Streaming Analysis of Information Diffusion

Extended Abstract

Peter M. Fischer, Io Taxidou
Univ. of Freiburg, CS Department, 79110 Freiburg, Germany
{peter.fischer,taxidou}@informatik.uni-freiburg.de

1 Background and Motivation

Modern social media like Twitter or Facebook encompass a significant and growing share of the population, which is actively using it to create, share and exchange messages. This has a particularly profound effect on the way how news and events are spreading.Given the relevance of social media both as sensor of the real world (e.g., news detection) and its impact on the real world (e.g., shitstorms), there has been a significant work on fast, scalable and thorough analyses, with a special emphasis on trend detection, event detection and sentiment analysis.

To understand the *relevance* and *trustworthiness* of social media messages deeper insights into *Information Diffusion* are needed: where and by whom a particular piece of information has been created, how it has been propagated and whom it may have influenced.

Information diffusion has been a very active field of research, as recently described in a SIGMOD Record survey [GHFZ13]. The focus has been on developing models of information diffusion and targeted empirical studies. Given the complexity of most of these models, nearly all of the investigations have been performed on relatively small data sets, offline and in ad-hoc setups. Despite all this work, there is little effort to tackle the problem of real-time evaluation of information diffusion, which is needed to assess the relevance.

These analyses need to deal with *Volume* and *Velocity* on both on messages and social graphs. The combination of message streams and social graphs is scarely investigated, while incomplete data and complex models make reliable results hard to achieve. Existing systems do not handle the challenges: Graph computation systems neither address the fast change rates nor the realtime interaction between the graph and other processing, while data streams systems fall short on the combination of streams and complex graphs.

2 Goals and Challenges

The goal of our research is to develop algorithms and systems to trace the spreading of information in social media that produce large scale, rapid data. We identified three crucial building blocks for such a real-time tracing system:

1) Algorithms and Systems to perform the tracing and influence assignment, in order to deliver paths that information most likely propagated 2) Classification of user roles, as to provide support for assessing their impact on information diffusion process 3) Predictions on the spreading rate in order to allow estimations of information diffusion lifetime and how representative evaluations on the current state will be.

Our first task is to design, implement and evaluate algorithms and systems that can trace information spreading and assign influence at global scale while producing the results in real-time, matching the volumes and rates of social media. This requires a correlation between the message stream and the social graph. While we already showed that real-time reconstruction of retweets is feasible when the social graph fragment is locally accessible [TF14], real-life social graphs contain hundreds of millions of users, which requires *distributed storage and operation*. Our approach keeps track of the (past) interactions and drives the partitioning on the communities that exist in this interaction graph. Additionally, since the information available for reconstruction is incomplete, either from lack of social graph information or from API limitations, we aim to develop and evaluate methods that *infer missing path information* in a low-overhead manner. In contrast to existing, model-based approaches, we rely on a lightweight, neighborhood-based approach.

Access to diffusion paths enables a broad range of analyses of the information cascades. Given this broad range, we are specifically focusing on features that provide the baselines for supporting relevance and trustworthiness, namely the identification of prominent user roles such as opinion leaders or bridges. An important aspect includes the interactions and connections among users, that lead towards identifying prominent user roles. Our approach will rely on stream-aware clustering instead of fixed roles over limited data.

The process of information spreading varies significantly in speed and duration: most cascades end after a short period, others are quickly spreading for a short time, while yet other group see multiple peaks of activity or stay active for longer periods of time. Understanding how long such a diffusion continues provides important insights on how relevant a piece of information is and how complete its observation is. Virality predictions from the start of a cascade are hard to achieve, while incremental, lightweight forecasts are more feasible. New observations can then be used to update and extend this forecast, incorporating temporal as well as strucutural features.

References

[GHFZ13] Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel A. Zighed. Information diffusion in online social networks: a survey. *SIGMOD Record*, 42(2):17–28, 2013.

[TF14] Io Taxidou and Peter M. Fischer. Online Analysis of Information Diffusion in Twitter. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '14 Companion, 2014.