# Towards Real-time Collaborative Filtering for Big Fast Data

Ernesto Diaz-Aviles, Wolfgang Nejdl
L3S Research Center
University of Hannover, Germany
{diaz, nejdl}@L3S.de

Lucas Drumond, Lars Schmidt-Thieme
Information Systems and Machine Learning Lab
University of Hildesheim, Germany
{ldrumond, schmidt-thieme}@ISMLL.de

## ABSTRACT

The Web of people is highly dynamic and the life experiences between our on-line and "real-world" interactions are increasingly interconnected. For example, users engaged in the Social Web more and more rely upon continuous social streams for real-time access to information and fresh knowledge about current affairs. However, given the deluge of data items, it is a challenge for individuals to find relevant and appropriately ranked information at the right time. Having Twitter as test bed, we tackle this information overload problem by following an online collaborative approach. That is, we go beyond the general perspective of information finding in Twitter, that asks: "What is happening right now?", towards an individual user perspective, and ask: "What is interesting to *me* right now within the social media stream?". In this paper, we review our recently proposed online collaborative filtering algorithms and outline potential research directions.

**Categories and Subject Descriptors:** H.3.3 [**Information Storage and Retrieval**]—*Information Filtering*
**General Terms:** Algorithms, Experimentation, Measurement
**Keywords:** Collaborative Filtering; Online Ranking; Twitter

## 1. INTRODUCTION

The collective effervescence of social media production has been enjoying a great deal of success in recent years. The hundred of millions of users who are actively participating in the *Social Web* are exposed to ever-growing amounts of information. The main problem in social streams is not the actual access to the content (e.g., *Twitter* or *YouTube*), but rather to transform this huge mass of data into useful insight. Effective recommender systems techniques, e.g., Collaborative Filtering (CF), are key elements in this context.

One of CF's most successful techniques is low dimensional linear factor models, which assume user preferences can be modeled by only a small number of *latent* factors [4]. The training procedure of latent factor models is performed in a *batch* mode, that is, they assume that all training examples are available before the learning task begins. This means that batch approaches need to be often retrained to cope with the changes in the data over time, which make batch models unsuitable for some real world scenarios where the training instances arrive sequentially and at high speed, as in the case of social web stream applications.

Copyright is held by the author/owner(s).
*WWW 2013 Companion,* May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

Our work deals with a more realistic scenario where observations come from a stream with a high temporal dynamics. One solution to this problem is to resort to online learning methods. In the presence of a continuous stream of incoming tweets, arriving at a high rate, our objective is to process the incoming data in bounded space and time, and recommend a short list of interesting topics that meet users' individual taste.

The high rate makes it harder to: (i) capture the information transmitted; (ii) compute sophisticated models on large pieces of the input; and (iii) store the amount of input data, which we consider significantly larger than the memory available to the algorithm.

This problem setting fits a streaming model of computation by Muthukrishnan [5], which establishes that by imposing a space restriction on algorithms that process streaming data, we may not be able to store all the data we see. The impact is that the data generated in real-time carries high-dimensional information which is difficult to process.

For instance, consider the scenario where topics of interest are captured by the hash-tagging behavior in Twitter. Hashtags are words or phrases prefixed with the symbol #, e.g., *#eurovision*, a form of metadata tag used to mark keywords or topics in a tweet. Hashtags were created by Twitter users as a way to categorize messages, the practice is now a Twitter standard. Any user can categorize or follow topics with the hashtags service. Hashtags evolve over time, reflecting the dynamics of user preferences in the social stream. Our approach seeks to incorporate these dynamics to produce a short list of interesting recommendations based on a matrix factorization model for CF, which is learned online.

In the next section, we give an overview of our approach for online CF.

## 2. ONLINE COLLABORATIVE FILTERING

Recently, we proposed an approach for online CF in the presence of large stream data. Experiments on the *476 million Twitter tweets* dataset [6] show that our online approach outperforms recommendations based on Twitter's *global trend*, and it is also able to deliver a recommendation quality at the level of state-of-the-art matrix factorization techniques for Collaborative Filtering, such as Weighted Regularized Matrix Factorization (WRMF) [3], much faster and more space efficient. Our approach features two important contributions:

- **Online Personalized Ranking based on Matrix Factorization.** In [1], we introduce `RMFO`, a method that creates, in real-time, user-specific rankings for a set of tweets,

based on individual preferences that are inferred from the user's past system interactions. Our novel framework for online collaborative filtering is based on a pairwise ranking approach for matrix factorization in the presence of streaming data.

- **Selective Model Updates for Collaborative Filtering.** In [2], we present RMFX, a novel approach that follows a selective sampling strategy to perform online model updates based on *active learning principles* that closely simulates the task of identifying relevant items from a pool of mostly uninteresting ones. The novelty of this approach lies in a selective sampling strategy to update the model based on personalized small buffers. Our empirical study showed that models updated using the selective sampling approach, significantly outperform online methods that use random samples of the data.

Our methods receive instances from a microblog stream, and update a matrix factorization model following a pairwise learning to rank approach for dyadic data. At the core of RMFO and RMFX is stochastic gradient descent which makes our algorithms easy to implement and efficiently scalable to large-scale datasets. From the RMFO's variants explored in our work, we found that the one using reservoir sampling technique performed the best.

RMFX represents a novel principled approach for online learning from streams. It builds upon the ideas of RMFO and extends them to consider a strategy that selects a subsample of the observed data, based on the objective function gradients, and uses this information to guide the matrix factorization.

We observe that RMFO is simpler to implement, since it does not require a selective model update as RMFX, but RMFO requires more iterations over the reservoir to achieve a competitive recommendation performance, when compared to batch models. We found that RMFX, using a single pass over the interactions captured in the reservoir, achieves a better performance than RMFO with a single iteration (e.g., epoch).

The selection of which approach to use depends on the concrete application scenario. For example, if the time spent by RMFO while learning the model using several iterations does not impact the timeliness of the information recommended, then its ease of implementation can be a strength. On the other hand, when timeliness is compromised, then RMFX becomes a better option.

# 3. FUTURE DIRECTIONS

There are several potential future directions we want to explore, particularly related to information filtering in the presence of highly dynamic data.

- **Better Learning Algorithms for Online Collaborative Filtering.** The success of collaborative filtering heavily relies upon the ability to translate the observed behavior to a meaningful cost function. We strongly believe that the Top-$N$ recommendation task needs to be treated as a ranking problem as discussed in [1, 2]. The exploration of directly optimizing information retrieval metrics for personalized ranking has started and may significantly improve recommendation performance.

- **Prediction of Individual and Collective Behavior in Real-Time Context.** Modeling complex nonlinear dynamics and high-dimensional data, such as social media streams, is an active area of research in machine learning and recommender systems. Many of the existing models, such as matrix factorization and neighborhood based algorithms have been widely used in practice. However, these models are limited in the types of structure they can model. What other methods could potentially capture nonlinear dynamics and also make multimodal predictions handling missing inputs?

- **Real-Time Experimentation.** How to conduct experimental evaluations at large scale in real-time networked settings involving users and their group interactions? A/B testing is a common practice in the industry to evaluate new project features and to support decision making processes, but such evaluations are expensive and time consuming. The exploration of new approaches that align long-term goals with the objective functions optimized by the learning models is an interesting research direction.

- **Integrating Heterogeneous Information Sources.** To improve recommendation quality, one can exploit a number of different information sources like the friendship graph, user demographic information and smartphone sensor data. Various approaches to incorporate specific types of side information exist. However, a general and principled framework that integrates different sources of side information is still missing.

# 4. CONCLUSION

Our research on online collaborative filtering for social media streams provides an example of integrating large-scale recommender systems with the real-time nature of Twitter.

We outlined RMFO and RMFX, approaches for recommending topics to users in presence of streaming data. Our online setting for collaborative filtering captures: "what is interesting to *me* right now within the social media stream", going beyond existing one-size-fits-all solutions.

We have presented several potential research directions, which we believe could lead us to better support users to conduct reliable assessments of dynamics topics on the Web, such as: views on political developments, economic events and crises, as well as pandemics or natural catastrophes.

**Acknowledgments.** Lucas Drumond is sponsored by a scholarship from CNPq, a Brazilian government institution for scientific development.

# 5. REFERENCES

[1] E. Diaz-Aviles, L. Drumond, Z. Gantner, L. Schmidt-Thieme, and W. Nejdl. What Is Happening Right Now ... That Interests Me? Online Topic Discovery And Recommendation In Twitter. CIKM '12, 2012.

[2] E. Diaz-Aviles, L. Drumond, L. Schmidt-Thieme, and W. Nejdl. Real-time Top-N Recommendation In Social Streams. RecSys '12, 2012.

[3] Y. Hu, Y. Koren, and C. Volinsky. Collaborative Filtering for Implicit Feedback Datasets. ICDM'08, 2008.

[4] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques For Recommender Systems. *Computer*, 2009.

[5] S. Muthukrishnan. *Data Streams: Algorithms And Applications*. Now Publishers, 2005.

[6] J. Yang and J. Leskovec. Patterns of Temporal Variation In Online Media. WSDM'11, 2011.