Crowd-powered recommendation for continuous digital media access and exchange in social networks

FP7-610594

D2.6

Third Reference Framework Release and Evaluation Report

Dissemination level:	Public
Contractual date of delivery:	Month 24, September 30, 2015
Actual date of delivery:	Month 24, September 30, 2015
Workpackage:	WP2
Task:	Task 2.3 Reference framework implementation Task 2.4 Reference framework evaluation
Type:	Prototype
Approval Status:	Draft
Version:	1.0
Number of pages:	22
Filename:	CrowdRec_WP2_D2.6_TUD.pdf

Abstract: This deliverable is a report accompanying the Third Release “Idomaar”, the CrowdRec Reference Framework at Milestone 3 of the project in Month 24. It reports on the larger context in which Idomaar exists, and the difference between Release 2 and Release 3. The root of the difference is the developing needs of recsys vendors, which have extended

beyond evaluation of algorithms, to evaluation of algorithms that are used in infrastructures. The deliverable contains the results of two evaluations, which demonstrate Idomaars ability to evaluate with respect to the 3D model used by CrowdRec. In sum, Idomaar is well positioned to move into the next iteration, which will culminate with its final release at Milestone 4 (M30), and with the NewsREEL 2016 challenge, through which we will promote the use of Idomaar broadly in the recommender system community.

co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	23 September	First draft online	Martha Larson
0.2	26 September	Gravity inputs	Domonkos Tikk
0.3	28 September	Moviri contribute to Section 2	Roberto Turrin
0.4	29 September	Moviri contribute to dataset section	Roberto Turrin
0.5	29 September	NewsREEL input	Till Plumbaum
1.0	30 September	Finalization	Martha Larson

Author list

Organization	Name	Contact Information
Moviri	Roberto Turrin	roberto.turrin@moviri.com
Moviri	Davide Malagoli	davide.malagoli@moviri.com
TUB	Till Plumbaum	Till.Plumbaum@dai-labor.de
TUD	Martha Larson	m.a.larson@tudelft.nl
Gravity	Domonkos Tikk	domonkos.tikk@gravityrd.com

Executive Summary

This deliverable accompanies the Third Release of Idomaar, the CrowdRec Reference Framework (http://rf.crowdrec.eu). It describes the evolving context of Idomaar, and then goes on to specify the differences between the Second and the Third Release. The evolution of Idomaar reflects the constantly developing needs of Moviri and Gravity, the recsys vendors, which represent the users served by Idomaar. Increasingly recommender system vendors must evaluate not just a recommender system algorithm, but the algorithm together with the entire infrastructure. Idomaar has tracked this development over the course of the project.

The deliverable reports on two evaluations carried out on the 30M music and playlist dataset NewsREEL 2015 data set. These evaluations demonstrate the Idomaar makes possible to evaluate algorithms according to the 3D recommender system evaluation model adopted by CrowdRec. In other words, evaluation is possible not only according to user-oriented metrics, but also simultaneously with respect to technical constraints and business metrics.

The main conclusion is that at the time of the Third Release, Idomaar fulfills the requirements that were defined for it. The goals of the next iteration, leading up to the final release, include adding more business metrics, exploring Apache Spark streaming, and integrating additional algorithms. Moving forward in the project, emphasis will be placed on disseminating Idomaar for use in the academic and research communities. The chosen vehicle is the NewsREEL 2016 challenge, which has been accepted to run again at CLEF 2016.

Table of Contents

1 Introduction

In this deliverable, we describe the Third Release of Idomaar, the CrowdRec Reference Framework (http://rf.crowdrec.eu), that took place in M24, at the Third Milestone of the project. This deliverable follows on from the D2.4 "Second Reference Framework Release and Evaluation Report" delivered in M18. Before continuing on to describe the release and evaluation, we devote this introduction to some observations concerning the evolution of the context in which Idomaar is being developed and used.

During Year 2 of CrowdRec, The importance of Idomaar has shifted from evaluating algorithms to demonstrating the feasibility of big data infrastructure for new clients. For recsys vendors, the data of the clients always takes the form of a stream. For this reason, not only big data, but also big data in stream form is essential. The computing environment is composed of infrastructure and algorithms.

Since the beginning of the CrowdRec project, important developments have taken place in terms of the Open Source infrastructure available for the purposes of processing big data. This includes Kafka and Flume. The exploitation partners in the consortium were faced with the necessity of making decisions about whether such resources were appropriate for them to adopt in order to serve the needs of their clients. Idomaar evolved into an essential tool in order to allow the recsys vendors to determine the appropriateness of infrastructures. Practically, Idomaar evaluation goes above and beyond the evaluation of the recommender algorithm itself. Instead, the evaluation is extended to the whole recommender system, which also includes the configuration of the architecture (e.g., the CPUs, the physical memory, the storage, etc.) and the environment (e.g., the operating system, the swap configuration, etc.) it is deployed in.

The necessary tests are directed at questions that are related to infrastructure such as knowing the number of back-end machines needed in order to achieve a certain processing time and response time for a new client.

Since the CrowdRec project start—two years ago—there have been progresses in the recommender systems community, and new solutions have either entered or got popular in the "market", competing with Idomaar as alternative reference frameworks. Among them, it is worthy mentioning RiVal and SCAR. We mention here the relationship of Idomaar to both of these.

The former—*RiVal* (http://rival.recommenders.net/)—is an open source evaluation framework created to evaluate recommendation algorithms in terms of user metrics (e.g., recall, precision, NDCG, RMSE). The main mission of RiVal is to grant *comparable* results. RiVal implements several evaluation strategies and metrics, and is already integrated with existing recommendation framework (e.g., LensKit, MyMediaLite). RiVal was already adopted by Idomaar in the first reference framework release (see D2.2, "First Reference

Framework Release and Evaluation"), and later replaced with other technologies. The main limitations of RiVal are that: it is not designed to work with streamed data and evaluate stream recommendations, but it focuses on static data set; furthermore, it is not scalable and though to work in a distributed environment.

The latter—*SCAR* (http://scar.know-center.tugraz.at/scar/)—is a recent recommender system framework released as Software-as-a-Service (SaaS), which is designed to manage streamed data by implementing highly-scalable technologies suitable to cloud-based environment. SCAR currently supports both user metrics and business metrics (e.g., allowing to test an algorithms under several request loads). Many features of SCAR are common with Idomaar. The key difference is the main goal of the two frameworks. Idomaar is meant to be an evaluation framework suitable for evaluating recommender systems in production-like environment with big streams of data, granting the consistency and reproducibility of results by comparing the functional and non-functional KPIs of different recommendation algorithms, different recommendation frameworks, different environments. SCAR, on the other hand, offers a very scalable recommender system, covering from the algorithms implementation to their evaluation; however, SCAR is not designed to compare, for instance, two recommendation algorithms implemented in two difference frameworks (e.g., MyMediaLite versus Mahout).

2 Reference Framework: Third Release

The Second Release of the CrowdRec Reference Framework at M18—as described in D2.4 "Second Reference Framework Release and Evaluation Report"—focused on all four components of Idomaar. The data container was integrated with FilmTweetings and NewsREEL data sets; the evaluator was re-designed as a two-task process, the former devoted to splitting the data set and the latter to evaluate the recommendation algorithms; the orchestrator was implemented by using recent, state-of-the-art, scalable and distributed open source technologies (e.g., Apache Kafka); finally, the computing environment was released with two communication interfaces: HTTP and Apache Kafka.

This section builds on the Second Release to describe the Third Release. As such, it continues Section 4 "Reference Framework Release" of D2.4 "Second Reference Framework Release and Evaluation Report"(cf. Task 2.3 Reference framework implementation), and also Section 6 "Reference Framework Evaluation" (cf. Task 2.4 Reference framework evaluation). The basis for the Third Release was the Third Iteration of Reference Framework requirements, formulated in D2.5, Section 6 "Idomaar".

2.1 Relation to the Reference Framework Requirements

In this subsection, we describe the connections with the most-recent Idomaar requirements presented in Section 6.3 "Idomaar" of D2.5 "Third Iteration Requirements" (see Section 2.1)

and the progress (see Section 2.2). In this sub-section, we assess the success of the Reference Framework in terms of its fulfillment of the requirements that were established during the requirements engineering process.

We start with the high-level goals of the reference framework, as encoded in the Business requirements of the two recommender system vendors in the consortium, Moviri and Gravity. These are the partners whose business is directly supported by the capabilities of Idomaar. The following table provides information about the relationship of Idomaar, at the moment of the Third Release, to the Moviri and Gravity business requirements (MB and GB).

Requirement (D2.5)	**Third Reference Framework release**
MB05 - Moviri must increase its credibility by comparing their algorithm performance against the competition in a standardized environment.	The evaluation needs of Moviri have moved beyond comparison of algorithms, to assessment of computing environments, which encompass both algorithms and architectures. The central emphasis on architectures is a result of the move to the use of Open Source tools to handle big data in streamed form.
MB09 - CrowdRec must provide Moviri with public evidence of the quality of their algorithms to improve their credibility for prospective clients.	
MB10 - CrowdRec must create a common environment to evaluate algorithms.	
MB16 - Moviri must exploit its academic performances to increase their attractiveness for new young talent.	The Idomaar workshop, the release of the 30M data set at RecSys 2015, and the rising profile of the NewsReel challenge contributes to Moviri academic visibility. All of these activities are connected to Idomaar.
GB07 - Gravity must collect clear evidence on the ROI of their products for both classified ads providers and e-commerce businesses.	Idomaar can serve as a benchmarking platform for Gravity's target clients—its uptake by the company, however, depends on the willingness of the client as well as competitors in case of a tender.
GB08 - Gravity must remain in close contact with state-of-the-art research and innovation projects on recommender	Through Idomaar Gravity can increase its academic visibility and potentially offer researchers to experiment with their algorithms on real-world data live. It

systems technology.	contributes to become actively involved in R&D related to Gravity's area of expertise, and maintains Gravity's innovative status.
GB11 - CrowdRec must develop an evaluation framework that can demonstrate the added value of Gravity's solutions for all customer segments.	The business focus of Gravity moved from TV and media towards online businesses (e-commerce and alike) where integration is not so costly for the client, and is thus a viable investment in the assessment of a new vendor. For this reason, Gravity's need for an external evaluation framework has reduced.

In sum, we see that Idomaar has succeeded in fulfilling the business requirements of Moviri and Gravity. This point is particularly critical, as the business needs with respect to the Reference Framework have shifted for both partners. This shift has meant that the business requirements are fulfilled in

Next, we turn to the technical requirements. The technical requirements are summarized in D2.5 Section 6 "Idomaar" in the requirements table on pp. 106-107. We do not repeat that table here, rather we return to discuss it below. Here, note that the key development in the technical requirements moving from D2.3 to D2.5 was the emphasis on Data sources. Data sources were given higher priority, since the ability to handle streamed data is essential to the recommender system vendors in the consortium. In fact, recommendation services deployed in real systems typically work in environments where the data are dynamic and flood into the system as flow of data. In such environment, data are always handled by queues that manage the concurrency of incoming messages and requests. As Idomaar is being designed as an evaluator to support the comparison of recommender systems in production-like settings, adopting streamed data sources is fundamental to validate its capabilities.

The evaluation capabilities of Idomaar—granting, among the others capacities, the consistency and reproducibility of the results—continue to be a key objective. It is worth noting that the evaluation is not limited to a "standalone" recommendation algorithm, rather it is extended to the whole recommendation system environment. Indeed, the recommender system vendors in the consortium daily face with the integration of recommendation services in their customers' environment, and the evaluation goes far over the generation of recommendations. The whole process is to be tested, and Idomaar is realized so that it can easily integrate into a real environment where the recommendation algorithm is one of the blocks necessary to serve the end users with personalized content. Furthermore, the data in real systems are never static, but they are continuously generated, requiring to manage, for instance, data queues. For such reasons, the whole data flow is to be tested, such as: the ingestion of data, the training of the recommendation algorithms, the

responses to recommendation requests, the update of the existing models, etc. Such priorities have driven all main implementation choices, focusing on highly scalable and distributed technologies for the communication and processing of streamed data, as remarked in the following section.

2.2 Description of the Release

The main updates of the third reference framework release are summarized in the following table, which sums up the key changes of each release (from Release 1 up to Release 3) by component (where "all" refers to the whole framework).

Component	**Release 1**	**Release 2**	**Release 3**
All	• First prototype	• Generalization of the evaluation process, • Adoption of advanced, state-of-the-art technologies.	• Full integration of all components and testing, • Released demo configuration.
Data container	• Defined data format (entities and relations) • Released MovieTweeting dataset	• Released FilmTweeting dataset, • Released NewsREEL 2015 dataset.	• Released 30M music and playlist dataset.
Orchestrator	• File-based communication • Message-based communication (ZeroMQ)	• Testing of stream-oriented, scalable and distributed communication technologies: Apache Flume, Apache Kafka, Zookeeper.	• Consolidated process and data flow.
Computing environment	• Automatic provisioning with Vagrant/Puppet	• Exposure of HTTP and ZeroMQ communication interfaces.	• HTTP interface verified for NewsREEL challenge.
Evaluator	• Simple integration with RiVal (file-based communication)	• Moved to scripting languages: released Python evaluation scripts.	• Evaluation implemented with Apache Spark scripts.

The Third Release of the reference framework focuses on the full integration of all the four components and the consolidation of the flow of data. Extensive bug fixes and tests were made to implement a robust and reliable process.

Furthermore, the HTTP interface, available since the second release, has been verified to work with the communication protocol used by the NewsREEL challenge.

A new dataset has been added to the data container: the 30M music and playlist dataset, described in Section 4. Data sets used in Reference Framework.

Towards the direction of managing large streams of data, also the evaluator adopted a scalable and distributed technology, Apache Spark, in order to process the responses of the recommendation algorithms and verify them against the ground truth data set.

Moreover, this release simplifies the first try by new users. In fact, the third release of Idomaar provides a demo script that covers the full process designed by Idomaar to evaluate recommendation algorithms:

- the orchestrator is started
- the data set is retrieved from the specified location (e.g., github)
- the computing environment is downloaded by Vagrant (if not already done) and started
- the computing environment is provisioned with the required packages, libraries by Puppet (if not already done)
- the recommendation algorithm is started (by the computing environment) and bootstrapped with the training data
- the recommendation algorithm is flooded by recommendation requests (optionally, further training data might be ingested in the meanwhile) on the basis of their timestamp
- the responses to the recommendation requests are stored (together with any required information, e.g., the timestamps)
- finally, the orchestrator invokes the evaluator scripts to evaluate the responses of recommendation algorithm

This demo script, tested in multiple platforms and environments, aims at describing the Idomaar's data flow, demonstrating its capabilities. Furthermore, it represents a starting point for new users to configure custom evaluation, e.g,. changing the dataset, the computing environment, the recommendation algorithms, the evaluation logic, etc.

3 Example Cases of Evaluation with the Reference Framework

3.1 NewsREEL

NewsREEL has been organized under the umbrella of the CLEF Initiative. CLEF seeks to evaluate various kinds of information access systems. NewsREEL had been added to the collections of labs for the 2015 edition. NewsREEL offers two tasks related to news

recommendation. One task is dedicated to a living-lab-style online evaluation while the other focuses on offline evaluation. NewsREEL provides heterogeneous data collected from a variety of news portals. Performance is measured in terms of precision and technical complexity. A total of 42 teams registered for NewsREEL 2015. Of these, 38 teams signed up for both tasks. Figure 1 illustrates the spread of teams around the Globe. Central Europe, Iran, India, and the United States of America engaged most. Network latency may negactively affect the performance in Task 1 of team located far from Europe. Five teams received virtual machines to run their algorithms and alleviate latency issues. In the final evaluation phase of Task 1, we observed 8 actively competing teams. Each team could run several algorithms. Some teams explored a larger segment of algorithms. This led to a total of 19 algorithms competing during the final evaluation round of Task 1.

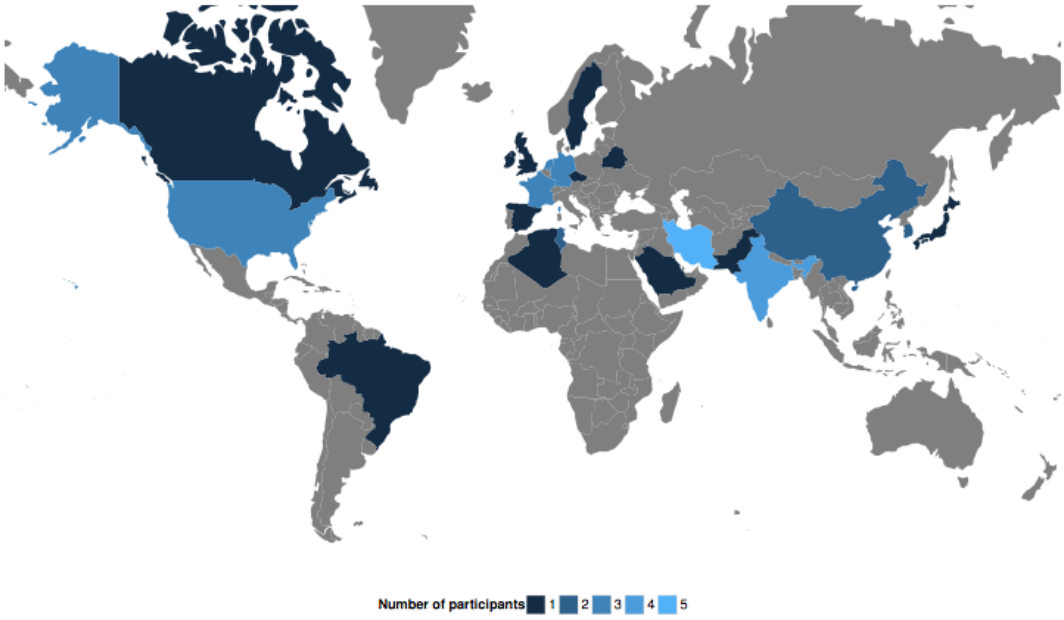


Figure 1: NewsReel participants of CLEF'15

Task 1 ("online evaluation/living lab")

We challenged participants from academia and industry to implement a recommendation service. An actual news recommender system delegates recommendation requests randomly to participants. Thereby, participants face detailed messages describing users' and items' properties. In addition, the system requires participants to respond within 100ms at most. Further, participants have to handle uncertainties regarding identifying users and shifting trends in news consumption. We observed a comparable volume of requests for all

algorithms active for the full evaluation period. These algorithms collected on average ≈ 1000 requests per day. We observed a wide spectrum of approaches contributed by actively engaging participants. Our baselines proved to be hard to beat. They considered a mixture of popularity and recency. Ensembles of several methods accounted for the most successful contributions. Figure 2 illustrates the performance of individual algorithms. We present performance on a plane defined by the number of clicks and requests. A point on this plane refers to a specific CTR. Points' colors refer to the respective team. The teams "cwi" and "riadi- gdl" deployed several algorithms. Two lines depict two CTR levels. A drawn through line marks the 1.0 % level. A dashed line represents the 0.5 % level. The illustration confirms that teams "abc" and "artificial intelligence" outperformed their competitors. More information about the algorithmic challenges can be found in D3.2, in section T 3.3, User Interaction Streams.

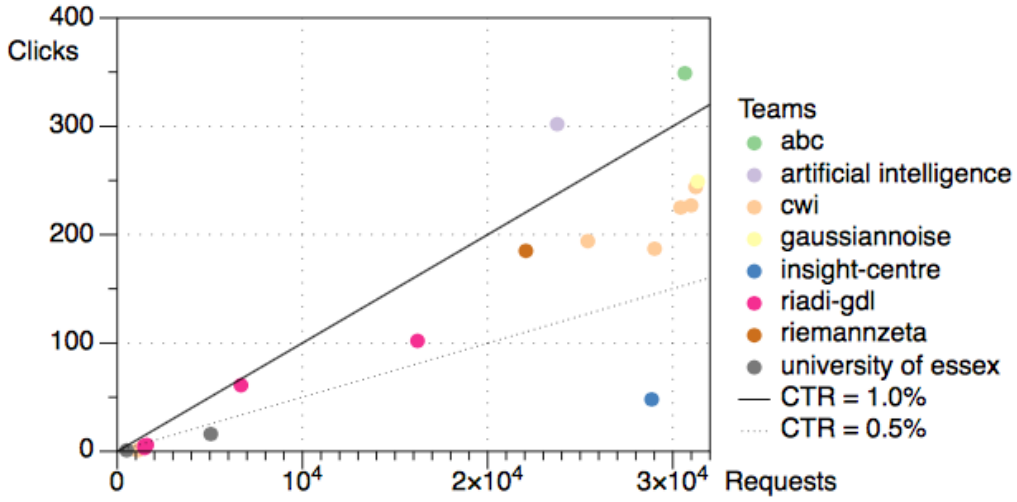


Figure 2: Teams were eligible to run several algorithms simultaneously. We observe some teams operating various recommenders. Teams "abc" and "artificial intelligence" man- aged to achieve a CTR of more than 1 %.

Task 2 ("offline/replay")

We released a comprehensive data set. The data set contains the logs of interactions on several news portals over the period of two months. Participants could download the ~100GB and replay the stream of events locally. The data set used in the offline evaluation has been recorded between July 1st, 2014 and August 31st, 2014 (see Table 1 for more information). A detailed overview of the general content and structure of the data set is provided by Kille et al. 2013 [1]. Thereby, they could observe the precision as well as

complexity of competing recommendation systems. A framework has been provided to support them herein. The framework creates a stream, collects recommendations, and computes evaluation metrics.

Participants reported that they struggled to use the full data set. First, they had to wait until the download finished. Second, the evaluation took long time due to the enormous amount of data. Nevertheless, we need such massive data to ensure significant evaluation results.

	item create/update	user-item interactions	sum
July 2014	618,487	53,323,934	53,942,421
August 2014	354,699	48,126,400	48,481,099
sum	973,186	101,450,334	102,423,520

Table 1: Data set statistics for Task 2.

Figure 3 shows the maximal achievable CTR for the three different domains in the offline data set. The graphs show that the CTR varies highly from day to day. In addition, the graphs show that the average offline CTR for each of the analyzed news portals is specific for each of the portals. This can be explained by the different user groups and the differences in the number of messages per day. Due to the definition of the offline CTR, the expected CTR correlates with the number of messages forwarded as requests to a participant.

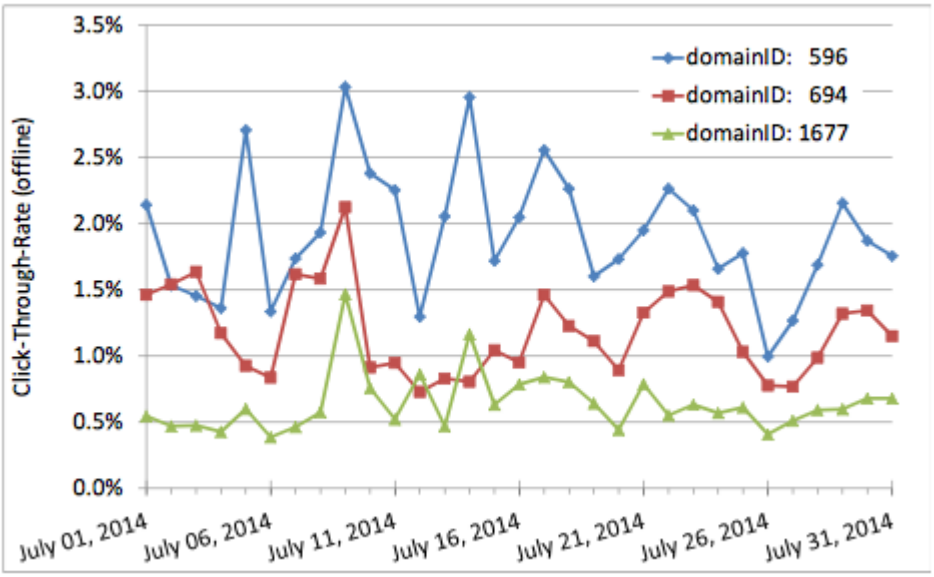


Figure 3: Visualization of the offline CTR for the optimal "recommender algorithm". The optimal recommendation strategy is implemented by looking up the items that will be rewarded by the evaluator. The strategy defines the upper bound of the CTR reachable in Task 2.

The evaluation with respect to scalability focused on maximizing the through- put. Since the teams in the competition used different hardware configurations, the measured results cannot be compared directly. A common optimization objective that has been addressed by the teams working on Task 2 is the effective synchronization of concurrently executed threads.

For the next year, we plan to use standardized virtual machines for the scalability evaluation, ensuring that all teams run the algorithms on exactly the same "virtual" hardware. In order to hide the complexity of building the evaluation environment, we plan to improve the Idomaar framework and facilitate getting started with it.

Relationship with Idomaar Idomaar supports the offline evaluation (task 2). It allows us to determine fine-grained differences between various recommendation algorithms. Using exactly identical settings (data, configuration, virtual machine) yields comparable results. Additionally, Idomaar enables simulating rare conditions such as load peaks. Thereby, we learn to manage such situations.

NewsREEL provides the offline data set in a format compatible to Idomaar. Idomaar requires 5-column tab-separated files as input. We developed NewsREEL-specific components mapping flume-based streams into http-based connections. Hence, participants could easily use their algorithms for both tasks without the need to adjust them.

Advantages of using Idomaar for the NewsREEL evaluation Idomaar enables us defining abstract execution environments. For instance, different virtual machine images let us abstract from operating systems or hardware settings. Additionally, we can customize different software releases (e.g., the java runtime version). Idomaar creates fresh virtual machines thus preventing evaluation instances to conflict. Generally, we could run evaluations on a hosted cloud environment.

3.2 30Music

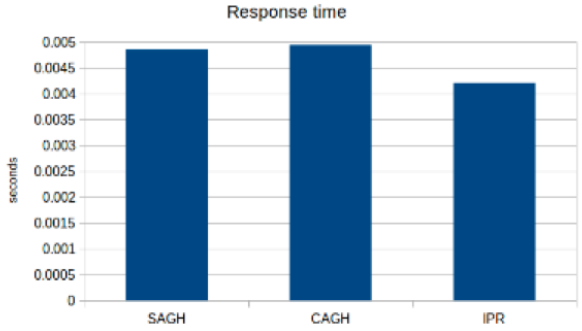
A family of experiments has been executed on the 30M music dataset - released in the current release of the reference framework (see Section 4. Data sets used in Reference Framework) - as reported in the paper “Large scale music recommendation” published in the proceedings of the 2nd Large Scale Recommender System (LSRS) workshop held in Vienna in conjunction with the 9th ACM Recommender Systems conference [3].

The work evaluated three recommendation algorithms proposed for the recommendation of the next song to listen, given the current user listening section. The activity has been led in collaboration with partners not part of the CrowdRec consortium (namely, the DEIB department at Politecnico di Milano). The external collaborators used a cluster composed by 9 Amazon EC2 nodes, manually deployed on Amazon AWS to serve recommendation requests - acting as computing environment. To be observed that the computing environment was not deployed using the Idomaar auto-provisioning technologies (i.e., Vagrant and Puppet), but manually defined using the Amazon AWS console, thus proving the flexibility of the process implemented by the reference framework, as generic as needed to be applied to experiment with custom environments.

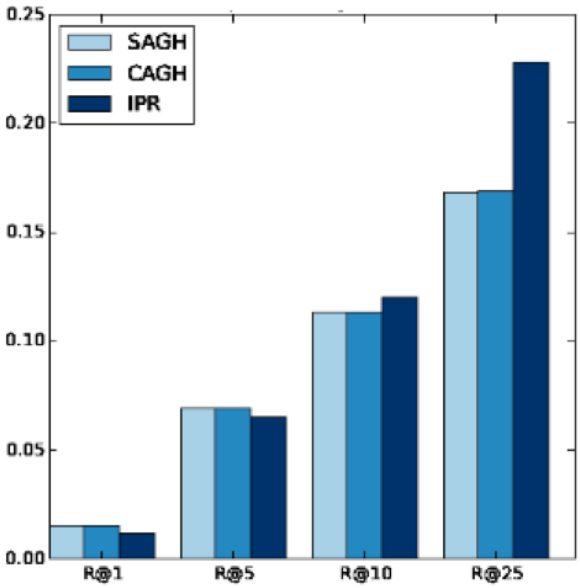
The following figures report the main metrics measured during the activity: the experiment computed both quality metrics - i.e., recall, as percentage of successful recommendations - and non-functional indicators - i.e., training time, as the time to train the recommendation model, and request response time, as the time required to reply to a single recommendation request. We only report in the following an partial extraction of the tests “without repetition”, i.e., where we discard from testing the tracks already consumed in the current user listening session.

Three algorithms have been tested: SAGH (same artist greatest hits), CAGH (Collocated artist greatest hits) and IPR (implicit playlist recommender). Further details can be found in the presented paper [3].

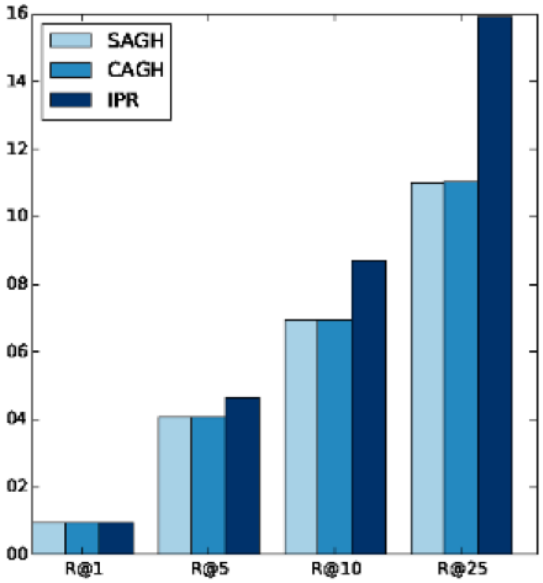
The first two figures below focus on business metrics, and plot, respectively, the training time (in minutes) and the response time of any single recommendation request (in seconds). The latter two figures focus on user metrics and report, respectively, the recall (for different top-N recommendation, where N varies from 1 to 25) of experiments with session length equals to 5 and session length equals to 10.



Business metrics



Recall with session length 5



Recall with session length 10

Comparing the three algorithms among the four figures, we can state that the SAGH and CAGH algorithms have good performance in terms of user metrics (i.e., recall) when the user session length is short (i.e., length equals to 5), when strategies based on popularity, while such KPI downgrades as more user signals (i.e., with session length larger than 5) are collected. On the other side, the IPR algorithm, which provides more personalized solutions with respect to SAGH and CAGH, outperforms the two "popularity-based" baselines as soon as the user listening session collects more than 5 feedbacks. Looking at the business metrics, both baselines SAGH and CAGH have higher training time with respect to IPR, as well as a slightly higher response time to serve recommendations (the main reason is that they do not only use "collaborative" data—i.e., ratings—but also content-based—i.e., the song artists).

4 Data sets used in Reference Framework

In this section, we provide a brief summary and update of the data sets that are used with the data set.

4.1 30M music and playlist dataset

The 30M music and playlist dataset is a collection of listening and playlists data retrieved from Internet radio stations through Last.fm API. Attractive features of the 30Music dataset that differentiate it from existing public datasets include, among the others, (i) the user listening sessions complete of contextual time information, (ii) the user playlists, and (iii) the positive user ratings, key information to experiment with the task of modeling user taste and recommending playlists.

The 30Music dataset has been obtained via Last.fm public API. Last.fm provides free API to track details of user listening sessions. In the case a user has connected his supported player to his Last.fm account, the player scrobbles the user listening activity, i.e., it transfers the play event to Last.fm that records such user interaction.

The dataset is formed by 45K users, 5.6M tracks, 50K playlists, 600K artists, 200K albums, and 280K tags. Relations model links between two (or more) entities. We have 31M user play events, 2.7M user play sessions, and 4.1M user love preferences.

The dataset can be downloaded from http://tinyurl.com/pjlg4tn.

Further details about the dataset can be found in [4] and in the related poster (available at https://goo.gl/fy6vUJ) presented at the ACM RecSys 2015 conference in Vienna on the 15th of September 2015.

4.2 MovieTweeting and FilmTweeting datasets

The MovieTweetings dataset is part of the reference framework since release 1 (see D2.2 "First Reference Framework Release and Evaluation Report", page 30, Section 7.3 "Data

release"). Such dataset was converted from the one originally published by Simon Dooms (https://github.com/crowdrec/datasets/tree/master/01.MovieTweetings) for the ACM RecSys 2014 challenge. Simon Dooms"—at that time PhD student at Ghent University, Belgium"—was notified of our intention to include the dataset into Idomaar and formally agreed for its inclusion in CrowdRec's reference framework.

The dataset is a collection of tweets posted to Twitter by a set of users playing with the mobile IMDb (Internet Movie Database) application. The dataset is available on github repository at https://github.com/crowdrec/datasets/tree/master/01.MovieTweetings. Among the others, the tweets contain the publication timestamp, the rating given by the user on a movie, and the movie IMDb identifier.

The FilmTweeting dataset was introduced in D2.4 "Second Reference framework release and evaluation report" (pag 21, Section 4, "Reference framework release"). It extends the MovieTweeing data set with additional information. More information on these data sets is contained in Loiacono et al. 2014 [5].

4.3 NewsREEL 2015 dataset

The data set was initially released for the CLEF NewsREEL challenge and consists of data about user interactions on news portals (German news portals at the moment). The data set includes both user and item features along with interactions between them. Interactions can be characterized as either clicks (a user clicked on a recommended article) or impressions (a user reads an article). The figure below shows an example of creates and updates.

The NewsREEL data set was mentioned in Deliverable D2.4 "Second Reference framework release and evaluation report" (page 21, Section 4, "Reference framework release") and also reported in Section 3.1. More information about NewsREEL data in general can be found in Kille et al. 2013 [1].

5 Conclusion and Outlook

In this deliverable, we have described the Third Release of Idomaar, and presented test results on the 30Music and NewsReel 2015 data sets. Idomaar has been able to track the evolving needs of the recsys vendors in the consortium, who must carry out evaluation on streamed data, and who also need to evaluate not just the recommender system algorithm, but the algorithm together with the entire infrastructure.

The specific ways in which Idomaar, in its Third Release, allow recommender system vendors to expand their client base are twofold:

- For recsys vendors Idomaar also provides a benchmark tool for testing new technologies. To keep up with the current speed of technological development, established recsys vendor needs to be flexibly and have the ability to quickly evaluate new technology. This is also increasingly important with the increasing emergence of new competitors that are not bound by existing technology and can quickly get in competitive advantage if not a breakthrough technology pops up, and not considered in time by established vendors. Therefore the shift of technology related benchmarking ability of Idomaar is backed up by the real business needs of recsys vendors in the consortium.
- Technological advances and thus, in the long term, the potential decrease of the cost of operation, is also beneficial for customer acquisition and retention. Cost descrease either allows to improve competitive pricing, or less dominantly to apply larger margin.

The final release of Idomaar will be directed at further supporting these goals.

The next steps will follow the roadmap that was set out in D2.5 Section 6 “Idomaar”(page 109), repeated here for convenience.

• M25: Finalization of supported business metrics (in collaboration with T6.3 Business modeling and exploitation)

• M26: Evaluation of whether Apache Spark Streaming can be integrated, and, if the evaluation is successful, proceed with its integration in the reference framework.

• M28: Preparation of Idomaar configurations to support experimentation with popular recommendation frameworks. Such configurations will allow practitioners to play with well-known existing frameworks where several algorithms are already available.

• M29: Final integration of CrowdRec algorithms (focusing on context-aware algorithms, in particular where the context is not explicit, but driven by the current user session).

• M30: D 2.7 Final Reference Framework Release and Evaluation Report (final release of Idomaar with code, configurations, and documentation)

• M36: Bug-fixing/improvement based on the received feedback; dissemination (in collaboration with Task 6.1).

Task 2.4 “Reference Framework Evaluation” will run until M30 and the goal is to determine if there are any further constraints that are limiting the community’s use of Idomaar, or extensions that could be made in order to make it easier for NewsReel participants to use or to increase the uptake. D 2.7 “Final Reference Framework Release and Evaluation Report” is at M30 and will include this information. Then, Task 2.3 “Reference framework implementation” will run until M36, which ensures that it can also react to any new developments that arise within NewsReel.

Our ultimate goal with Idomaar, is to ensure that as large an audience as possible finds its capacity to evaluate stream-based recommendation processes easy-to-use and helpful. We do this by focusing on making functionalities easy to use, rather that including complex functionalities that would be of interest only to a limited audience. In the third year, the NewsReel 2016 challenge is the key vehicle by which we will expand the uptake of Idomaar, both in industry and academia.

6 References

[1] B. Kille, F. Hopfgartner ,T. Brodt, and T. Heintz, “The Plista dataset”, NRS’13: Proceedings of the International Workshop and Challenge on News Recommender Systems, p. 14-21, ACM, October 2013.

[2] B. Kille, A Lommatzsch, R. Turrin, A. Sereny, M. Larson, T. Brodt, J. Seiler, and F. Hopfgartner , “Overview of CLEF NEWSREEL 2015: News Recommendations Evaluation Labs”, Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September, 2015

[3] R. Turrin, A. Condorelli, P. Cremonesi, R. Pagano, and M. Quadrana. “Large Scale Music Recommendation”. Large Scale Recommender System 2015 workshop, in conjunction with ACM RecSys 2015, Vienna, September 2015.
url: http://bit.ly/1KHDTWb

[4] R. Turrin, M. Quadrana, A. Condorelli, R. Pagano, P. Cremonesi “*30Music Listening and Playlists Dataset*”. Poster Proceedings of the 9th ACM Conference on Recommender Systems (RecSys 2015)**,** Vienna, Austria, September 16, 2015.
url: http://ceur-ws.org/Vol-1441/recsys2015_poster13.pdf”

[5] Daniele Loiacono, Andreas Lommatzsch, and Roberto Turrin. 2014. An analysis of the 2014 RecSys Challenge. In *Proceedings of the 2014 Recommender Systems Challenge* (RecSysChallenge '14).