Social Data Analysis

ten Thij, M.C.

2018

document version
Publisher's PDF, also known as Version of record

Summary

In recent years, there have been many significant changes in how people communicate and interact with each other, mostly due to social media. This has revolutionised the Internet into a more personal and participatory medium, where social networking is one of the top on-line activities. The abundance of information, that is generated by these on-line platforms, has given rise to a lot of new approaches and fields of study in the research community; such as complex science and computational social science. In this thesis, our aim is to leverage the information that is contained in the aforementioned abundance of available information, which we call social data. Thus, we explore the field of social data analytics. In particular, we study three aspects of this field. The first aspect relates to one of the prominent fields in social data analysis that is driven by the interest in the question: "What drives trending behaviour in on-line platforms?" The second aspect relates to the prediction of future events based on present behaviour and activity; and the third aspect relates to real-time processing of information from on-line platforms.

Our work on trending behaviour focuses on the underlying networks describing the social ties between users of Twitter. Specifically, we consider a graph of users, where an edge means that one of the users has retweeted a message of a different user. Thus, we investigate the dynamics of retweet graphs with the goal to mimic the effects of peaks in activity on Twitter and thereby reveal the nature of trends. Since we want to model trending behaviour without using additional information of the users, we employ random graphs to design a model that captures the development of trending behaviour and mimics the progression of a trend through a (social) network. Therefore, we focus only on the pattern of progression of the trends and not their content. To do so, we define a random graph model, which we call the *retweet graph model* (RGM), that describes the possible changes that can occur to the retweet graph when new messages arrive that mention the topic of interest. In our modelling, we assume that the diffusion pattern of a single message follows a given structure, which we produce using the so-called superstar model. Using the definition of the RGM, we analytically derive the expected value and the variance of both

the average degree and the number of retweets per message.

Furthermore, it is important to know how likely it is that a message is shared frequently. Therefore, we analyse the degree distribution of the RGM. Since the set-up of the model contains several time-dependent aspects, we analyse the degree distribution of a simplified version of the RGM and obtain that high degree nodes are fairly common in this simplified version. After comparing these results to simulations of the RGM, we find that both the simplified version and the RGM itself produce very similar degree distributions.

If a topic is trending, a large group of individuals discusses the topic of interest at a particular time. Knowing how large these groups are and if there is more than one large group in this discussion could provide another helpful insight into the progression of trends in a social network. Because of this, we study the component size distribution of the RGM. We show that the component size distribution of the RGM is either identical or approximated by the bin size distribution of a Pólya process, dependent on the parameters used in the RGM. We also study the size of the largest connected component (LCC) of the retweet graph, proving that, under certain conditions, this LCC is a so-called giant component. Furthermore, we identify four regions of different behaviour of the size of the LCC.

An analysis of multiple retweet graphs indicates that the size of the LCC and its edge density, which is defined as the fraction of edges over nodes, are the most informative characteristics for predicting the peak in activity on Twitter. We discover an interesting phenomenon in the development of the edge density of the LCC, which we call the *densification* of the LCC. This densification captures the phenomenon that disjoint components of the retweet graph join together to form the LCC, which is the first time when several groups find out that other groups have been mentioning the same topic. This perception can then trigger a heightened discussion of the topic, which in turn is perceived as trending behaviour. Finally, we analyse the performance of the model in a predictive setting using data that is acquired from Twitter. After obtaining estimators of the model parameters, we perform simulations of the RGM and conclude that our model captures the density of the LCC and that our model performs better on the data sets that have a single peak in activity rather than a series of peaks.

Studying the activity dynamics for on-line platforms, which is the second aspect we study, has lead to interesting results in the quest to understand human dynamics. In our work, we investigate the activity dynamics on two platforms: Twitter and Wikipedia. As a by-product of our analysis, we show that the global activity pattern follows a circadian rhythm on both these platforms.

We study the influence of promotion on the page-view activity of Wikipedia pages. After comparing several languages (i.e., English, Spanish, Dutch and German), we show that the number of page-views decays exponentially and can be modelled with two parameters. Based on our model formulation, we link our model to the Poisson Process. Furthermore, we present an algorithm that reproduces the diffusion pattern of a tweet to study the diffusion patterns of individual tweets through the network of users of Twitter. Our analysis

shows that the majority of these diffusion patterns have a star-like structure, indicating that a large fraction of the retweets that a message receives are shares of that original message. These results confirm the aforementioned assumption used in our retweet graph model regarding the structure of the diffusion pattern of a single message.

For the third aspect of our work, we address the challenge of obtaining value for the horticulture sector through social media analytics, in which we focus on data that is obtained from Twitter. The horticulture sector is a traditional sector in which growers are focused on production and in which many traders use their own transactions as the main source of information. Tracking how and when the products of the industry are mentioned in a social media feed, is an important addition to current techniques used in the horticulture industry to actively listen to customers.

Our preliminary case studies indicate a correlation between sales data and Twitter mentions of two products and confirm that our data can be used to track the effects of promotional campaigns. These examples show that there is potential for the horticulture to use social data as an additional source of information for the daily business. In our efforts to obtain the information that is contained in a real-time feed, we built a system to obtain tweets that mention products of the horticulture industry in real-time, which we call the *Hortiradar*. This system uses techniques from the field of Natural Language Processing to process the data in real-time and subsequently visualises the data per product. Furthermore, we developed the *storify* algorithm that detects new topics that are discussed in the social media feed in real-time and tracks their development. A comparison of the output of our algorithm with a post-hoc clustering of the same data shows that our algorithm performs as well as if we clustered the data after we obtained all messages. Then, we study how the performance of the algorithm changes when a different set of parameters is used. Based on these results, we obtain a preferred parameter setting that we implement in the Hortiradar.

According to the feedback from our industry partners, the Hortiradar provides a good overview of the current activity on social media with regards to the products of their business. Especially in the case where one of the products is a topic in the main-stream media, the Hortiradar can give a deeper insight into the current public opinion that is expressed on Twitter.