Bo, L, Zhao, J. and **Huang, J. X.** "Graph-Enhanced Multi-Task Learning of Multi-Level Transition Dynamics for Session-based Recommendation" (full research paper and invited for oral presentation), *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021)*, Vancouver, Canada. February 2-9, 2021. 4123-4130 (21.0% acceptance rate: 1,692 papers accepted out of a record 9,034 submissions. C. Huang was a PhD student and J. Zhao was a former PhD student in my research lab. This research is supported in part by my NSERC DG grant).
- Chen, X., Xiong, K., Zhang, Y., Xia, L., Yin, D. and **Huang, J. X.** "Neural Feature-aware Recommendation with Signed Hypergraph Convolutional Network", *ACM Transactions on Information Systems (TOIS)*. ACM Publisher. Vol 39, No. 1, Article 8, 22 pages. November 2020. ISSN: 1046-8188 and EISSN:1558-2868 (ACM TOIS is a top tier journal in information systems. L. Xia is a postdoc fellow working in my research lab. This research was supported by my NSERC DG grant). ISI Journal Impact Factor: 4.797 (2020).
- Zhou, L., Xia, L., Gu, Y., Liu, W., **Huang, J. X.** and Yin, D. "Neural Interactive Collaborative Filtering" (full paper), *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, Xi'an, China. July 25-30, 2020. 749-758 (26.49% acceptance rate: 147 regular papers accepted out of 1,180 submissions which include 555 long papers and 512 short papers etc. **ACM SIGIR** is the top tier conference in the fields of Information Retrieval and Web search. L. Xia is a postdoc fellow working in my research lab. This research was supported by my NSERC DG grant).
- Wang, P., Fan, Y., Xia, L., Zhao, W. X., Niu, S. and **Huang, J. X.** "KERL: A Knowledge-Guided Reinforcement Learning Model for Sequential Recommendation" (full paper), *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, Xi'an, China. July 25-30, 2020. 209-218 (26.49% acceptance rate: 147 regular papers accepted out of 1,180 submissions which include 555 long papers and 512 short papers etc. **ACM SIGIR** is the top tier conference in the fields of Information Retrieval and Web search. L. Xia is a postdoc fellow working in my research lab. This research was supported by my NSERC DG grant).
- Xie, Z., Zhou, G.Y., Liu, J. and **Huang, J.X.** "ReInceptionE: Relation-Aware Inception Network with Joint Local-Global Structural Information for

[1] AAAI stands for the Association for the Advancement of Artificial Intelligence (formerly the American Association for Artificial Intelligence)

Knowledge Graph Embedding" (full paper and invited for oral presentation), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Seattle, Washington, USA. July 5-10, 2020. 5929-5939 (25.4% long paper acceptance rate for presentation: 571 regular full papers accepted for presentation at ACL out of 3,429 paper submissions with 2,244 for long papers and 1,185 for short papers. **ACL** is the top tier conference in the fields of Computational Linguistics and NLP).

- Tan, X. and **Huang, J. X.** "On Computational Complexity of Pickup-and-Delivery Problems with Precedence Constraints or Time Windows" (full research paper and invited for oral presentation), *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pp. 5635-5643. Macao, China. August 10-16, 2019 (13.7% acceptance rate: 650 papers accepted out of 4752 paper submissions. **IJCAI** is a premier international conference for AI researchers and practitioners and IJCAIs were held biennially in odd-numbered years since 1969. X. Tan was a postdoctoral fellow from 2016 to 2019 in my research lab. This research is supported by my NSERC DG grant & CREATE award).
- Chen, Q., Hu, Q. and **Huang, J. X.**. "TAKer: Time-Aware Microblog Search with Kernel Density Estimation", *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. IEEE Publisher. Vol 30, No 8, pp. 1602-1615, 2018. ISSN: 1041-4347 (IEEE TKDE is a top tier journal in data mining. I am the contact author of this paper. Q. Chen was a visiting Ph.D. student in my research lab. Q. Hu was my former Ph.D. student. This research was supported by my NSERC DG grant & CREATE award). ISI Journal Impact Factor: 3.438 (2017).
- **Huang, J. X.**, He, B. and Zhao, J. "Mining Authoritative and Topical Evidence from Blogosphere for Improving Opinion Retrieval", *Information Systems (IS)*. Vol 78, pp. 199-213, 2018. ISSN: 0306-4379 ELSEVIER Publisher. In press (IS is a leading peer-reviewed journal covering data-intensive technologies underlying database systems, business processes, social media, and data science. I am the contact author of this paper. B. He was a postdoctoral fellow and J. Zhao was my Ph.D. student under my supervision and completed her Ph.D. in March 2015. This research was supported by my NSERC DG grant and NSERC CREATE Award). Impact Factor: 2.777 (2016) and 5-Year Impact Factor: 2.822 (2016).
- Chen, Q., Hu, Q. and **Huang, J. X.** "CAN: Enhancing Sentence Similarity Modeling with Collaborative and Adversarial Network" (full paper), *Proceedings of the 41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018)*, Ann Arbor, Michigan, U.S.A. July 8-12, 2018 (21% acceptance rate: 86 regular papers accepted out of 409 long paper and 327 short paper submissions. **ACM SIGIR** is the top tier conference in the field of Information Retrieval. Q. Chen was a visiting Ph.D. student in my research lab. Q. Hu was my former Ph.D. student. This research was supported by my NSERC DG grant & CREATE award).
- Chen, Q., Hu, Q. and **Huang, J. X.** "CA-RNN: Using Context-Aligned Recurrent Neural Networks for Modeling Sentence Similarity" (full research paper and invited for oral presentation), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, New Orleans, Louisiana, USA. February 2-7, 2018 (24.5% acceptance rate: 933 papers accepted out of over 3800 paper submissions. **AAAI** is the top-tier and premier interdisciplinary conference in the field of artifical intelligence. Q. Chen was a visiting PhD student in my research lab from November 17, 2017 to March 17, 2018. This research is supported by my NSERC DG grant & CREATE award).
- Gan, G. and **Huang, J. X.** "A Data Mining Framework for Valuing Large Portfolios of Variable Annuities" (full paper and invited for oral presentation), *Proceedings of the 23rd ACM SIGKDD International Conference of Knowledge, Discovery and Data Mining (KDD 2017)*, Halifax, Nova Scotia, Canada. August 13-17, 2017 (8.65% acceptance rate for oral presentation: 99 accepted full papers with oral presentation in which 35 come from data science track and 64 come from research track, and 10.23% acceptance rate for posters: 117 accepted posters in which 50 come from data science track and 67 come from research track out of 1,144 full paper submissions where 396 were submitted from data science track and 748 were submitted from research track. **ACM SIGKDD** is the top-tier and premier interdisciplinary conference bringing together researchers and practitioners from data science, data mining, knowledge discovery, large-scale data analytics, and big data. This research is supported by my NSERC DG grant).
- Zhou, G. and **Huang, J. X.** "Modeling and Mining Domain Shared Knowledge for Sentiment Analysis", *ACM Transactions on Information Systems (TOIS)*. ACM Publisher. Vol 36, No. 2, Article 18, 40 pages. August 2017. ISSN: 1046-8188 and EISSN:1558-2868 (ACM TOIS is a top tier journal in information systems. I am the contact author of this paper. G. Zhou was a visiting professor in my research lab. This research was supported by my NSERC DG grant & CREATE award).
- Zhou, G. and **Huang, J. X.** "Modeling and Learning Distributed Word Representation with Metadata for Question Retrieval", *IEEE Transactions on Knowledge and*

Data Engineering (TKDE). IEEE Publisher. Vol 29, No. 6, 1226-1239, 2017 (IEEE TKDE is a top tier journal in data mining. I am the contact author of this paper. G. Zhou was a visiting professor in my research lab from Oct. 2016 to Oct. 2017. This research was supported by my NSERC DG grant & CREATE award). ISI Journal Impact Factor: 2.285.

- Miao, J., **Huang, J. X.** and Zhao, J. "TopPRF: A Probabilistic Framework for Integrating Topic Space into Pseudo Relevance Feedback" [2], *ACM Transactions on Information Systems (TOIS)*. ACM Publisher. Vol 34, No. 4, Article 22, 38 pages. September 2016. ISSN: 1046-8188 and EISSN:1558-2868 (TOIS is a top tier journal in information systems. I am the contact author of this paper. J. Miao and J. Zhao are my PhD student from August 2009 to June 2016 and August 2008 to April 2015 respectively under my supervision. This research has been supported by my NSERC DG grant & ERA award and this paper is a part of James Miao 's PhD thesis).
- Zhou, G.Y., **Huang, J. X.** and Zhao, J. "Bi-Transferring Deep Neural Networks for Domain Adaptation" (full paper and invited for oral presentation), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany. August 7-12, 2016. 322-332 (12.9% long paper acceptance rate for oral presentation: 231 regular full papers accepted for presentation at ACL — 116 as oral presentations and 115 as poster presentations out of 897 long paper and 463 short paper submissions. **ACL** is the top tier conference in the fields of Computational Linguistics and NLP).
- Zhao, J., **Huang, J. X.** and Ye, Z. "Modeling Term Associations for Probabilistic Information Retrieval" [3], *ACM Transactions on Information Systems (TOIS)*. ACM Publisher. Vol 32, No. 2, Article 7, 47 pages. April 2014. ISSN: 1046-8188 and EISSN:1558-2868 (TOIS is a top tier journal in information systems. I am the contact author of this paper. J. Zhao and Z. Ye are my PhD student from August 2008 to July 2013 and PDF since November 1, 2011 respectively under my supervision. This research has been supported by my NSERC DG grant & ERA award and this paper is a part of Jessie Zhao's PhD thesis).
- Ye, Z. and **Huang, J. X.** "A Simple Term Frequency Transformation Model for Effective Pseudo Relevance Feedback" (full paper), *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*, Gold Coast, Australia. July 6-11, 2014 (21% acceptance rate: 82 regular papers accepted out of 387 long paper and 275 short paper submissions. **ACM SIGIR** is the top tier conference in the field of Information Retrieval. Z. Ye is my postdoctoral fellow starting from January 2012. This research is supported by my NSERC DG grant).
- Yin, X, **Huang, J. X.**, Li, Z. and Zhou, X. "Survival Modeling Approach to Biomedical Search Result Diversification Using Wikipedia" (14 pages), *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. Vol 25, No 6, pp. 1201-1212, June 2013. IEEE Publisher (TKDE is a top tier journal in data mining. I am the contact author of this paper. X. Yin was my PhD student from April 2008 to April 2010. This research has been supported by my NSERC DG grant & ERA award and this paper is a part of her PhD thesis. X. Zhou is my Ph.D. student in Math & Statistics since September 2008). ISI Journal Impact Factor: 2.285. DOI Bookmark:
- **Huang, X.**, Miao, J. and He, B. "High Performance Query Expansion Using Adaptive Co-training", *Information Processing & Management: An International Journal (IPM)*. 49(2):441-453, 2013. ELSEVIER Publisher (IPM is a top tier journal in information retrieval, information systems and information management. I am the contact author of this paper. My contribution to this paper is more than 80%. J. Miao is my PhD student and B. He was my PDF). ISSN: 0957-4174. ISI Journal Impact Factor: 1.783
- Yu, X, Liu, Y., **Huang, X.** and An, A. "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. 24(4): 720-734, April 2012. IEEE Publisher (TKDE is a top tier journal in data mining. Y. Liu is my PhD student and this paper published is a part of her PhD thesis). ISI Journal Impact Factor: 2.285
- He, B, **Huang, X.** and Zhou, X. "Modeling Term Proximity for Probabilistic Information Retrieval Models", *Information Sciences Journal*. 181(14): 3017-3031, July 15, 2011. Elsevier Publisher (Information Sciences Journal is a top tier journal in information Science. I am the contact author of this paper. B. He was my postdoctoral fellow from March 2009 to July 2010 and X. Zhou is currently my PhD student. This research was supported by my ERA award and NSERC DG grant). ISSN: 0020-0255. ISI Journal Impact Factor: 3.291 (2010)[4].
- Zhu, J., **Huang, X.**, Song, D. and Ruger, S. "Integrating Multiple Document Features in Language Models for Expert Finding", *Knowledge and Information Systems: An International Journal (KAIS)*. 23(1): 29-54, 2010.

[2]Top 1 downloaded article for ACM TOIS (past 6 weeks) on October 17, 2016.
[3]Top 1 downloaded article for ACM TOIS (past 6 weeks) on June 9, 2014.
[4]https://journalinsights.elsevier.com/journals/0020-0255/impact_factor

Springer-Verlag Publisher, January 2010 (KAIS is one of the most esteemed journals in knowledge systems, data mining and advanced information systems. My contribution to this paper is 35%.). ISSN: 0219-1377 (Print) and ISSN (Online): 0219-3116. ISI Journal Impact Factor: 2.211

- **Huang, X.** and Hu, Q. "A Bayesian Learning Approach to Promoting Diversity in Ranking for Biomedical Information Retrieval" (full paper), *Proceedings of the 32nd Annual International* **ACM SIGIR** *Conference on Research and Development in Information Retrieval*, Boston, USA, July 19-23, 2009 (15.8% acceptance rate: 78 regular papers accepted out of 494 submissions. **ACM SIGIR** is the best conference in the field of Information Retrieval. Q. Hu is my PhD student).
- Yang Liu, **Xiangji Huang**, Aijun An, and Xiaohui Yu "Modeling and Predicting the Helpfulness of Online Reviews" (full paper), *Proceedings of the 2008 IEEE International Conference on Data Mining*, Pisa, Italy, December 15-19, 2008. (9.7% acceptance rate: 70 regular papers accepted out of 724 submissions. **IEEE ICDM** is the best conference in the field of Data Mining. Y.Liu was my PhD student).
- Liu, Y., **Huang**, X., An, A. and Yu, X. "ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs" (full paper), *Proceedings of the 30th Annual International* **ACM SIGIR** *Conference on Research and Development in Information Retrieval*, Amsterdam, July 23-27, 2007. (17.5% acceptance rate: 86 regular papers accepted out of 491 submissions. **ACM SIGIR** is the best conference in the field of Information Retrieval. Y. Liu is my PhD student).
- **Huang**, X., Huang, Y., Wen, M., An, A., Liu, Y. and Poon, J. "Applying Data Mining to Pseudo-Relevance Feedback for High Performance Text Retrieval", *Proceedings of the 2006 IEEE International Conference on Data Mining*, Convention and Exhibition Centre, Hong Kong, December 18-22, 2006. (9.4% acceptance rate: 73 regular papers accepted out of 776 submissions. **IEEE ICDM** is the best conference in the field of Data Mining. Y.Huang, M.Wen and Y.Liu are my graduate students).
- **Huang, X.**, Peng, F., An, A., Schuurmans, D. "Dynamic Web Log Session Identification with Statistical Language Models", *Journal of the American Society for Information Science and Technology (JASIST)*, Special Issue on Webometrics, 55(14):1290-1303, 2004. John Wiley & Sons, Inc. (JASIST is the most prestigious journal in information science and technology. My contribution to this paper is 80%). ISSN (Print): 1532-2882 and ISSN (Online): 1532-2890. ISI Journal Impact Factor: 2.300
- **Huang, X.**, Peng, F., Schuurmans, D., Cercone, N. and Robertson, S.E. "Using Machine Learning for Text Segmentation in Information Retrieval", *Information Retrieval*, 6 (3-4), pp. 333-362, 2003. Kluwer Academic Publisher. (Information Retrieval is the most prestigious journal in information retrieval. My contribution to this paper is 80%). ISSN (Print): 1386-4564 and ISSN (Online): 1573-7659. ISI Journal Impact Factor: 1.841

Contact Information

Director:
Dr. Jimmy Huang, Professor
Tier I York Research Chair
School of Information Technology
York University
Toronto, Ontario
Canada M3J 1P3
E-mail: jhuang@yorku.ca

Phone: +1 416-736-2100 ext.30149
Fax: +1 416-736-5287

Website:
https://www.yorku.ca/jhuang/irlab/