EUROPEAN
SCIENCE
FOUNDATION

Research Networking Programmes

Science Meeting – Scientific Report

The scientific report (WORD or PDF file - maximum of seven A4 pages) should be submitted online <u>within two months of the event</u>. It will be published on the ESF website.

<u>Proposal Title</u>*:** ***Conference Tutorial: Evaluating Recommender Systems – Ensuring Replicability of Evaluation

<u>Application Reference N°</u>: 5515

1) Summary (up to one page)

Recommender systems have become a popular and widely used means of assisting users in online services. Whether for listening to music (Spotify, Last.fm), watching movies and videos (Netflix, YouTube) or interacting with more heterogeneous items (Amazon, eBay, etc.), recommendations are ubiquitous parts of the experience. With the growing expansion of recommender systems in recent years, it comes as little surprise that the recommender system-related research field has grown exponentially, and today most top-tier research venues feature tracks on recommender systems (or closely related). The related industry sector has had a parallel development and many of the currently top-ranked positions in data science list knowledge of recommendation design, techniques and frameworks as a top priority. This unrivaled gain in popularity has led to an overwhelming amount of research being conducted and published over the last few years. With this in mind, it becomes increasingly important to be able to measure different recommendation models against each other in order to objectively estimate their qualities. In today's recommender systems-related literature, the overwhelming majority of papers state what datasets, algorithms, baselines and other potential settings are used in order to (theoretically) ensure replication, many research manuscripts additionally present running times, hardware and software infrastructure, etc. In light of the vast variety of methods to implement and evaluate a recommender system this is a very positive aspect. It should however be critically analyzed in order to ensure some form of advance in the field, i.e. whether the details concerning data, algorithm and general settings are enough to reproduce the same or (more realistically) similar results.

One of the goals in this tutorial is, hence, to gather researchers and practitioners interested in defining clear guidelines for their experimental needs to allow fair comparisons to related work. The tutorial will define and present evaluation metrics, methodologies and experimental configurations used in the literature. We seek to present in a clear way specific guidelines towards reporting experimental results, specifically in the recommendation area. As a particular focus of interest, in the tutorial we will address the main datasets and benchmarking frameworks available, and how they can (and should) be used to improve the research being published in the topic to overcome limitations related with the lack of reproduction and reproducibility of the experiments.

We will additionally present and discuss recent results found in the area where different benchmarking frameworks are compared – using a default experimental setting and also an external, fine-grained controlled experimental setting –, evidencing in a practical view the importance of the evaluation parameters when reporting results. Furthermore, we will introduce an external recommender system evaluation toolkit (*RiVal*, http://rival.recommenders.net) that can be used to create comparable and replicable results.

The following link will lead to the slides and results presented in the talk:
http://www.slideshare.net/abellogin/ht2014-tutorial-evaluating-recommender-systems-ensuring-replicability-of-evaluation

2) Description of the scientific content of and discussions at the event (up to four pages)

The tutorial was planned for a half day split into two presentation parts and an interactive discussion part; however the conference allocated only two hours for the meeting, hence, we removed the discussion from the schedule and focused on the presentation parts. Both parts of the tutorial were given by Alejandro Bellogín, Assistant Professor at Universidad Autónoma de Madrid, in Spain. The first part of the presentation was focused on the basics of recommendation and its evaluation: core recommendation concepts, definition of metrics and methodologies. In the second part, practical aspects about reproducibility and replication related with the evaluation were presented, such as biases arisen from incorrect configurations. In this part, results where different configurations are used were presented, so that discussion could arise about expected against obtained experimental outcomes.

3) Assessment of the results and impact of the event on the future directions of the field (up to two pages)

In general the tutorial was an interesting, engaging event. As it was developed in a related but not too close conference (ACM Hypertext) complementary visions and experiences from other communities (user modeling, user interfaces, privacy, etc.) were discussed. In this sense, it was interesting to observe that in those communities the problem of reproducibility is started being addressed, along with the fact that some of the examples presented from the recommender systems literature could be applied to these other areas. Therefore, the notion of reproducibility and replication can (and should) be extended in the future to these related areas, as well as to others such as contextual search or personalized mobile services. In this way, we envision initiatives driven by these communities where reproducibility would be the focal point; one example of this could be the “Reproducibility IR” track in the next ECIR conference.

The funding of ESF ELIAS has been used for the payment of the lunches and the conference registration of the tutorial speaker.

4) Annexes 4a) and 4b): Programme of the meeting and full list of speakers and participants

Annex 4a: Programme of the meeting

See Part 2 of this report.

Annex 4b: Full list of speakers and participants

Added to the ESF website.