Analyzing and Evaluating Data Concerns for Log-based e-commerce recommendation system

Katarina Smolikova, 1528686

***Abstract*—The Log-based e-commerce recommendation system stores the information about user behavior in a datastorage. The service uses this information to automatically train the recommendation system. Personalized recommendation for a user can be then accessed through simple API. The data governance in this scenarion is essential. Not only are the data stored in the service private, if they do not have satisfiable quality the whole service becomes useless. Another important factor to consider is the availability of the data. In this paper, we will not only describe the data concerns, we will also design components for evaluation and utilization of the concerns.**

***Index Terms*—data concerns, data quality, daas, data evaluation**

I. Introduction

IN this paper we will analyze the data concerns for proposed solution for Log-based e-commerce recommendation system. First, there will be data concerns determined. In the next section we will design the components for evaluation of the concerns. Further, we will analyze how to utilize theses concerns. Last but not least, we will introduce one example scenario of how data concerns influence end-user.

II. Determining data concerns

The main data concern in our service is the quality of data. The quality of the recommendations depends heavily on number of previous interactions from the user. This can cause great inaccuracies for customers with low traffic. Although, in the prototype we will simulate a website with high traffic, this might not be the case for all the potential customers. Moreover, the quality of data depends on the ability to assign an event to the specific user, even if they are not logged in. This is usually done through a specific user ID saved in the browser cookie. However, there is a high risk of losing the identificator on the user side. The user can delete the cookie, it can expire or the user can use multiple devices or browsers. All these factors lead to losing the assocciation between the user and their events and therefore the data gained from their previous interactions are useless.

Together with the data quality, we also need to consider the quality of our service. The Log-based recommendation system should provide accurate personalized recommendations. This depends not only on the data quality but also on the speed of data processing. When user makes an action, the data is stored in elasticsearch and resend to PredictionIO to train the recommender. This process needs to be done within milliseconds so that the recommender can provide the user recommended items based on their newest interactions. Moreover, the recommendation algorithm can be tuned and the best possible set of parameters needs to be found to achieve to best possible precision.

Last but not least, the data availability and the service performance are absolutely vital. Since, the recommended items are usually displayed somewhere within the page, the response time of the service should be as low as possible. If there are some delays, the page load time increases. Moreover, if the service is currently unavailable, there are no recommendations displayed and the customer loses possiblity to advertise theire products and therefore their revenues are impacted. Therefore, the service needs to be almost always available and if there needs to be some maintenance done, it should be done in the time of the day, when traffic is lowest.

III. Components for evaluating concerns

All these concerns will be evaluated by a single Data concern evaluation tool, which will be self-developed for the specific purposes of our service. The general pattern for evaluting data concerns is Push, pass-by-values shown in figure 1 Since all the concerns are fairly similar we decided

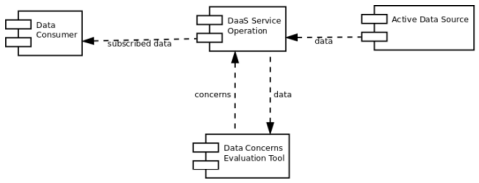

Fig. 1. Push, pass-by-values data concerns evaluation pattern

to evaluate them in the same component. However, they will not be evaluated in the same manner.

The quality of data will be evaluated by observing the number of interactions per user which are sent to the service. For each user ID it will be determined how many interactions are saved. Moreover, average number of events per user will be evaluated. The higher, the number of interactions, the better the data quality. Furhter, we will also track the number of recommendation request for a user without any interaction. This number should be as low as possible.

The quality of service will be evaluted through measuring the performance of the recommendations service. Not only will we track the click-through rate (how many percent of users clicked on the recommended items), we will also track the conversion rate (how many of users purchased a recommended item). Furthermore, we will also track the average time in which the data is processed by the service, i.e the time from when an event happened until the recommendations system recommends items based also on this event. User can for

example buy an item and after the purchase he is shown more recommended items. The recommendations displayed on the page after the purchase should also be based on what items were purchased in the previous step. Therefore, the service needs to process the data very quickly.

The performance of the service will be measure through average response time for a recommendation request. For measuring the availability we will track in how many percents of the cases were the requests unsuccessful.

All these evaluated concerns will be returned through the common user interface. Through, this interface, the owner of the website can track the concerns and utilize them, as will be described in the next chapter.

IV. Utilization of concerns

Utilization of quality of the data is fairly challenging. When the website does not reach high enough traffic, it is extremely difficult to provide good recommendations. However, in this case, it is possible to set up the algorithm to be more item-based. Therefore, if user visits a detail of an item, the recommended items will be similar to the viewed item, rather then on their previous interactions. The challenge of identifying user through cookie can be solved by modifying the registration policy. If for example before each purchase log in would be required, then the percentage of logged in users would raise and the identification of the user would no longer depend on the cookie or browser. This would enable the service to collect more precise data. The quality of service can be increased by changing the algorithm parameters. Finding the ideal setup will lead to better recommendations precision. The speed of data processing can be increased through modifying the middleware which is sending the data from elasticsearch to PredictionIO. The data can be sent in different format or using different technology, if the data concern is too high. The response time and availability can be increased by setting up better infrastructure. It is a great advantage, that all the data concerns in our service can be at least to some extent utilized.

V. Concrete function of data concerns

In this scenario, the end-user is the owner or manager of the website. Most of the data concerns can be utilized directly by the end-user. If the owner properly sets up the algorithm paramterers, he will increase the quality of service without needing to modify it. The quality of the results therefore depends heavily on the ability of the end-user do find the proper parameter setup. This is a great advantage of this service, since the user does not need to pay more to increase the precision. If the user is not satisfied with the response time of the recommendation system, he can modify the infrastructure and use more powerful machines. The other option would be to use the recommendation system asynchronously. If the response from the recommender would be too slow, the content of the webpage would be displayed without the recommended items and the recommendations would be displayed later, when the response from the service arrives. By providing the user single interface for evaluating the data concerns, they have the overview on the service performance and therefore can easily take measures to improve the data governance.

VI. Conclusion

The main data concerns for Log-based ecommerce recommendation system are data quality, service quality, availability and performance. Some of these concerns can be evaluated through measuring the performance of some parts of the service (response time, data processing time etc.). Others can measured by crowds. The main Key performance indicators (KPIs) of the service are click-through rate and conversion rates. The quality of the results of the service can only be measured by the fact, if the recommendations are fitting for the user and that can only be decided by the user alone. The evaluation of the concerns can be done through a simple user interface and should be done regularly by the website owner. When the website owner is aware of the concerns and evaluates them properly, he can take corresponding actions and therefore be able to fully utilize the service.