PhD FORUM

**An Enriched Collaborative Recommendation Engine for Arabic Content**

Recommendation, personalization, user profiling, enrichment, collaborative

# Introduction

In an era of personalization, content publishers are in a race to better understand their visitors in to push the right content in the right time. Also, increased literacy about the data value chain called for better user profiling to enhance user targeting. Different recommendation algorithms are used, each with a special logic, demonstrating strengths relative to some types of readers and for different content styles. However, there is a big limitation in recommendation engines that target the complicated Arabic language and analytics/insights relative to it. In my PhD, an Arabic recommendation engine that deploys four different algorithms will be developed (content based, collaborative, trending and random). The direction is to enhance existing evaluation models by emphasizing the relationship between users and content deploying natural language enrichment techniques.

**Theoretical framework & Research questions**

To better understand online content consumers without violating their privacies and allow content publishers to make actionable insights, this work aims to develop a smart, enriched recommendation engine, fully functional on online articles, with 4 different logics (collaborative, content based, trending and random) along with a dashboard that allows deep diving into the analytics-highlighting the value of the recommendations. Collaborative filtering is among the most commonly used algorithms for recommending content, deriving the logic of the crowd and using the witnessed patterns to suggest it to lookalikes (Gupta and Gadge 2014, 3) and (Cheung, Lam, She & Tao 2014, 3). This logic has not been efficiently tested on Arabic content for unknown users and in compliance with GDPR. Similarly, the content based logic recommends new pieces of content with high similarity functions assuming that one can be interested in topics he usually reads. When it comes to Arabic tested logics, the focus is always on the user similarities and not on the relationship between items or using cosine vector similarity as presented by many papers (Tewari, Kumar & Barman 2014, 1) and (Ren, Zhang & Zhu 2016, 2). The effect of geographic locations (Gupta and Gadge 2014, 6), when combined with trending topics in the collaborative approach are rarely tested on Arabic content as presented in the case study about Sudan (Kashef, Osman, Sabr & Samani 2017, 5). In the examined literature, we hardly found assessment models that stress the relationship between user behavior/consumption of content and user engagement. Available analytics platforms such as Google analytics are too complicated for deep analytics by journalists especially segmentation and clustering steps needed to derive well-structured insights. Traditionally, adopted approaches rely on calculating the difference between predicted ratings and the actual ratings (Aggarwal, Parhi & Pal 2017, 2) and forecast test scores (Xu & Zhang 2017, 3) or ( Kharita, Kumar and Singh 2018, 3). Finally, majority of existing Arabic natural language processing works are still being improved and thus and require more testing as mentioned in recent journals (Armouty & Tedmori 2019, 4) and (Al-Ghamdi, Al-Khalifa & Khabti 2017, 2). So will collaborative filtering algorithms give good results for Arabic recommendations vs content based approaches? When combined to these algorithms, will the geographic locations of users and trending tags on the website affect the click through rate? How effective will these approaches be when tested on anonymous users to remain compliant with GDPR? How can the existing NLP limitations on Arabic language be mitigated?

**Methodology & Discussion**

After an extensive literature review along with an assessment to possible testing capabilities on real websites(leading websites in the MENA region addressing different themes such as Kooora.com, strive.me, yawmiyati.com for 6 months) , I outline the main technical steps after benchmarking with the lambda architecture (Numnonda 2017, 3). Data will be collected from websites using a scalable backend java web service that handles millions of requests, sending it to a messaging system where a scalable (on demand) processing server ingests the logs. In this phase, an Apache beam (a parallel processing library) is used and I deploy it to Google dataflow. The clean processed data is saved to big query(map reduce) which is used as a data warehouse that can query petabytes of data in seconds building the blind profiles of users. My recommendation engine reads the data from big query and generate user profiles, similar to the approach presented in the Correlation based similarity calculation formula by Liu and Wu (Liu & Wu 2019, 3) and then it generates recommendations based on latest articles, serving the results in widgets, matching between the two profiles. Different optimization techniques to user group selection were considered before adopting this approach, including the clustering approach presented by Xia (Hong-xia 2019, 3) and (He, Huangg, Yang &Zheng 2016, 3). Impressions and clicks of recommendations are stored as events to be used in widget performance reports. For enriching the profiles, web crawlers are used to bring in new articles and read posts metadata, save it to big query to be used in the pushing new recommendations.

# REFERENCES

Al-Ghamdi, Sharefah A., Joharah Khabti, and Hend S. Al-Khalifa. 2017. “Exploring NLP Web APIs for Building Arabic Systems.” *2017 Twelfth International Conference on Digital Information Management (ICDIM)*. doi:10.1109/icdim.2017.8244649.

Armouty, Batool, and Sara Tedmori. 2019. “Automated Keyword Extraction Using Support Vector Machine from Arabic News Documents.” *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. doi:10.1109/jeeit.2019.8717420.

“Developing a Recommendation System for the Students Applying for Sudanese Higher Education Institutions Using Graph Databases and Myers Briggs Type Indicator Personality Test.” 2017. *2017 Joint International Conference on Information and Communication Technologies for Education and Training and International Conference on Computing in Arabic (ICCA-TICET)*. doi:10.1109/icca-ticet.2017.8095289.

Gupta, Jyoti, and Jayant Gadge. 2014. “A Framework for a Recommendation System Based on Collaborative Filtering and Demographics.” *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*. doi:10.1109/cscita.2014.6839276.

Hong-Xia, Wang. 2019. “An Improved Collaborative Filtering Recommendation Algorithm.” *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*. doi:10.1109/icbda.2019.8713205.

Kharita, Mukesh Kumar, Atul Kumar, and Pardeep Singh. 2018. “Item-Based Collaborative Filtering in Movie Recommendation in Real Time.” *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. doi:10.1109/icsccc.2018.8703362.

Liu, Gaojun, and Xingyu Wu. 2019. “Using Collaborative Filtering Algorithms Combined with Doc2Vec for Movie Recommendation.” 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). doi:10.1109/itnec.2019.8729076.

Numnonda, Thanisa. “A Real-Time Recommendation Engine Using Lambda Architecture.” Artificial Life and Robotics 23, no. 2 (2017): 249–54. https://doi.org/10.1007/s10015-017-0424-8.

Pal, Ashish, Prateek Parhi, and Manuj Aggarwal. 2017. “An Improved Content Based Collaborative Filtering Algorithm for Movie Recommendations.” *2017 Tenth International Conference on Contemporary Computing (IC3)*. doi:10.1109/ic3.2017.8284357.

Ren, Caixia, Ping Zhu, and Hua Zhang. 2016. “A New Collaborative Filtering Technique to Improve Recommendation Diversity.” *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. doi:10.1109/compcomm.2016.7924910.

Tao, Zhaojie, Ming Cheung, James She, and Ringo Lam. 2014. “Item Recommendation Using Collaborative Filtering in Mobile Social Games: A Case Study.” *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*. doi:10.1109/bdcloud.2014.73.

Tewari, Anand Shanker, Abhay Kumar, and Asim Gopal Barman. 2014. “Book Recommendation System Based on Combine Features of Content Based Filtering, Collaborative Filtering and Association Rule Mining.” *2014 IEEE International Advance Computing Conference (IACC)*. doi:10.1109/iadcc.2014.6779375.

Thorson, Emily. “Changing Patterns Of News Consumption And Participation.” Information, Communication & Society 11, no. 4 (2008): 473–89. https://doi.org/10.1080/13691180801999027.

Xu, Xiangshen, and Yunhua Zhang. 2017. “Collaborative Filtering Recommendation Algorithm Based on Hybrid Similarity.” *2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)*. doi:10.1109/iccsec.2017.8447000.

Yu, Peng. 2019. “Merging Attribute Characteristics in Collaborative Filtering to Alleviate Data Sparsity and Cold Start.” *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. doi:10.1109/itnec.2019.8729461.

Zheng, Longshuai, Shengqi Yang, Jian He, and Zhangqin Huang. 2016. “An Optimized Collaborative Filtering Recommendation Algorithm.” *2016 2nd International Conference on Cloud Computing and Internet of Things (CCIOT)*. doi:10.1109/cciot.2016.7868309.

Ziani, Amel, Nabiha Azizi, and Yamina Tlili Guiassa. 2015. “Recommender System for Sports Articles Based on Arabic Opinions Polarity Detection with a Hybrid Approach RSS-SVM.” *2015 3rd International Conference on Control, Engineering & Information Technology (CEIT)*. doi:10.1109/ceit.2015.7233152.