FinIR 2020: The First Workshop on Information Retrieval in Finance

Fuli Feng
National University of Singapore
fulifeng93@gmail.com

Cheng Luo
MegaTech.AI
luochengleo@gmail.com

Xiangnan He
University of Science and Technology of China
xiangnanhe@gmail.com

Yiqun Liu
Tsinghua University
yiqunliu@tsinghua.edu.cn

Tat-Seng Chua
National University of Singapore
dcscts@nus.edu.sg

ABSTRACT

This half-day workshop explores challenges and potential research directions about Information Retrieval (IR) in finance. The focus will be on stimulating discussions around the accessing, searching, filtering, and analyzing financial documents in banking, insurance, and investment, such as the financial statements, analyst reports, filling forms, and news articles. We welcome theoretical, experimental, and methodological studies that aim to advance techniques of managing and understanding financial documents, as well as emphasize the applicability in practical applications. The workshop aims to bring together a diverse set of researchers and practitioners interested in investigating relevant topics. Besides, to facilitate developing and testing some relevant techniques, we hold a data challenge on quantifying analyst reports and news articles for the prediction of commodity prices.

website: http://finir2020.github.io/.

KEYWORDS

Information Retrieval, Finance, Unstructured Data

ACM Reference Format:
Fuli Feng, Cheng Luo, Xiangnan He, Yiqun Liu, and Tat-Seng Chua. 2020. FinIR 2020: The First Workshop on Information Retrieval in Finance. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401462

1 MOTIVATION AND FIT FOR SIGIR

In the last decade, digitalization and automation have empowered fast technology development in finance, enabling various financial applications such as peer-to-peer lending, online insurance, robo-advisor, and algorithmic trading. Owing to such development, various large scale financial data and documents are generated in different financial scenarios, including banking, insurance, and investment. As the volume of such unstructured financial data (*i.e.,* financial documents) continuously increases, advanced IR techniques for information accessing, searching, filtering, and understanding within these documents have become one of the essential requirements to improve the efficiency of the financial working process and the quality of financial services. For instance, information filtering techniques are largely required to alleviate the information overload in finance, which can filter out irrelevant documents for financial practitioners and thus save time and labor-cost of traversing large-scale texts. Moreover, due to the importance of timing in financial services, advanced IR techniques are required to retrieve and extract information from financial documents in real-time. Unfortunately, existing IR techniques cannot fully satisfy these requirements since lacking considering the unique properties of financial documents, for instance:

- **Number intensity:** Numbers (*e.g.,* price and date) are widely used in financial documents to express financial information. Most existing IR methods lack numeracy to recognize and understand the numbers correctly, and thus can fail to retrieve or filter the documents.
- **Heterogeneity:** In addition to textual contents, financial documents typically contain figures and tables, requiring the IR approaches to jointly modeling the non-textual contents and their surrounding text.
- **Curse of PDF:** Financial documents are commonly in PDF format, which is a barrier for applying the existing IR approaches in processing the documents.

To bridge the gap, it is vital to adapt IR techniques to be suitable for finance documents to serve the financial applications of great practical value. From a broader perspective, exploring IR technology in finance has the potential to benefit the world economy and human society. This is because finance plays an important role in the world economy and is one of the cornerstones of human society. We believe that, as one of the premier conferences on information retrieval, the SIGIR community has the responsibility to lead the IR research in finance. Towards this end, this workshop is organized to bring together a diverse set of researchers and practitioners interested in exploring IR technology in finance. To facilitate developing and testing of IR approaches, this workshop also hosts a *data challenge* with a large-scale corpus of multi-lingual financial documents.

SIGIR ’20, July 25–30, 2020, Virtual Event, China
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401462

2 THEME AND PURPOSE OF THE WORKSHOP

FinIR will be a forum for discussion about the challenges in applying IR technologies to unstructured data in finance. The purpose of this workshop is to establish a bridge for communication between industrial researchers and academic researchers, and provide an opportunity for people to present new work and early results, and discuss future directions. The themes of focus for the workshop include but not limited to the following topics:

2.1 Financial Information Filtering

Information filtering [2] aims to alleviate the information overload by only delivering to users the information that is relevant to them. Aiming to alleviate the information overload of financial practitioners, it is of vital importance to a line of relevant topics, including financial practitioner profiling, financial recommendation, financial topic distillation, and financial text quantification.

2.2 Financial Search and Ranking

Search and ranking are core research topics in the SIGIR community [8], which focused on the plain text of documents. Considering the unique properties of financial documents (mentioned in Section 1), it is an interesting setting for the scientific exploration of searching and ranking financial documents, including query analysis, retrieval models, ranking algorithms, indexing, *etc.*.

2.3 Financial Semantics

To facilitate filtering, searching, and analyzing financial documents, a fundamental research direction is to understand the semantics of financial documents. For instance,

- **Financial Representation Learning.** The document representation learning model is a crucial building block of various IR approaches. The pre-trained language model [4] is an emergent technique for document representation learning, which has led to significant performance gain in many IR tasks [3, 10, 12]. This wave of research has also aroused great interest in training language models to properly encode financial documents especially the numerical information [14].
- **Financial QA and Dialog.** Question answering and dialog system have achieved remarkable success in many fields such as e-commerce. This wave of research has also aroused great interest in finance which is of great value to save human resources and accelerate financial services.
- **Financial Knowledge Graph.** Domain knowledge that complements the data-driven models has shown great potential in various information retrieval tasks [6, 13]. Prior study [7] also demonstrates the success of domain knowledge in finance, which requires of constructing, maintaining, and utilizing knowledge graphs in finance.

2.4 Financial Applications

Research on financial applications where IR or relevant techniques are properly applied is also welcomed. For instance, we elaborate on the research relevant to numbers, which is highly related to financial applications.

- **Number Parsing.** As a crucial part of financial documents, understanding the numbers in-depth, *e.g.,* category and semantic role, is beneficial for understanding the contents in financial documents [1]. Moreover, towards the goal of comprehensively understanding documents, number parsing is a valuable and interesting research direction that complements the conventional word-oriented parsing such as dependency parsing [9].
- **Numeric Information Extraction.** Robotic process automation can bring immediate value to the core financial working processes such as accounting services [11], where accurately extracting structured information especially numerical one from financial documents (*i.e.,* information extraction) is the fundamental technology. Despite the recent development of information extraction in open domain texts [5], there is still a long way to go towards this goal.
- **Financial Table Information Accessing.** In addition to numbers embedded within texts, a large number of tables are presented in financial documents to organize and manipulate data. As such, the research explores information accessing tasks are valuable for many downstream applications, *e.g.,* financial table extraction, financial table interpretation, and financial table search [15].

3 ACTIVITIES AND TENTATIVE SCHEDULE

FinIR will be a highly interactive half-day workshop, featuring a mix of presentation and interaction formats.

FinIR will feature presentations of the following types:

(1) One invited keynote (academic/industrial)
(2) Three invited talks (academic and industrial)
(3) Six contributed talks.

WorkShop Schedule

○ **09:00-09:05** Welcome and Opening

○ **09:05-09:35** Keynote

○ **09:35-09:55** Invited Talk 1

○ **09:55-10:00** Data Challenge Introduction

○ **10:00-10:30** Contributed Talk * 3 (Winning Solutions)

○ **10:30-11:00** Coffee Break

○ **11:00-11:20** Invited Talk 2

○ **11:20-11:40** Invited Talk 3

○ **11:40-12:10** Contributed Talk * 3

○ **12:10-12:30** General Discussion and Closing

Considering that the final schedule of the SIGIR conference has not been fixed, we are still inviting speakers to give talk. We plan to invite experts with experience of IR in finance from academia (two) or industry (two).

4 SELECTION PROCESS FOR PARTICIPANTS AND/OR PRESENTERS AND MAXIMUM NUMBER OF PARTICIPANTS

We will solicit submission of papers up to four pages through an open call for papers, representing reports of original research, preliminary research results, proposals for new work, and position papers. All papers will be peer reviewed by the program committee and judged by their relevance to the workshop, especially to the topics, and their potential to generate discussion. To facilitate interaction and simulate discussion, we limit the number of participants to 40.

5 LIST OF ORGANIZERS

- Fuli Feng received his Ph.D. in the Computer Science from National University of Singapore in 2019. His research interests include information retrieval, data mining, and multi-media processing. He has over 20 publications appeared in several top conferences such as SIGIR, WWW, and MM. His work on Bayesian Personalized Ranking has received the Best Poster Award of WWW 2018. Moreover, he has been served as the PC member and external reviewer for several top conferences including SIGIR, ACL, KDD, IJCAI, AAAI, WSDM etc.
- Cheng Luo received his Ph.D. in Computer Science from Tsinghua University in 2017. His major research interests are in Web Search, Machine Learning and Natural Language Processing. He has published more than 20 papers in several top conferences such as WSDM, SIGIR, WWW, and IJCAI. He is serving as a co-PI at NUS-Tsinghua NExT++ center, as well as a co-founder of a start-up, MegaTech.AI.
- Xiangnan He is a professor with the University of Science and Technology of China (USTC). He received his Ph.D. in Computer Science from National University of Singapore (NUS) in 2016. His research interests span information retrieval, data mining, and multi-media analytics. He has over 60 publications appeared in several top conferences such as SIGIR, WWW, KDD, and MM, and journals including TKDE and TOIS. His work on recommender systems has received the Best Paper Award Honourable Mention in WWW 2018 and ACM SIGIR 2016. Moreover, he has served as the PC chair of CCIS 2019, area chair of MM 2019 and CIKM 2019, and PC member for several top conferences including SIGIR, WWW, KDD etc., and the regular reviewer for journals including TKDE, TOIS, TMM, etc.
- Yiqun Liu is now working as professor and Department co-Chair at the Department of Computer Science and Technology in Tsinghua University, Beijing, China. His major research interests are in Web Search, User Behavior Analysis, and Natural Language Processing. He is also a visiting research professor of National University of Singapore and a visiting professor of National Institute of Informatics (NII) of Japan, as well as a member of Tiangong AI Research Center which is founded by Tsinghua and Sogou Inc.
- Tat-Seng Chua is the KITHCT Chair Professor at the School of Computing, National University of Singapore. He was the Acting and Founding Dean of the School from 1998-2000. Dr. Chua's main research interest is in multimedia information retrieval and social media analytics. In particular, his research focuses on the extraction, retrieval and question-answering (QA) of text and rich media arising from the Web and multiple social networks. He is the co-Director of NExT, a joint Center between NUS and Tsinghua University to develop technologies for live social media search. Dr. Chua is the 2015 winner of the prestigious ACM SIGMM award for Outstanding Technical Contributions to Multimedia Computing, Communications and Applications. He is the Chair of steering committee of ACM International Conference on Multimedia Retrieval (ICMR) and Multimedia Modeling (MMM) conference series. Dr. Chua is also the General Co-Chair of ACM Multimedia 2005, ACM CIVR (now ACM ICMR) 2005, ACM SIGIR 2008, and ACMWeb Science 2015. He serves in the editorial boards of four international journals. Dr. Chua is the co-Founder of two technology startup companies in Singapore. He holds a PhD from the University of Leeds, UK.

6 POTENTIAL PC MEMBERS

- Longbing Cao (University of Technology Sydney)
- Chung-Chi Chen (National Taiwan University)
- Hsin-Hsi Chen (National Taiwan University)
- Brian Davis (Maynooth University)
- AndrÃĺ Freitas (University of Manchester)
- Siegfried Handschuh (University of Passau)
- Macedo Maia (University of Passau)
- Ross McDermott (National University of Ireland, Galway)
- Ming-Feng Tsai (National Chengchi University)
- Manel Zarrouk (National University of Ireland, Galway)
- Edgar Meij (Bloomberg L.P.)
- Ridho Reinanda (Bloomberg L.P.)
- Abhinav Khaitan (Bloomberg L.P.)
- Miles Osborne (Bloomberg L.P.)
- Diego Ceccarelli (Bloomberg L.P.)
- Xiaomo Liu (S&P Global)
- Armineh Nourbakhsh (S&P Global)
- Quanzhi Li (Alibaba Group)
- Jun Zhou (Ant Financial Group)

7 RELATED WORKSHOPS

ECONLP[1], KDD-ADF[2], FinNLP[3] are the related workshops that are hosted in ACL, KDD, and IJCAI. These workshops either largely focus on a specific application of financial document analysis or applying natural language processing techniques. In contrast, FinIR focuses on the IR problems of financial documents as well as technical problems on understanding numerical information.

REFERENCES

[1] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*. 19–27.

[2] Paolo Ciaccia and Marco Patella. 2011. Metric information filtering. *Information Systems* 36, 4 (2011), 708–720.

[3] Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 985–988.

[1] https://julielab.de/econlp/2018/

[2] https://sites.google.com/view/kdd-adf-2019

[3] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp/

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[5] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Commun. ACM* 51, 12 (2008), 68–74.

[6] Fuli Feng, Xiangnan He, Yiqun Liu, Liqiang Nie, and Tat-Seng Chua. 2018. Learning on partial-order hypergraphs. In *Proceedings of the 2018 World Wide Web Conference*. 1523–1532.

[7] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 1–30.

[8] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.

[9] Aishwarya Kamath and Rajarshi Das. 2018. A survey on semantic parsing. *arXiv preprint arXiv:1812.00978* (2018).

[10] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1101–1104.

[11] Somayya Madakam, Rajesh M Holmukhe, and Durgesh Kumar Jaiswal. 2019. The future digital work force: robotic process automation (RPA). *JISTEM-Journal of Information Systems and Technology Management* 16 (2019).

[12] Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1113–1116.

[13] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. 2018. Neural compatibility modeling with attentive knowledge distillation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 5–14.

[14] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5310–5318.

[15] Shuo Zhang and Krisztian Balog. 2020. Web Table Extraction, Retrieval, and Augmentation: A Survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 2 (2020), 1–35.