Real-time, Scalable, Content-based Twitter Users Recommendation

Julien Subercaze
Univ Lyon, UJM Saint-Etienne,
CNRS, Laboratoire Hubert Curien
UMR 5516, France

Christophe Gravier
Univ Lyon, UJM Saint-Etienne,
CNRS, Laboratoire Hubert Curien
UMR 5516, France
first.last@univ-st-etienne.fr

Frederique Laforest
Univ Lyon, UJM Saint-Etienne,
CNRS, Laboratoire Hubert Curien
UMR 5516, France

ABSTRACT

Real-time recommendation of Twitter users based on the content of their profiles is a very challenging task. Traditional IR methods such as TF-IDF fail to handle efficiently large datasets. In this paper we present a scalable approach that allows real time recommendation of users based on their tweets. Our model builds a graph of terms, driven by the fact that users sharing similar interests will share similar terms. We show how this model can be encoded as a compact binary footprint, that allows very fast comparison and ranking, taking full advantage of modern CPU architectures. We validate our approach through an empirical evaluation against the Apache Lucenes implementation of TF-IDF. We show that our approach is in average two hundred times faster than standard optimised implementation of TF-IDF with a precision of 58%. The work presented here has been published in The Web Intelligence Journal [1].

CCS CONCEPTS

• **Information systems** → **Social networks**; **Recommender systems**; **Information extraction**;

KEYWORDS

Twitter recommendation, binary footprint, large scale approach, information retrieval, real-time recommendation

ACM Reference Format:
Julien Subercaze, Christophe Gravier, and Frederique Laforest. 2018. Real-time, Scalable, Content-based Twitter Users Recommendation. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3184558.3191587

Microblogging websites such as Twitter produce tremendous amounts of data each second. For instance, Twitter was known to publish an average of 140 millions tweets per day as of march 2011. In a single year, this number has increased up to 340 millions tweets. In this context, we address the problem of building a good user profile model, in order to exploit it in applications such as recommending users to users.

We expect a good user profile model that maximizes the following characteristics:

- Distance-preserving: The distance between users in the user profile space should preserve the interests proximity perceived by users with their peers as much as possible.
- Explainability: Recommender systems should be able to produce human-understandable justifications of recommendations.
- Extensibility: The set of users and their interests evolve with time. However, it should be possible to update the model without complete reprocessing.
- Scalability: Algorithms to compute and compare user profiles should present a complexity as low as possible.
- Real-Time: The profile must be often updated, there its computation must be as fast as possible. These algorithms should also be easily parallelized, to take advantage of the advent of Map/Reduce and related paradigms.

We present a new approach to create and compare profiles of social network users. The solution exploits user-generated contents. In this model, we generate a binary footprint a hash of the user profile that preserves the distance between users profiles in the binary space. Using a binary footprint provides both scalability and parallelization for computing and comparing user profiles. Computing Hamming distance between two hashes is a very fast operation that is computed at the processor level 3 on commodity machines. Moreover, computing newcomers profiles does not require to recompute others profiles.

REFERENCES

[1] Julien Subercaze, Christophe Gravier, and Frédérique Laforest. 2016. Real-time, scalable, content-based Twitter users recommendation. *Web Intelligence* 14, 1 (2016), 17–29. https://doi.org/10.3233/WEB-160329

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191587