Leveraging Automated Search Relevance Evaluation to Improve System Deployment: A Case Study in Healthcare

Yizhao Ni†
Ferosh Jacob
Kaiser Permanente Digital
KP Information Technology
Oakland, CA, USA
yizhao.x.ni@kp.org
ferosh.x.jacob@kp.org

Priya Gopi Achuthan
Kaiser Permanente Digital
KP Information Technology
Oakland, CA, USA
priya.x.gopiachuthan@kp.org

Hui Wu
Faizan Javed
Kaiser Permanente Digital
KP Information Technology
Oakland, CA, USA
jason.x2.wu@kp.org
faizan.x.javed@kp.org

ABSTRACT

Over the last year, a digital initiative has been focused on reengineering the search engine for kp.org, a health web portal serving over 12 million members. However, traditional software testing techniques that rely on limited use cases and consistent behavior are neither comprehensive nor specific for capturing complex user search behaviors. To support system deployment, we utilize information retrieval (IR) technologies to monitor search performance, identify areas of improvement and suggest actionable items. In this case study we share industrial experience on building an IR evaluation pipeline and its usage to inform deployment and improve system development. The work emphasizes domain specific challenges, best practices and lessons learned during system deployment in a healthcare setting. It features the ability of IR techniques to strengthen collaboration between data scientists, software engineers and product managers in making data-driven decisions.

ACM Reference format:

Yizhao Ni, Ferosh Jacob, Priya Gopi Achuthan, Hui Wu, & Faizan Javed. 2022. Leveraging Automated Search Relevance Evaluation to Improve System Deployment: A Case Study in Healthcare. In *Proceedings of the 31st ACM Int'l Conf. on Information and Knowledge Management (CIKM '22), Oct. 17-21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA. 2 pages. https://doi.org/10.1145/3511808.3557517

1 RELEVANCE TO CIKM

We present a real-world case study that leverages IR technologies to improve system deployment in a healthcare organization with over 12 million members. The work is in conformity with CIKM topics of interest including: 1) studies using user behavior to improve systems in practice, 2) measurement techniques used at scale to understand performance of industrial systems, and 3) domain specific challenges in healthcare. Different from research papers published in CIKM that report results from experimental settings, this work presents findings and lessons learned from a real-world system deployment process. Our work emphasizes the application of IR technologies in the healthcare industry, which has unique challenges in data sourcing due to strict regulations, knowledge management and user behavior analysis. The talk will engage both academic researchers and industrial practitioners to discuss and address technical challenges in healthcare informatics.

†Corresponding speaker.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

CIKM'22, October 17-21, 2022, Atlanta, GA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9236-5/22/10. https://doi.org/10.1145/3511808.3557517

2 PRESENTER BIOGRAPHY

Dr. Yizhao Ni has over 15 years of experience in machine learning (ML), natural language processing, IR, and their applications in health informatics. His research is application-oriented and the overall objective is to improve the quality of healthcare by: 1) providing more effective provisioning of usable data, 2) assisting clinicians to generate more objective clinical decisions, and 3) providing more reliable proactive prediction of clinical outcomes. Ni is currently a principal data scientist at Kaiser Permanente, where he leads the development of ML-based IR solutions (e.g., learning to rank) for content, search, and personalization. Before joining Kaiser Permanente, Ni was an assistant professor in Biomedical Informatics at Cincinnati Children's Hospital Medical Center. His laboratory had led and participated in over 10 NIH-funded projects that develop and deploy AI solutions in clinical settings. Ni is a fellow and member of the American Medical Informatics Association (AMIA) and he has given professional presentations in the AMIA annual symposium since 2014. He served as session chair of ECML & PKDD'12, co-chair of IDM'20, and session chair of AMIA clinical informatics conference'21. Ni has over 50 publications in top health informatics journals and conferences [1-3].

3 PRESENTATION OUTLINE

3.1 Introduction

Kp.org is a web portal providing a variety of health resource (e.g., healthcare programs, health/wellness articles, drug encyclopedia) to patients and healthcare professionals. With the rapid growth of data sources, finding relevant information in kp.org is increasingly challenging. As an ongoing effort to improve user experience, a digital initiative has been focused on reengineering the search engine for kp.org. However, traditional testing techniques in software development that rely on limited use cases and consistent behavior are neither comprehensive nor specific [4]. Our motto "if you cannot measure it, you cannot improve it" has taken us to the importance of system evaluation. By utilizing IR evaluation technologies [5-7], we present an automated pipeline to evaluate search relevance performance based on user behavior data to im-

prove system deployment. In this session, we present unique characteristics of health data sources, discuss limitations of traditional software validation techniques, and specify the study objective.

3.2 Method

The automated evaluation system is built upon Microsoft Azure and Tableau services and it consists of four steps: reference-standard (RS) data creation, search result extraction, relevance evaluation and summarization, and performance visualization for causality analysis (Figure 1). In this session, we present each step in terms of methodology and technical implementation.

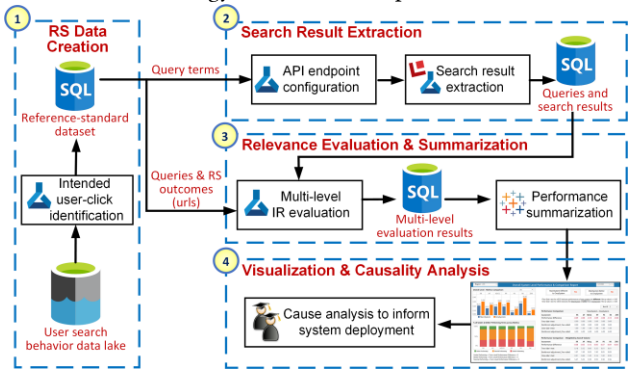

Fig. 1: The automated search relevance evaluation pipeline.

3.2.1 RS data creation (step 1 in Figure 1). User search behaviors on kp.org (e.g., queries, clicks, page views) are routinely collected and stored in an Azure data lake. To mitigate noise in the data, we adopt a literature approach to analyze a search session and identify intended user clicks (defined as RS outcome urls). In this part we specify 1) the health data sources and unique challenges to IR, 2) the approach adopted to identify RS urls based on web types and user interactions, and 3) the algorithm used to normalize relevance levels of outcome urls for a query to build the RS term-url dataset. For technical implementation, we illustrate the extract, transform, and load process of the user behavior data and the pipeline framework using Azure ML and Azure SQL database.

3.2.2 Search result extraction (step 2). Unique query terms from the RS term-url dataset are submitted to the search engine to retrieve top K relevant urls. In this part we introduce 1) the search engine developed in-house for kp.org, 2) its current ranking strategies and the comparison with a legacy system adopted by kp.org, and 3) the approach to standardize the retrieved urls so that different domain names (e.g., kp.org and kaiserpermanente.org) directing to the same web source are considered identical. For technical implementation, we discuss development of Azure ML/SQL interfaces for querying the search engine in different deployment environments (e.g., user acceptance testing environment), its challenges (e.g., credential delivery, scheduled update) and solutions.

3.2.3 Relevance evaluation and summarization (step 3). To synthesize a human-oriented conceptualization of search performance, the retrieved urls are compared with the RS urls at 3 levels: 1) term-result level, where a binary decision (match/not-match) is assessed for each retrieved url on a query term, 2) term level, where binary decisions are aggregated for a term to assess its IR performance, and 3) system level, where the performances are aggregated across all terms. In this part we specify 1) the IR evaluation metrics used, including zero result rate, precision at K, recall at K, recall-precision, average precision, reciprocal rank, and normalized discounted cumulative gain at K [8], 2) their implications and priorities in different deployment stages, and 3) statistical tests adopted to assess significance of performance difference when comparing two systems (e.g., systems before and after enhancement). For technical implementation, we present the designed bridge between Azure ML/SQL and Tableau for computing and delivering evaluation results.

3.2.4 Performance visualization and causality analysis (step 4). The performance summaries are visualized via the Tableau dashboards to inform engineers and managers. In this part we demonstrate three use scenarios during system deployment: 1) validating the search engine's functional modules such as spelling correction, 2) identifying deployment bugs such as missing data sources, and 3) detecting low-performing queries and causes to inform enhancement. For technical implementation, we describe our iterative design of the dashboards using focus group meetings between data scientists, application developers, and managers.

3.3 A Real-world Case Study in Healthcare

The case study utilized user search behavior data on kp.org between 01/01/2021 and 01/01/2022, with 3,527,777 search counts and 621,443 user clicks for 36,230 query terms. In RS data creation, we selected top 10,000 terms and top 10 RS urls per term, which covered 91% of the total searches and 80% of clicks. During system deployment, the pipeline ran weekly to monitor search performance, identify areas of improvement, and suggest actionable enhancements when appropriate. In this session, we present performance trend of the search engine from week 4 to week 16 of deployment and detail how the pipeline improved this process. We feature its ability to strengthen collaboration between data scientists, software engineers and product managers in making data-driven decisions. Finally, we conclude the talk with future applications of the pipeline such as supporting A/B testing and facilitating implementation of sophisticated ranking strategies (e.g., unbiased learning to rank) for the search engine [9].

REFERENCES

[1] Yizhao Ni, Alycia Bachtel, Katie Nause, and Sarah Beal, 2021. Automated detection of substance use information in pediatric electronic health records. *JAMIA* 28, 10 (2021), 2116-2127.
[2] Lei Liu, Yizhao Ni, Andrew Beck, et al., 2021. Understanding pediatric surgery cancellation: geospatial analysis. *JMIR* Vol. 23, 9 (2021), e26231.
[3] Yizhao Ni, Todd Lingren, Eric Hall, et al. 2018. Designing and evaluating an automated system for real-time medication administration error detection in a neonatal intensive care unit. *JAMIA* Vol. 25, 5 (2018), 555-563.
[4] Antonia Bertolino, 2007. Software testing research: achievements, challenges, dreams. *Future of Software Engineering (FOSE'07)*. 2007, 7705-7716.
[5] Himanshu Sharma and Bernard Jansen, 2005. Automated evaluation of search engine performance via implicit user feedback. *In Proceedings of ACM SIGIR'05*. ACM Press, Salvador, Brazil, 649-650.
[6] Aleksandr Chuklin and Maarten de Rijke, 2016. Incorporating clicks, attention and satisfaction into a search engine result page evaluation model. *In Proceedings of ACM CIKM'16*. ACM Press, Indianapolis, IN, 175-184.
[7] Katja Hofmann, Lihong Li and Filip Radlinski, 2016. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval* Vol. 10, 1 (2016), 1-117.
[8] Christopher Manning, Prabhakar Raghavan and Hinrich Schütze, 2009. Evaluation in information retrieval. In *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.
[9] Tie-Yan Liu, 2011. *Learning to rank for information retrieval*. Springer Science & Business Media, 2011.