An Industry Based Evaluation of Recommender System Performance

Mohamed Hussain, MHL, Latif

CognativeX, University College of London, mohamed.latif.10@ucl.ac.uk

This paper examines the analysis of recommender system performance with specific focus on the context and customer behaviour. We begin by identifying the difference between performance and value. This serves as a foundation for understanding the Value-Assessment of Recommender Systems (VARS) framework for evaluation that we have created. We present the methodology behind the assessment of implicit feedback recommender systems concluding with a general framework that will assist future researchers in assessing recommender systems based on value when using implicit feedback. Implicit user behaviour assessment heavily depends on the context of the recommender system, as a result, we discuss the application of our framework in an online news environment. Finally, we present our results for assessing recommender system algorithms in an online environment.

CCS CONCEPTS • General and reference □Cross-computing tools and techniques □Evaluation • Information systems □Information retrieval □Evaluation of retrieval results • Information systems □Information retrieval □Retrieval tasks and goals □Recommender systems • Information systems □World Wide Web □Online advertising □Content match advertising

Additional Keywords and Phrases: Value Based Evaluation, Multi-Stakeholder Systems, Beyond-Accuracy Measures, Implicit User Feedback

1 Introduction

Traditionally a recommender system (RS) performance has been evaluated by looking at algorithmic improvements of performance analysed in offline environments such as prediction accuracy or processing time [18,41]. This kind of approach has been successful originally because of the historic outlook on RS to provide users with resources that are more suitable to their needs [38]. More recently, researchers have shown that there are more crucial elements for recommendations to be assessed based on such as fairness [11], diversity [43], novelty [29], coverage, and serendipity [16].

As RS become an integral part of many online businesses and with the need to cater for more than just the user, the evaluation of these systems has become more complex. Multi-stakeholder RS focus on the fact that there are multiple stakeholders that a recommendation produced must satisfy the requirements for [3]. A good example here is a job RS, where the recommended jobs should not only satisfy the candidate's requirements and experience, but also the recruiter's requirements and interests [19,31,42]. RS are being applied for various different purposes in industry, from recommending new social connections [20], to recommending potential dating partners on dating sites [36]. The success of the recommendations are no longer measured on the accuracy of rating prediction, but rather on the mutual match between two parties.

The majority of recommender systems make use of explicit feedback [35]. Explicit feedback is feedback that is collected through the user explicitly. An example of this is user ratings provided by the user such as movie ratings [6]. Explicit feedback is a reliable source of information on user preferences. This type of feedback uses no assumptions and infers no interests on users based on their actions [23]. Explicit feedback also provides transparency in the user recommendation experiences [12].

This form of feedback was the foundational driver for the growth of the RS world, however, it has been agreed upon that this form of data is scarce, limited, noisy, and sometimes inaccurate when used for user preference representation [22,30]. This is partly the reason researchers have recently taken more interest in implicit feedback for RS. Implicit Feedback reduces the burden of requiring the user to respond to a call of action. It usually infers logical assumptions from user behaviour collected through web logs and scripts. Implicit feedback can take various forms such as the history of purchases, the history of items in basket, navigation history, time spent on a page, links clicked on by the user, response to e-mail, button clicks, and comments, among others. Implicit feedback has been used in various different ways to collect feedback from users. Authors in [32], for example, use click behaviour data to personalise news recommendations. For a music platform, a measure of the frequency a track is played is regarded as a positive rating according to [25].

Implicit user feedback is more readily available and can be used in various ways to enhance recommendations and give context to the assessment. For example, paper [21]explores implicit feedback as a secondary data source for their RS and showcase the superiority in performance to the state-of-the-art models. Nonetheless, Although research in the implicit user behaviour feedback has been increasing over the years, it still has a long way to become well-established. A core factor that hinders the compounding progress in research for the world of RS has been the evaluation of these systems. Whilst this has been a long ongoing research aspect of RS, there has been great development in key-concepts driving this area such as beyond-accuracy measures [8,28], and multi-stakeholder RS [1,2,45]. The following are contributing elements to the problem:

1. Limited data sets that are available for researchers to use that include implicit user behaviour
2. Misrepresentation in offline evaluation or technical difficulty and high costs in online evaluation
3. Absence of a formalised comprehensive evaluation model

Publicly available datasets have known issues and limitations. There has been an effort to include user logs or user behaviour data, however, these datasets may suffer from biases and ambiguity with regards to the circumstances under which the data was collected [27]. Unfortunately, these may sometimes not even be made known, for example, a running promotion during the data collection, an existing recommendation on the site, the order of recommendations presented, or even a change in the RS user experience.

Given the complexity and costs involved in running online field tests, the most widely used method for evaluation is using offline tests [24]. This is normally carried out by using a subset of the data to train the model and the rest of the data to check the prediction accuracy of the model such as user rating preferences, click behaviour, or purchases. Interestingly, it has been reported by a growing body of work that RS that are the highest performing for accuracy in offline tests, did not actually lead to a superior performance or better perceived accuracy in the online test when the real users interacted with the system [9,15,17,39,40]. When taking a closer look at this problem, it breaks down into explicit based and implicit based RS. Explicit RS have over 80 different approaches for evaluation, however all present major shortcomings with no agreed best-practice for evaluation [10]. Implicit feedback RS have a greater problem with evaluation, this is due to the inability of comparing recommendations with actual user preferential data. However, there has been a growing body of work to explore the meaning of user feedback and how that can be incorporated in the assessment of RS [26]. Nonetheless, the greater problem which has been established for a while is that accuracy measures are “not enough”, or indicative for that matter, of the real value of a RS [5,16,18,28,34,37,43].

With the rise in application of RS in academia and businesses, it has become imperative to create a formalised evaluation framework that:

1. Methodises the approach to RS assessment and evaluation
2. Enables empirical metrics from beyond-accuracy measures as a proxy for value representation
3. Offers inclusive evaluation for both explicit and implicit RS, and offline or online experimentation

4. Builds on a multi-stakeholder RS model to include all user requirements, and accordingly the success criteria
5. Systematises the research and discussions around RS evaluation

In this paper, we therefore present a framework for RS assessment and evaluation. We consider the reproducibility of this framework across algorithms, industries, and experimental settings. We ensure that the framework is applicable for implicit as well as explicit RS. Our framework inherently builds in the concepts of multi-stakeholder RS by looking at all users involved in the success of the RS. We also advocate for beyond-accuracy measures by taking a value based assessment perspective on RS. We present a case of how implicit feedback recommender systems can look at the human psychology of user behaviour and devise practical empirical beyond-accuracy metrics such as Engagement (E), Bounce Rate (BR), and Click-Through Rate (CTR) for evaluation. Finally, we conclude this paper by showcasing the application of the framework on an implicit feedback recommender system in a news website environment and discussing the results. A secondary objective achieved by this framework is a clearer understanding of the RS design. By formalising the evaluation criteria, researchers and system developers – by reverse engineering – will be able to design superior RS.

2 Defining Performance and Value

With the popularity of recommendation engines rising both commercially and in the research community, it has become increasingly more important to decide on the right algorithm that matches the domain and task at hand [18]. This could not be feasible without a clear understanding of how to measure these algorithms against each other. As a result, it is critical to distinguish the terms performance and value.

As online industries become more aware of the opportunities Recommender Systems create, it becomes critical for the research community to create optimized approaches, appropriate metrics, and rapid experimentation frameworks [7]. Personalisation is a core element of online industries and recommender systems are powering it. A recommender system uses knowledge of the user to enhance a particular service which in turn provides a specific value [17,24].

In this paper, we refer to value as the positive increase in the key-performance indicator (KPI) of interest. This can only be done with a clear view on what task the recommender system is required to assist in. According to [18] the standard way of choosing an algorithm is by comparing a number of algorithms offline using some evaluation metric. This may be particularly difficult when looking at implicit feedback data with no explicit knowledge of user rating. As a result, this only allows for evaluation of a specific problem within recommendation systems, for example, accuracy of predicting user actions (e.g. a purchase) or interests (e.g. ratings). Below, are the definitions of performance and value used in this paper.

A. **Performance:** the abstract measurement or evaluation of algorithmic factors for an algorithm. This is normally an assessment carried out away from the context of user satisfaction and preferences or business/stakeholder and industry requirements. Examples are measuring the speed of a recommendation, or the accuracy of rating prediction.
B. **Value:** the contextual assessment of the utility of the recommendation algorithm. The context and utility can only be derived by a clear understanding of how and where the recommendation algorithm should be applied and the impact desired. The value-based evaluation of a recommendation algorithm for online industries will therefore be measured by added value to the stakeholders - either the end user and/or the service provider or sometimes both.

3 VARS Framework

In this section, we give an in-depth explanation of the Value Assessment Recommender System (VARS) Framework. As can be seen in Figure 1, in this approach all the stakeholders (e.g. business , service provider, user) will define the problem, and by that, the objective & utility of the recommender system. The industry and application of the recommender system will inform the system architect of the data available for utilising in recommendation process. The objective and utility will be translated into measurable quantitative Key-Performance Indicators (KPIs) which will be monitored for evaluation. Finally, the online testing environment is an online real-world sandbox where user behaviour is

being tested and evaluated in a scientific way. In this circumstance adaptation of algorithms and testing can happen dynamically and iteratively to ensure conclusive granular results.

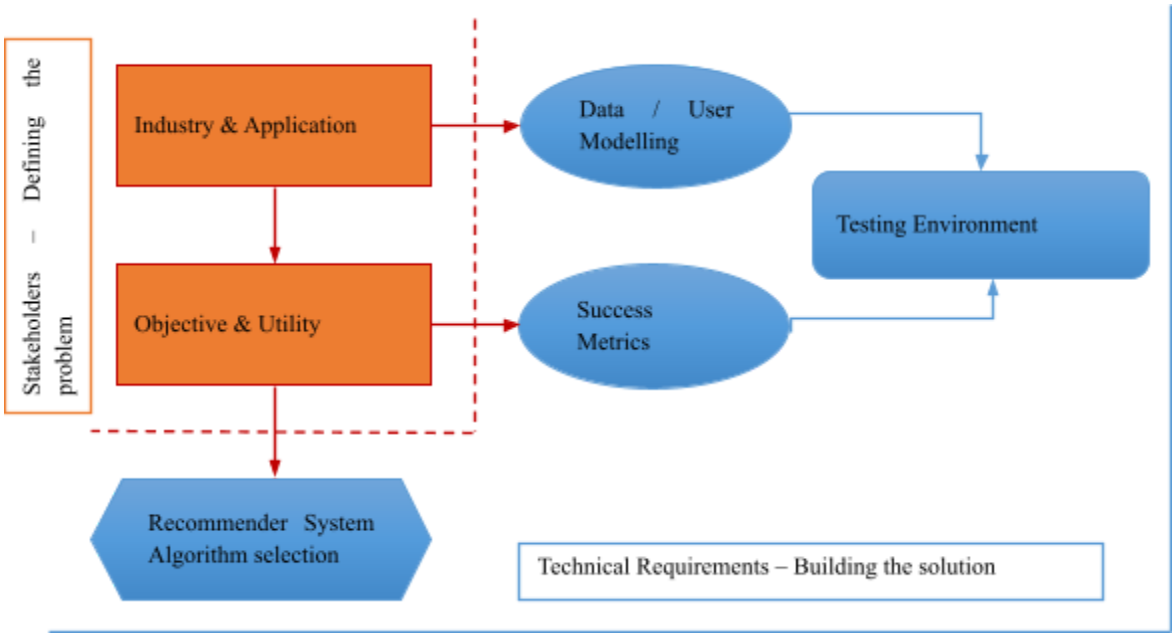

Figure 1: **The components of the VARS Framework that require attention of the system architect**

4 Industry & Application

It is important to acknowledge the industry that the recommender systems algorithm is to be implemented in. The industry will give rise to different objectives/utility the recommendation algorithm can be used for. The industry will also dictate the type of content and context in which the recommender system will be placed in, which in turn will dictate the type of data available for use. Another reason the industry is particularly important is for understanding the psychological intentions and use of users for the service. For example, someone streaming music may be inherently more open to exploring new music, than someone browsing for news, who generally would have a specific focus or categorical interest when consuming content. The behaviours of the same user on these two different services will be very different due to the nature of the service.

In paper [33], the author categories the applications of recommender systems into eight main categories (e-government, e-business, e-commerce/e-shopping, e-library, e-learning, e-tourism, e-resource services and e-group activities) and summarises the related recommendation techniques used in each category.

Paper [14] provides a categorisation of recommender systems from an industry perspective. They categorise the common recommender system applications into the following groups:

- **Entertainment:** Recommending movies, music, IPTV, etc.
- **Content:** Personalisation of newspapers, web-pages, e-learning applications, email filters, etc.
- **E-Commerce:** Recommendation of consumer products to buy, such as clothes, books, etc.
- **Services:** Recommendations of travel services, experts for consultation, or matchmaking services.

5 Objective & Utility

The objective defines the role of the recommender system. This determines two important factors, the recommender system algorithm of choice, and the metrics in which the success of the recommender system algorithm can be assessed based on.

Objectives of recommender systems in general are dependent on the stakeholders involved. More recently, there has been an increased interest in regarding recommender systems as multi-stakeholder environments. Multi-stakeholder Recommender Systems (MSRS) is a new perspective proposed which formalises the understanding that recommendations may have multiple stakeholders involved with varying interests and utility [13]. When considering utility, one should take into consideration all the stakeholders' interests in to account [3].

In general for web applications, the service providers and consumers should have a mutual benefit when applying/interacting with recommender systems and setting objectives. In general, all business objectives should add some value to the end user. This is very prominent in online applications of recommender systems. Take for example recommending an ad online, it is a requirement for the user seeing an ad to be interested in the advertiser, but additionally, the advertiser must also be interested in advertising to this user. For instance, a. user may be interested in automobiles, but the advertiser might be only interested in users who are interested in automobiles and considering to purchase [44]. Take for instance another example is online e-commerce. While the business owner wants to surface more products to increase their sales, the user expects to be presented with potentially relevant and interesting products they need, thereby saving them time .

Based on [14], the most common reasons for the application of recommender systems by service providers can be categorised by the categories below:

- **Increase the number of items sold:** This can only be achieved by understanding what the user wants. In terms of products, this could be recommending relevant products or products that best match the user's desires. In web-page based applications, the revenues are based on the number of pages viewed, so recommending more relevant and interesting items will increase sales.
- **Explore new items:** Sometimes, the service provider would like to make new irrelevant items apparent to their consumers. By recommending these items to users based on their pattern of behaviour, recommender systems can assist in making a more targeted and personalised exposure of this service.
- **Increase the user satisfaction:** by enabling a recommender engine, users may find it more satisfying as they require to sift through less items/content/products/services to arrive at a relevant and desired outcome. This can greatly enhance the user experience and avoid bombardment of the user with irrelevant messaging.
- **Increase user fidelity:** recommender systems give users the opportunity to influence their experience by learning from action and inaction on items presented to the user. Furthermore, it allows for returning users to be recognised as valuable customers by the enhanced personalised service they receive.
- **Understand user preferences:** recommender systems work by creating user profiles. These user profiles can be very useful to business decision makers if exposed in the right way. It can assist in improving stock management, or even inform of what content/services to invest more in.

6 Data

The data that is available for use is directly related to the industry and application of the recommender system. For example, web-based recommender systems use a different set of data metrics to music recommender engines. Understanding what data is available can create a more holistic and accurate evaluation of the value of recommender systems.

Due to the complexity of human psychology and behaviour, it is critical in taking a human approach to understanding the psychology of online behaviour. This gives the system architect an opportunity to surface patterns in affecting data dimensions that would usually be overlooked, or even dismissed due to its technical irrelevance. For example, the choice

of a movie by a user could be highly influenced by the actors in the movie, leading to anomalous behaviour if solely looking at user's genre preferences.

In any implicit recommender engine problem, there are two uses for data:

- **Training:** the training data set is used in the training of the algorithm as input and feedback.
- **Evaluation:** the evaluation data is used to evaluate the performance of the recommendation engine.

In this paper, we will focus on the evaluation of the recommender system.

7 Success Metrics

The success metrics chosen by the system architect should directly tie to a quantitative evaluation of the objectives dictated by the business/stakeholders. This gives context and utility to the application of the recommender system algorithms. Furthermore, it offers opportunity to benchmark algorithm performance from a utility critical perspective rather than an engineering optimisation perspective.

Implicit feedback recommender systems rely on assessing non-explicit indicators that indirectly infer a preference of some form for the user. For example, assessing how much a user enjoyed a new recommended music can be derived from indirect indicators such as adding to playlist, repeating or sharing the music rather than looking solely at explicit ratings.

In industry, and due to the diverse application of recommender systems, not all applications of this technology have the advantage of using explicit rating data to measure the quality of recommendations. Implicit feedback recommender systems are critical for two major cases:

1. When the luxury of explicit ratings is unavailable
2. When solving for anything other than predicting ratings

8 Online Testing Environment

The authors of [7] describe the need by the research community to create optimized approaches, appropriate metrics, and rapid experimentation frameworks for online industries as they become more aware of the opportunities Recommender Systems create. As a stakeholder solving for a particular need/problem, it is critical that the choice of recommendation algorithm and evaluation metrics are carefully made. However, these efforts will produce debatable results if not applied in a fair scientific environment for testing and assessing [4].

The online testing environment should allow the deployment of the recommender system algorithms to randomly selected user groups. The testing environment should take all affecting variables into account and strive for complete control and reporting over the variables being tested at any given time. Effectively, users consuming the same content or service should be offered recommendations from the variations of algorithms/recommendation techniques to be able to evaluate the effect of an algorithm on the behaviour of the group. Additionally, it is highly recommended for two control groups to always be present in any experiment. A group that gets offered no recommendations and another group that gets offered bad recommendations. This critical element ensures the explicit awareness by the system architect of the value of recommendations vs. no recommendation, and the attractiveness of recommendations for their content Vs. the natural behaviour and habit of users to consume recommendations.

The following elements should be taken into account when building an online testing environment:

- **User Experience:** The consistency of visibility and experience in the presentation of recommendations to the user groups. For example, recommendations on the checkout page of an e-commerce platform via a pop-up will undoubtedly be seen, and hence interacted with, far more than a recommendation presented in a grid on a categories page of the site.
- **Industry Variables:** All affecting industry based variables should be taken note of by the system architect to ensure accurate derivation of conclusions. For example, the impact of recommending a specific artist's music who has recently received publicity for it can tremendously skew the results regardless of the utility of the recommendation that is being tested.
- **User Groups:** groups undergoing the live test should be equally distributed and randomly selected. As mentioned above we highly recommend that two user groups are always included in any test to allow fair and scientific benchmarking. The two control. Groups can flag the unaccounted variables

as a general user behaviour thereby avoiding misinterpreted spikes or troughs in the recommendation value performance.

- **Period and time of testing:** Due to the nature of online tests there are many variables outside of the control of the system architect. For example, news consumption is heavily dependent on national and international affairs outside the remit of the system architect. The sales of products online fluctuate based on seasons for example Christmas or summer.

9 Motivating Example - the VARS Framework in a news environment

The VARS framework organises the System Architect's thinking and approach in solving the problems faced by the industry application, whilst ensuring taking into account the business and stakeholders demand for value and performance. In this section, we present the application of the VARS Framework in an online news environment. Using the VARS framework, we first begin by discussing the application of the recommender engine on the website. We then go on to derive a set of Key Performance Indicators (KPIs) to assess the recommender engine performance in terms of value.

9.1 Industry & Application – Online News Environment

The majority of online content and in specific news business models depend on advertising. Advertising business models revolve around users. Websites heavily rely on users visiting their website and consuming content. However, with the vast numbers of content pieces and sources available to users, it becomes very hard and competitive to be seen by users. As a result, there is a critical need for personalisation of content. By reducing the time needed for the users to find relevant pieces of content, online business owners can achieve higher loyalty and retention rates, hence increasing their revenues.

The online news environment suffers from some of the most prominent problems recommender engines face. The cold-user and cold-item problem. Online news websites offer content to users for free without the need to login or sign-up. This leaves business owners with limited accumulation of historic knowledge about a user's preferences. Furthermore, news websites publish over 75 content pieces per day, with an average life-span of 48 hours before it becomes outdated. This presents recommender engines with a further challenge of having a large item inventory to be recommended with a very limited outlet space to surface these recommendations

9.2 Objective and Utility – Increase Revenues and Audience Engagement

Online content serving websites who rely on advertising business models usually auction their audience to advertisers through an ecosystem of ad technology. The value of the user depends on both demographic and behavioural profiles. The demographic characteristics of an audience of a site are heavily affected by the brand and general content strategy of the site. However, behavioural profiles can be heavily influenced by the Recommender System.

Personalisation can assist in increasing the time spent on the website, increasing the average number of pages consumed per session, and decreasing the bounce rate of pages. These metrics directly affect the behavioural profile of a user, and hence the value of the user. Improving the behavioural profile of a website's users will thereby increase the revenues of the website. Therefore, in this experiment the value of the recommendation engine will be a measurement of the added value to the behavioural profile – user engagement.

9.3 Data

In this experiment, we have access to user generated data along with contextual information about the content pages being consumed. This data can be useful for the training and evaluation stages of the recommendation engine building. We split the data into three parts, contextual data, behavioural data, and experiential data. In this paper, the data of

interest are only those that can directly be used in the evaluation of the recommender engine. The following are the most prominent data points that can be used in a calculation or in the raw format for evaluation purposes:

- Length of Article – Calculated as average time to complete
- Time Spent on Page – The time a user spends on an article measured in seconds
- Article Clicked – Clicked to open an article
- Article viewed – spent more than 10 seconds in the article page after clicking on it
- Exit page – The article on which the user then exited the site
- Session length - The duration of a visit from entry to exit
- Widget Impression – Counted on a successful load of a recommendation element on a page
- Widget Viewed – The appearance of the recommendation element in the active view of the user's screen
- Widget Click – Counted when the user has clicked on a recommended article within the recommendation element

9.4 Success Metrics

The success metrics below are all a directly or indirectly a calculation of the metrics above. For evaluation, we focus on data that indicates engagement of the user, which in turn results in increased revenues. The success metrics will be used a change in these numbers can be attributed to a successful application of the recommender system, in addition to the evaluation of the recommender system

The following is a list of measures that indicate engagement:

- **Page Views Per Visit:** The number of pages consumed per session can demonstrate the added value of the recommendation engine surfacing the right content to the user at the right time in order to extend their stay on the site.
- **Bounce Rate:** The bounce rate here is calculated as the rate of people who exit the article before spending at least 5% of the time required to consume the full article. On average, we found that the optimal time is 10 seconds. This can indicate whether the recommended article was actually of interest to the user or not. By personalising the content the user bounce rate on articles should decrease as a result of understanding the user's interests and behaviour of content consumption.
- **Average Time Spent on Page:** The time a user spends reading an article is directly related to the interest of the user in that specific article. If a RS is increasing this metric then it is successfully recommending an article of user interest.
- **Average Session Duration:** The average time spent on the website should also give an evaluation of the effects of personalisation on the user's behaviour. However, not only will this overlap with the pattern of behaviour of the average time spent on a page, but it will also be heavily affected by the news and content beyond the remit of personalisation.
- **Click-Through Rate:** Recommendations are presented to users via a widget either on the side or at the end of the page. Click-Through Rate is the percentage of users who view the recommendations and click and click on a recommended article. This is a strong indication that the recommendation engine was able to capture a user's attention *and* assist them in engaging with more content. This measure can inform us of the success of personalisation on the audience from a recommendation attractiveness and user-acceptance perspective.

9.5 Online News Website Testing Environment

The online testing environment was created to assess the value of implicit feedback recommender algorithms against each other. For this experiment, there were:

- 42,117 unique user profiles
- 2,413 unique articles
- 20,579 unique users in the ALS matrix
- 1,024 unique articles in the ALS matrix

To be able to attribute the change to the Recommendation System, the test is carried out on four different audiences in a A/B testing environment. The total population of the site was divided equally into six different populations:

- **Population A:** Presented with Content-Based recommendations
- **Population B:** Presented with the ALS-WR Collaborative-Based recommendations
- **Population C:** Presented with Trending-Based recommendations
- **Population D:** Presented with bad recommendations
- **Population E:** Presented with random recommendations
- **Population F:** Presented with no recommendations at all

To ensure credibility of the results, these populations were split equally and randomly. The test operated on the same articles with no change to the UI/UX for a fixed period of time at the same time. Populations D, E, and F acted as a control groups to account for any breaking news, viral news, or temporary change in overall audience behaviour that may occur during the evaluation process.

9.6 Results

For the purpose of this paper, in this section we will present some results based on the above methodology. The purpose of this section is not to answer an experimental question, but rather, showcase how the results derived can be used in the explainabilty of the evaluation and conclude with a clear outcome on the value of the RS.

The results in Figure 2 demonstrate that the Collaborative Filtering (CF) RS positively affected the performance of the website when compared to no personalisation. The average time spent on a page per day for the website visitors was approximately doubled as can be seen in Figure 2. This figure clearly demonstrates that the CF Algorithm positively contributes to users spending more time on average on articles when the RS is deployed. As a result the CF RS has succeeded in the objective of increasing user engagement.

Another example of evaluation is Figure 3, where multiple RS algorithms are deployed in an A/B testing environment mentioned above. In this case it was a CF, Content Based (CB), and Trending (TR) algorithm. The objective to be tested here is which RS is producing the highest CTR. The reason CTR was selected is because it is a measure of the recommendation attractiveness and showcases the ability of the RS to present the users with recommendations that are considered as potentially interesting news stories. Using Figure 3, it can be concluded that no algorithm had a visibly higher impact. Interestingly, the Trending algorithm performed the best in this test. Meaning users are more likely going to interact with a trending recommendation than they are to interact with an article recommended to them based on their interests.

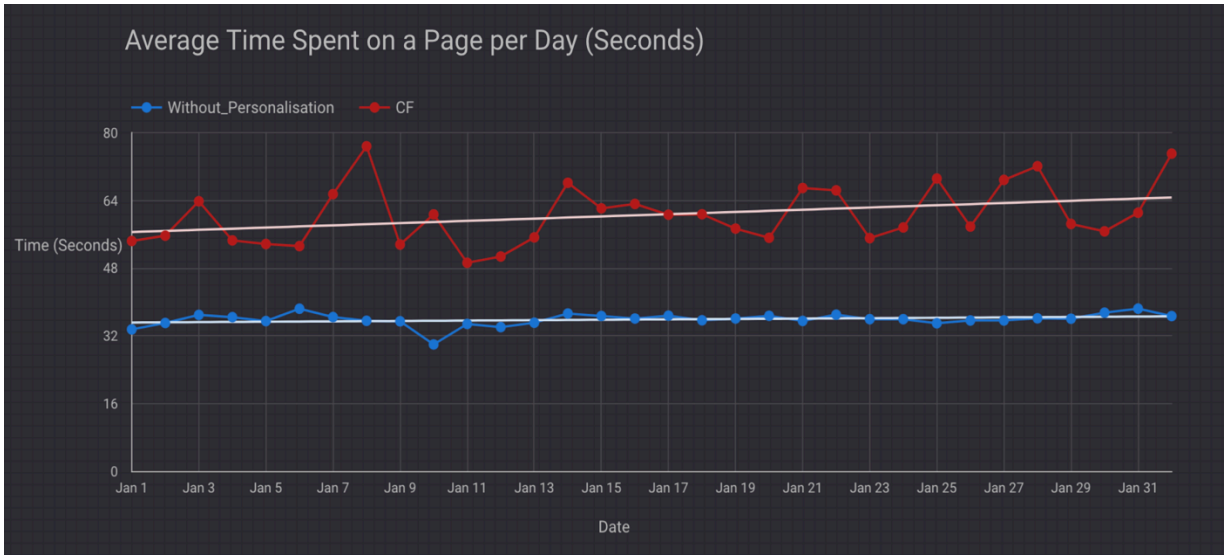

Figure 2 - **Average time spent by users on a page when arriving through a recommendation vs. when arriving without a recommendation.**

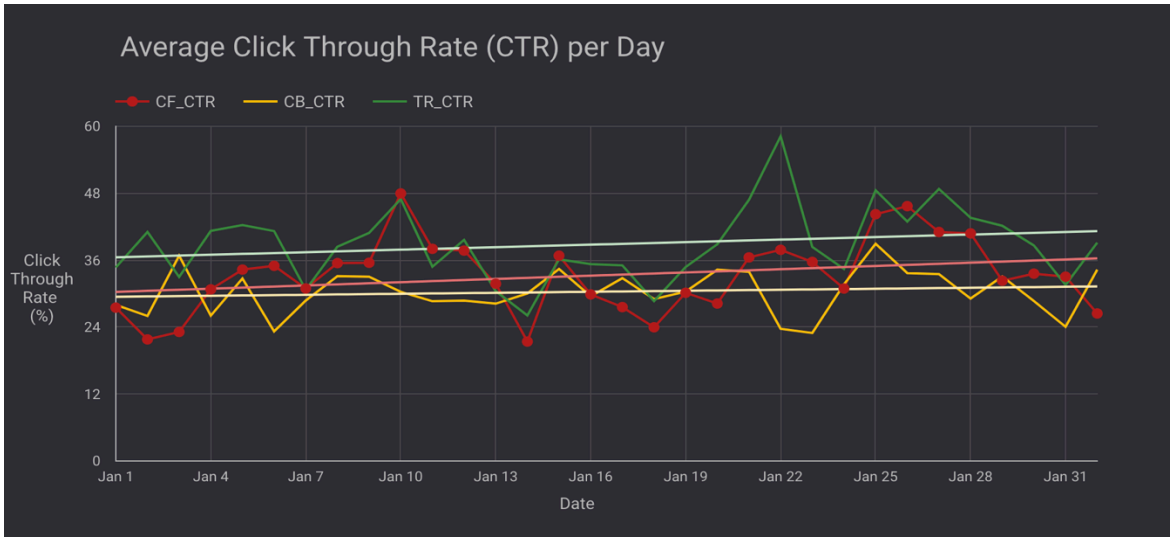

Figure 3 – **The average click through rate per day for three algorithms, Collaborative Filtering (CF), Content-Based Filtering (CB), and Trending (TR) algorithms**

10 Conclusion and Future Work

In this paper we present a framework for evaluating and assessing recommender systems. We incorporate the value and industry in our framework and utilise well researched concepts such as behind-accuracy measures, and multi-stakeholder systems. We aim in our approach to bring the various research concepts in evaluation of RS under a single framework

and devise a more inclusive method for both implicit and explicit RS. Furthermore, we explain the opportunity and pitfalls of offline and online experimentation to allow for a more rigorous selection on where to evaluate RS. We invite all researchers to contribute to this model and extend it into their research so that we may collectively achieve a common ground that extends beyond industry walls for RS evaluation.

A number of questions remain for future work. We have presented a framework that can accept any new concepts that might emerge in the field of RS evaluation. We explained beyond-accuracy measures with regards to value based evaluation. However, there would be great interest in correlating commonly used performance based evaluation to value. Furthermore, with multiple stakeholders defining their objectives they are most certainly going to compete at some stage. An interesting topic is competing objectives defined by the various stakeholders involved in the system. Would that impact the evaluation criteria? Would it even be possible to arrive at a single evaluation metric that acts as the common denominator that you would optimise for?

Acknowledgement

This work has been supported by CognativeX.

References

<bib id="bib1"><number>[1]</number>Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. User Modeling and User-Adapted Interaction 30, 1 (May 2020), 127–158. DOI:https://doi.org/10.1007/s11257-019-09256-1</bib>
<bib id="bib2"><number>[2]</number>Himan Abdollahpouri and Robin Burke. 2022. Multistakeholder Recommender Systems. Recommender Systems Handbook (2022), 647–677. DOI:https://doi.org/10.1007/978-1-0716-2197-4_17</bib>
<bib id="bib3"><number>[3]</number>Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Recommender systems as multistakeholder environments. Association for Computing Machinery, Inc, New York, NY, USA. DOI:https://doi.org/10.1145/3079628.3079657</bib>
<bib id="bib4"><number>[4]</number>Panagiotis Adamopoulos and Alexander Tuzhilin. 2015. The Business Value of Recommendations: A Privacy-Preserving Econometric Analysis. (2015).</bib>
<bib id="bib5"><number>[5]</number>Gediminas Adomavicius and Youngok Kwon. Overcoming Accuracy-Diversity Tradeoff in Recommender Systems: A Variance-Based Approach. Retrieved October 17, 2018 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.423.3952&rep=rep1&type=pdf</bib>
<bib id="bib6"><number>[6]</number>Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17, 6 (June 2005), 734–749. DOI:https://doi.org/10.1109/TKDE.2005.99</bib>
<bib id="bib7"><number>[7]</number>Xavier Amatriain and Justin Basilico. 2016. Past, Present, and Future of Recommender Systems. In Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16, 211–214. DOI:https://doi.org/10.1145/2959100.2959144</bib>
<bib id="bib8"><number>[8]</number>Vito Walter Anelli, Alejandro Bellogin, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (July 2021), 2405–2414. DOI:https://doi.org/10.1145/3404835.3463245</bib>
<bib id="bib9"><number>[9]</number>Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. 2013. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In ACM International Conference Proceeding Series, ACM Press, New York, New York, USA, 7–14. DOI:https://doi.org/10.1145/2532508.2532511</bib>
<bib id="bib10"><number>[10]</number>Joeran Beel, Stefan Langer, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, and Andreas Nürnberger. 2013. Research paper recommender system evaluation: a quantitative literature survey. RepSys 20, April (2013), 1–35. DOI:https://doi.org/10.1145/2532508.2532512</bib>
[1]<bib id="bib11"><number>[11]</number>Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Interplay between upsampling and regularization for provider fairness in recommender systems. User Modeling and User-Adapted Interaction 31, 3 (July 2021), 421–455. DOI:https://doi.org/10.1007/s11257-021-09294-8</bib>
<bib id="bib12"><number>[12]</number>Jürgen Buder and Christina Schwind. 2012. Learning with personalized recommender systems: A psychological view. Computers in Human Behavior 28, 1 (2012), 207–216. DOI:https://doi.org/10.1016/j.chb.2011.09.002</bib>
<bib id="bib13"><number>[13]</number>Robin Burke, Himan Abdollahpouri, Bamshad Mobasher, and Trinadh Gupta. 2016. Towards multi-stakeholder utility evaluation of recommender systems. In CEUR Workshop Proceedings.</bib>
<bib id="bib14"><number>[14]</number>Paul B. Kantor Francesco Ricci, Lior Rokach, Bracha Shapira. 2010. Recommender Systems Handbook. DOI:https://doi.org/10.1007/978-0-387-85820-3</bib>
<bib id="bib15"><number>[15]</number>Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. 2014. Offline and online evaluation of news recommender systems at swissinfo.ch. In RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems, Association for Computing Machinery, 169–176. DOI:https://doi.org/10.1145/2645710.2645745</bib>
<bib id="bib16"><number>[16]</number>Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. Proceedings of the fourth ACM conference on Recommender systems - RecSys '10 (2010). DOI:https://doi.org/10.1145/1864708</bib>
<bib id="bib17"><number>[17]</number>Carlos A. Gomez-Uribe and Neil Hunt. 2015. The netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems 6, 4 (December 2015). DOI:https://doi.org/10.1145/2843948</bib>

[1]

<bib id="bib18"><number>[18]</number>Asela Gunawardana and Guy Shani. 2009. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. The Journal of Machine Learning Research 10, (2009), 2935–2962. DOI:https://doi.org/10.1145/1577069.1755883</bib>
<bib id="bib19"><number>[19]</number>Shiqiang Guo, Folami Alamudun, and Tracy Hammond. 2016. RésuMatcher: A personalized résumé-job matching system. Expert Systems with Applications 60, (2016), 169–182. DOI:https://doi.org/10.1016/j.eswa.2016.04.013</bib>
<bib id="bib20"><number>[20]</number>Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. WTF: The Who to Follow Service at Twitter. In [WWW2013]Proceedings of the 22nd international conference on World wide web, 505–514. DOI:https://doi.org/10.1145/2488388.2488433</bib>
<bib id="bib21"><number>[21]</number>Yutian Hu, Fei Xiong, Dongyuan Lu, Ximeng Wang, Xi Xiong, and Hongshu Chen. 2020. Movie collaborative filtering with multiplex implicit feedbacks. Neurocomputing 398, (2020), 485–494. DOI:https://doi.org/10.1016/j.neucom.2019.03.098</bib>
<bib id="bib22"><number>[22]</number>Abdelrahman H. Hussein, Qasem M. Kharma, Faris M. Taweel, Mosleh M. Abualhaj, and Qusai Y. Shambour. 2022. A hybrid multi-criteria collaborative filtering model for effective personalized recommendations. Intelligent Automation and Soft Computing 31, 1 (2022), 661–675. DOI:https://doi.org/10.32604/IASC.2022.020132</bib>
<bib id="bib23"><number>[23]</number>F O Isinkaye, Y O Folajimi, and B A Ojokoh. 2015. Recommendation systems: Principles, methods and evaluation. (2015). DOI:https://doi.org/10.1016/j.eij.2015.06.005</bib>
<bib id="bib24"><number>[24]</number>Dietmar Jannach and Michael Jugovac. 2019. Measuring the business value of recommender systems. ACM Transactions on Management Information Systems 10. DOI:https://doi.org/10.1145/3370082</bib>
<bib id="bib25"><number>[25]</number>Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. 2010. Comparison of implicit and explicit feedback from an online music recommendation service. Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems - HetRec '10 (2010), 47–51. DOI:https://doi.org/10.1145/1869446.1869453</bib>
<bib id="bib26"><number>[26]</number>Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling User Preferences in Recommender Systems: A Classification Framework for Explicit and Implicit User Feedback. ACM Transactions on Interactive Intelligent Systems 4, 2 (2014), 8:1--8:26. DOI:https://doi.org/10.1145/2512208</bib>
<bib id="bib27"><number>[27]</number>Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2018. Unbiased Learning-to-Rank with Biased Feedback. (2018).</bib>
<bib id="bib28"><number>[28]</number>H. A.N. Jungkyu and Hayato Yamana. 2017. A survey on recommendation methods beyond accuracy. In IEICE Transactions on Information and Systems, Institute of Electronics, Information and Communication, Engineers, IEICE, 2931–2944. DOI:https://doi.org/10.1587/transinf.2017EDR0003</bib>
<bib id="bib29"><number>[29]</number>Komal Kapoor, Vikas Kumar, Loren Terveen, Joseph A. Konstan, and Paul Schrater. 2015. "i like to explore sometimes": Adapting to dynamic user novelty preferences. RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems (September 2015), 19–26. DOI:https://doi.org/10.1145/2792838.2800172</bib>
<bib id="bib30"><number>[30]</number>Daniel Kluver, Tien T Nguyen, Michael Ekstrand, Shilad Sen, and John Riedl. 2012. How Many Bits Per Rating? Retrieved October 20, 2018 from http://delivery.acm.org/10.1145/2370000/2365974/p99-kluver.pdf?ip=144.82.8.75&id=2365974&acc=ACTIVE%20SERVICE&key=BF07A2EE685417C5%2ED93309013A15C57B%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1540046965_8712ebd90f997db6484eb5cd21b29b25</bib>
<bib id="bib31"><number>[31]</number>May Fern Koh and Yew Choong Chew. 2015. Intelligent Job Matching with Self-learning Recommendation Engine. Procedia Manufacturing 3, Ahfe (2015), 1959–1965. DOI:https://doi.org/10.1016/j.promfg.2015.07.241</bib>
<bib id="bib32"><number>[32]</number>Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2009. Personalized News Recommendation Based on Click Behavior. (2009), 31–40. DOI:https://doi.org/10.1145/1719970.1719976</bib>
<bib id="bib33"><number>[33]</number>Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. 2015. Recommender system application developments: A survey. Decision Support Systems 74, (2015), 12–32. DOI:https://doi.org/10.1016/j.dss.2015.03.008</bib>
<bib id="bib34"><number>[34]</number>Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. Conference on Human Factors in Computing Systems - Proceedings (2006), 1097–1101. DOI:https://doi.org/10.1145/1125451.1125659</bib>
<bib id="bib35"><number>[35]</number>Denis Parra, Alexandros Karatzoglou, Xavier Amatriain, and Idil Yavuz. 2011. Implicit feedback recommendation via implicit-to-explicit ordinal logistic regression mapping. CEUR Workshop Proceedings 791, 1 (2011). Retrieved October 20, 2017 from http://ceur-ws.org/Vol-791/paper4.pdf</bib>
<bib id="bib36"><number>[36]</number>Luiz Pizzato, Tomek Rej, Thomas Chung, Irena Koprinska, and Judy Kay. 2010. RECON: A Reciprocal Recommender for Online Dating. In Proceedings of the 4th ACM Conference on Recommender Systems (RecSys '10), 207–214. DOI:https://doi.org/10.1145/1864708.1864747</bib>
<bib id="bib37"><number>[37]</number>Shaina Raza and Chen Ding. 2021. Deep Neural Network to Tradeoff between Accuracy and Diversity in a News Recommender System. Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021 (2021), 5246–5256. DOI:https://doi.org/10.1109/BigData52589.2021.9671467</bib>
<bib id="bib38"><number>[38]</number>Paul Resnick and Hal R. Varian. 1997. Recommender Systems. Commun ACM 40, 3 (1997), 56–58. DOI:https://doi.org/10.1145/245108.245121</bib>
<bib id="bib39"><number>[39]</number>Marco Rossetti, Fabio Stella, and Markus Zanker. 2016. Contrasting offline and online results when evaluating recommendation algorithms. In RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems, ACM, New York, NY, USA, 31–34. DOI:https://doi.org/10.1145/2959100.2959176</bib>
<bib id="bib40"><number>[40]</number>Alan Said and Alejandro Bellogín. 2014. Comparative recommender system evaluation: Benchmarking recommendation frameworks. In RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems, Association for Computing Machinery, 129–136. DOI:https://doi.org/10.1145/2645710.2645746</bib>
<bib id="bib41"><number>[41]</number>Guy Shani and Asela Gunawardana. 2015. Evaluating recommendation systems. Recommender systems handbook (2015), 265–308. DOI:https://doi.org/10.1007/978-0-387-85820-3_8</bib>
<bib id="bib42"><number>[42]</number>Zheng Siting, Hong Wenxing, Zhang Ning, and Yang Fan. 2012. Job recommender systems: A survey. 2012 7th International Conference on Computer Science & Education (ICCSE) Iccse (2012), 920–924. DOI:https://doi.org/10.1109/ICCSE.2012.6295216</bib>
<bib id="bib43"><number>[43]</number>Celina Treuillier, Sylvain Castagnos, Evan Dufraisse, Armelle Brun, Celina Treuillier, Sylvain Castagnos, Evan Dufraisse, Armelle Brun, Being Diverse, Célina Treuillier, Université de Lorraine, Université de Lorraine, Evan Dufraisse, Armelle Brun, and Université de Lorraine. 2022. Being Diverse is Not Enough : Rethinking Diversity Evaluation to Meet Challenges of News Recommender Systems. In UMAP '22 Adjunct, ACM, New York, NY, USA, 222–233. DOI:https://doi.org/10.1145/3511047.3538030</bib>

<bib id="bib44"><number>[44]</number>Weinan Zhang, Shuai Yuan, and Jun Wang. 2014. Optimal Real-Time Bidding for Display Advertising. (2014). DOI:https://doi.org/10.1145/2623330.2623633</bib>
<bib id="bib45"><number>[45]</number>Yong Zheng. 2019. Multi-stakeholder recommendations: Case studies, methods and challenges. In RecSys 2019 - 13th ACM Conference on Recommender Systems, Association for Computing Machinery, Inc, New York, NY, USA, 578–579. DOI:https://doi.org/10.1145/3298689.3346951</bib>