Ain Shams University
Faculty of Computer and Information Sciences
Computer Science Department

Developing a Predictive Model for Message Propagation on Online Social Networks

A thesis submitted to the department of computer science, faculty of computer and information sciences, Ain Shams university, in partial fulfilment of the requirements for the degree of Doctor of Philosophy in computer and information sciences

By:

Sarah Abdelwahab Ali Elsharkawy

M.Sc. in Computer Science,
Faculty of Computer and Information Sciences,
Ain Shams University.
Cairo, Egypt

Supervised By:

Prof. Dr. Mohamed Ismail Roushdy

Head of Computer Science Department and Former Dean,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Ghada Nasr Ali Hassan

Associate Professor in Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Tarek Mohamed Nabhan

Research Director,
Research and Development Department,
ITWORX

Cairo - 2018

Abstract

In online social networks such as Twitter, tweeting allows users to share a variety of content to their own followers. As tweets are retweeted from user to user, large cascades of tweets propagation are formed. The growth of cascades over time signals the popularity or lack thereof of the subject matter. The k-core of an information graph is a common measure of a node connectedness in diverse applications. The k-core decomposition algorithm categorizes nodes into k-shells based on their connectivity. Previous research claimed that the super-spreaders are those located at the k-core of a social graph and the nodes become of less importance as they get assigned to a k-shell away from the k-core.

A meme represents an idea or a topic that spreads among users of an online social network. Current research on modelling information diffusion in social media focuses on studying retweet cascades of individual tweets independently. However, as a meme spreads, it evolves, and users adopt the meme in varying manners. While retweet cascades can model the propagation of a single piece of information among users, they are not useful in studying the propagation of the whole meme.

In this thesis, we aim to study the information diffusion from a wider perspective where the information propagation of a meme is tracked rather than individual tweets. And also, investigate the influence effect of the super-spreaders, located at the k-core, on the meme cascade growth.

First, the cascade growth of retweet cascades and the various features that govern the growth are studied. We pose the question of whether the same feature set can be used for cascade growth prediction of any dataset on Twitter. Two types of growth prediction are addressed: structural and temporal. First, a definition of structural and temporal growth is devised. Then, an approach to select the best of these features based on the dataset for better accuracy results is proposed. We present and discuss the results of the most discriminating features

in predicting cascades' growth and provide evidence that the pre-selection of features improved the accuracy of the prediction task on the datasets. Moreover, an evidence that the features governing the cascade growth vary from one dataset to another is found.

Next, we generalize the modelling of retweet cascades to a modelling of the diffusion of a meme. To construct the meme adoption graph (MAG), messages related to a meme are identified from the social network stream. Then, a recent clustering algorithm is utilized to automatically extract and cluster tweets. Next, three epidemic cascade construction models are evaluated and compared to construct the MAG and represent a meme diffusion. Also, a set of structural characteristics derived from the MAG that describe the underlying meme adoption pattern are proposed. An empirical study, using four real-world Twitter datasets, is performed to demonstrate the effectiveness of the proposed MAG.

Moreover, we work towards evaluating the influence span of the social media super-spreaders, located at the k-core, in terms of the number of k-shells that their influence is capable of reaching. Our methodology is based on the observation that the k-core size is directly correlated to the graph size under certain conditions. These conditions are explained and the correlation is utilized to assess the effectiveness of the k-core nodes for influence dissemination. The results of the carried out experiments show a high correlation between the k-core size and the sizes of the inner k-shells in the examined datasets. However, the correlation starts to decrease in the outer k-shells. Further investigations have shown that the k-shells, that were less correlated, exhibited a higher presence of spam accounts.

Finally, the effectiveness of using the k-core nodes, as seed nodes, for influence maximisation is inspected. A measure is proposed to estimate the relative strength of the k-core as an influence source among other sources of influence contributing to the cascade development. And, we propose combining that measure along with the correlation between the inner k-core size and the cascade size to determine the influence domination of the k-core nodes, and hence the effectiveness of targeting these specific nodes for influence maximization.

Publications

- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy, "Studying the K-core Influence Dissemination in Twitter Cascades", The International Conference on Artificial Intelligence Applications and Innovations (AIAI). Rhodes, Greece. 2018.
- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy. (2017) "Effectiveness of the K-core Nodes as Seeds for Influence Maximisation in Dynamic Cascades". International Journal of Computers, 2, 187-194.
- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy, "On the Reliability of Cascade Size as a Virality Measure", Proceedings of the European Conference on Electrical Engineering and Computer Science. Bern, Switzerland. 2017.
- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy, "Towards Feature Selection for Cascade Growth Prediction on Twitter", Proceedings of the 10th International Conference on Informatics and Systems (INFOS). Cairo, Egypt. 2016.
- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy, "Modelling Meme Adoption Pattern on Online Social Networks", International Journal of Web Intelligence, 2018. **(Pending)**

Acknowledgements

My deep gratitude, appreciation and sincerest thanks go to Prof.Dr.Mohamed Roushdy, former dean of the faculty of computer and information sciences, Ain Shams University, for his guidance, assistance and advice throughout the thesis.

I would like to express my sincere gratitude to my advisor Dr.Ghada Hassan for the continuous support of my Ph.D study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and of writing this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I also would like to express my deepest appreciation and sincerest thanks to Dr.Tarek Nabhan for proposing the idea of this thesis and providing continuous constructive feedback from an industrial perspective. He is always committed to provide his guidance in a very short time. His fruitful discussions guided me from the first day in this thesis.

Besides my advisors, I would like to thank Prof.Dr.Abdelbadie Salem who provided me with an opportunity to publish a research paper in an international scientific journal.

Last but not the least, I would like to thank my family: my husband, my two beloved daughters and my parents for supporting me spiritually throughout writing this thesis and my life in general.

Contents

List of Tables

List of Figures