Overview of the Health Search and Data Mining (HSDM 2020) Workshop

Carsten Eickhoff
Brown University
USA
carsten@brown.edu

Yubin Kim
UPMC Enterprises
USA
kimy10@upmc.edu

Ryen W. White
Microsoft Research
USA
ryenw@microsoft.com

ABSTRACT

We present HSDM, a full-day workshop on Health Search and Data Mining co-located with WSDM 2020's Health Day. This event builds on recent biomedical workshops in the NLP and ML communities but puts a clear emphasis on search and data mining (and their intersection) that is lacking in other venues. The program will include two keynote addresses by key opinion leaders in the clinical, search, and data mining domains. The technical program consists of 6 original research presentations. Finally, we will close with a panel discussion with keynote speakers, PC members, and the audience.

This workshop aims to help consolidate the growing interest in biomedical applications of data-driven methods that becomes apparent all over the search and data mining spectrum, in WSDM's spirit of collaboration between industry and academia.

1 DESCRIPTION

Annual healthcare costs are increasing sharply due to diagnostic errors [11], adverse drug reactions or drug-drug interactions [4, 7], missing, outdated or wrong information [18], and cognitive overload on physicians [9]. There has been a growing interest in biomedical and clinical AI approaches to mitigating these effects. Online data, e.g., from search engines [19, 20], tweets [12, 14], news articles [2], etc. has been mined for health-related applications.

There are many interesting challenges in delivering intelligent decision support in the health domain. Collections of documents such as health records, scholarly publications, clinical trials, or drug orders grow at high rates and are distributed around the globe in a fragmented manner. Health data is highly multi-modal (clinical notes, time series, medical images, genomics *etc.*) and its interpretation is domain specific. Users of health information systems have different levels of expertise, and information needs, *e.g.*, a patient *vs.* a primary care physician *vs.* cancer researcher. At the same time, the data is highly sensitive and subject to legal requirements regarding privacy, security, and confidentiality. This breadth of challenges requires interdisciplinary approaches. The Information Retrieval (IR) and Data Mining (DM) communities are particularly well-positioned to tackle these problems.

Search, recommendation, and information extraction systems help lay and expert users explore ever-growing collections. Decision support systems assist in complex decision making processes. Intelligent user interfaces present the right information at the right time and allow for unobtrusive interaction all the way from the lab to the bedside. Mobile device applications and other sensors help provide a more holistic view on the patient's case than what can be gleaned in an 10-minute physician interview [10].

Whereas existing workshops and benchmarks either focus on machine learning methods or solicit solutions to concrete, narrowly defined tasks (e.g. [15]), the primary goal of HSDM is to foster interest and research into search and related data mining techniques (e.g., analysis of online data [1, 13, 21]) in the health domain. We aim to connect and provide a discussion platform for researchers working in these areas who are also interested in healthcare. Health-related topics of interest include, among others:

- Search over images/genomics/structured data
- Federated multi-modal search combining different data sources
- User interfaces for biomedical/clinical search supporting complex information needs
- Analysis of search logs and social media
- User search behavior studies
- Building and use of medical knowledge bases or ontologies
- Privacy-preserving techniques for clinical data
- Adverse event detection and prediction
- Mobile (mHealth) applications
- Wearables
- Spoken interaction with health data
- Whole exposome modeling and estimation
- Applications of data mining and machine learning
- Ethics, bias, and fairness

2 RATIONALE

While there are many healthcare related workshops and benchmarking tasks, there is currently no workshop specifically focused on health search and its intersection with data mining. Given the considerable research and industry interest in this domain (shared biomedical tasks at CLEF and TREC tend to receive the highest overall numbers of submissions), we expect the WSDM 2020 audience to enjoy a full day of scientific discussion and networking opportunities.

3 WORKSHOP ACTIVITIES

3.1 Keynotes

The workshop has two keynote speakers. Dr. William Hersh, MD is a Professor and Chair of the Department of Medical Informatics &

WSDM '20, February 3–7, 2020, Houston, TX, USA

ACM ISBN 978-1-4503-6822-3/20/02.
https://doi.org/10.1145/3336191.3371879

Clinical Epidemiology in the School of Medicine at Oregon Health & Science University (OHSU) in Portland, Oregon. He is a global leader and innovator in biomedical informatics both in education and research and an established member of the IR community.

Dr. Zachary Lipton is an Assistant Professor with joint appointments in the Tepper school of Business and the Machine Learning Department at Carnegie Mellon University (CMU) in Pittsburgh, Pennsylvania. His research spans core ML methods and theory, their applications in healthcare and natural language processing, and critical concerns, both about the mode of inquiry itself, and the impact of the technology it produces on social systems.

3.2 Research Presentations

We accepted 6 research papers for publication.

- "Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records" by Jan Trienes, Dolf Trieschnigg, Christin Seifert and Djoerd Hiemstra [17]
- "Healthcare NER Models Using Language Model Pretraining" by Amogh Kamat Tarcar, Aashis Tiwari, Dattaraj Rao, Vineet Naique Dhaimodker, Penjo Rebelo and Rahul Desai [16]
- "Lung nodule classification using Convolutional Autoencoder and Clustering Augmented Learning Method (CALM)" by Soumya Suvra Ghosal, Indranil Sarkar and Issmail El Hallaou [6]
- "A Query Taxonomy Describes Performance of Patient-Level Retrieval from Electronic Health Record Data" by Steve Chamberlin, Steven Bedrick, Aaron Cohen, Yanshan Wang, Andrew Wen, Sijia Liu, Hongfang Liu and William Hersh [3]
- "Streaming Gait Assessment for Parkinson's Disease" by Cristopher Flagg, Ophir Frieder, Sean MacAvaney and Gholam Motamedi [5]
- "Clustering Large-scale Diverse Electronic Medical Records to Aid Annotation for Generic Named Entity Recognition" by Nithin Haridas and Yubin Kim [8]

3.3 Panel Discussion

The final substantial event of the day will be a panel discussion with both keynote speakers, additional senior members of the community to be recruited from among the PC members, and the audience. The organizers will prepare a set of questions around the overarching theme of data sharing, privacy and academia-hospital-industry collaborations to kick start the discussion. The floor will be open to any participants to discuss around these and other relevant topics.

4 SPONSORSHIP

We would like to thank UPMC Enterprises[1] for their generous sponsorship.

5 CONCLUSION

HSDM 2020 is the first edition of a workshop series focusing on search and data mining in healthcare. In close collaboration with the WSDM 2020 Healthcare Day, it will bring together researchers and practitioners from a wide range of backgrounds and foster exchange and discussion around the use of data-driven techniques in the health domain.

[1] https://enterprises.upmc.com/

REFERENCES

[1] Tim Althoff, Eric Horvitz, Ryen W White, and Jamie Zeitzer. Harnessing the web for population-scale physiological sensing: A case study of sleep and performance. In *Proceedings of the 26th international conference on World Wide Web*, pages 113–122. International World Wide Web Conferences Steering Committee, 2017.

[2] John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. Digital disease detection-harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157, 2009.

[3] Steve Chamberlin, Steven Bedrick, Aaron Cohen, Yanshan Wang, Andrew Wen, Sijia Liu, Hongfang Liu, and William Hersh. A Query Taxonomy Describes Performance of Patient-Level Retrieval from Electronic Health Record Data. *HSDM 2020 Workshop on Health Search and Data Mining*, 1, 2020.

[4] I Ralph Edwards and Jeffrey K Aronson. Adverse drug reactions: definitions, diagnosis, and management. *The lancet*, 356(9237):1255–1259, 2000.

[5] Christopher Flagg, Ophir Frieder, Sean MacAvaney, and Gholam Motamedi. Streaming Gait Assessment for Parkinson's Disease. *HSDM 2020 Workshop on Health Search and Data Mining*, 1, 2020.

[6] Soumya Suvra Ghosal, Indranil Sarkar, and Issmail El Hallaou. Lung nodule classification using Convolutional Autoencoder and Clustering Augmented Learning Method (CALM). *HSDM 2020 Workshop on Health Search and Data Mining*, 1, 2020.

[7] Richard M Goldberg, John Mabee, Linda Chan, and Sandra Wong. Drug-drug and drug-disease interactions in the ed: analysis of a high-risk population. *The American journal of emergency medicine*, 14(5):447–450, 1996.

[8] Nithin Haridas and Yubin Kim. Clustering Large-scale Diverse Electronic Medical Records to Aid Annotation for Generic Named Entity Recognition. *HSDM 2020 Workshop on Health Search and Data Mining*, 1, 2020.

[9] David A Jopp and Christopher B Keys. Diagnostic overshadowing reviewed and reconsidered. *American Journal on Mental Retardation*, 106(5):416–433, 2001.

[10] Nanon HM Labrie and Peter J Schulz. Exploring the relationships between participatory decision-making, visit duration, and general practitioners' provision of argumentation to support their medical advice: results from a content analysis. *Patient education and counseling*, 98(5):572–577, 2015.

[11] J Michael McGinnis, Leigh Stuckhardt, Robert Saunders, Mark Smith, et al. *Best care at lower cost: the path to continuously learning health care in America*. National Academies Press, 2013.

[12] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM 2011*, 2011.

[13] Adam Sadilek, Stephanie Caty, Lauren DiPrete, Raed Mansour, Tom Schenk, Mark Bergtholdt, Ashish Jha, Prem Ramaswami, and Evgeniy Gabrilovich. Machine-learned epidemiology: real-time detection of foodborne illness at scale. *npj Digital Medicine*, 1(1):36, 2018.

[14] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *ICWSM 2012*, 2012.

[15] Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, et al. Overview of the clef ehealth evaluation lab 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2018.

[16] Amogh Kamat Tarcar, Aashis Tiwari, Dattaraj Rao, Vineet Naique Dhaimodker, Penjo Rebelo, and Rahul Desai. Healthcare NER Models Using Language Model Pretraining. *HSDM 2020 Workshop on Health Search and Data Mining*, 1, 2020.

[17] Jan Trienes, Dolf Trieschnigg, Christin Seifert, and Djoerd Hiemstra. Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records. *HSDM 2020 Workshop on Health Search and Data Mining*, 1, 2020.

[18] Justin M Weis and Paul C Levy. Copy, paste, and cloned notes in electronic health records. *Chest*, 145(3):632–638, 2014.

[19] Ryen W White, Rave Harpaz, Nigam H Shah, William DuMouchel, and Eric Horvitz. Toward enhanced pharmacovigilance using patient-generated data on the internet. *Clinical Pharmacology & Therapeutics*, 96(2):239–246, 2014.

[20] Ryen W White and Eric Horvitz. Evaluation of the feasibility of screening patients for early signs of lung carcinoma in web search logs. *JAMA oncology*, 3(3):398–401, 2017.

[21] Elad Yom-Tov, Diana Borsa, Ingemar J Cox, and Rachel A McKendry. Detecting disease outbreaks in mass gatherings using internet data. *Journal of medical Internet research*, 16(6):e154, 2014.