Trend Recognition and Information Quality in Social Networks - an Exposé

Leon Edelmann

March 23, 2017

The proposed study would examine several topics related to social networks and their impact as an information medium.

First subject to be examined is the recognition of newly emerging trends in social media. The study will then try and observe the development of these trends, from the instance of their emergence and throughout the process of their gradual diffusion in a social network. Key players in the spreading of said trends are to be observed and studied. This work will further try and characterize the different aspects of these individuals as to which factors are influential in determining the magnitude of influence of said individuals. This part should further provide answers to questions of the sort *what makes a person a trend setter ? can the influence of a given person be quantified ? can a scheme of rating be applied to a person in terms of the quality of information she spreads ?*

The methodology of trend detection would observe trend recognition techniques such as **bursty topic** recognition using moving-averages as summarized by Guille et al (2013) [1], **tf-idf**, **normalized term-frequency** and **entropy** as discussed by Benhardus and Kalita (2013)[2]. For the purpose of studying data-reliability, **Sentiment analysis** as demonstrated by Go, Bhayani and Huang (2009)[3] is to be explored.

As of today, a large body of literature exists on the subject. This study would provide a survey of the existing state of research as well as examine, compare and experiment with implementing the different proposed techniques on sample data from social networks. A comparative survey in statistical terms, could then be compiled.

A further topic of interest is, the analysis of different aspects of the data obtained from social networks. This part will try to quantify the quality of the information being spread in social networks. The analysis would be conducted in the form of **text mining** and examine the usefulness of the obtained data. For example, classification of gathered facts to categories such as {News, Chatter, Spam }. A further view to be explored is the credibility of the gathered information as users are being exposed to it.

One possible experiment to study the credibility of information could be held in the form of an **electronical survey**. In the first stage, participants will fill out general information about themselves, such as age, gender, level of education, some measure for activeness in social networks and so on. A key-partition would be, erudition in the theme-subject, here being business and economics. The level of learnedness is to be differentiated relying on whether the said person, has had any experience academic, job-related or otherwise in the field. Afterwards, the participants would be presented with 10-15 posts consecutively. Each such post would be drawn at random from an existing pool. This pool would be compromised of posts extracted randomly from social media outlets, however being concentrated around a specific genre in this case business-related schemata. The participants would then be asked to classify each post to a category such as News, Chatter, Spam . Furthermore, each post should be rated in terms of **perceived credibility** on a numerical scale.

The results could then be studied to determine how this newly spread information is being received by members of social networks. Additionally, what factors make a post more/less credible. Possible characteristics of the post could be divided to into categories such as: Characteristics-of-User, Characteristics-of-post, Characteristics-of-topic. Moreover, a further point-of-interest is whether proficiency in the field of economics is decisive in this regard.

As a case study, the micro-blogging platform **Twitter** would be used. The platform offers an API which allows tapping into the Twitters data stream and sending queries to their servers. These services are being offered free of charge to some extent. Namely, it is possible to tap to up-to 1% of the real-time stream of data, which flows through Twitter. These should be more than sufficient for the spectrum of this study. Alternatively, several corpora of previously collected Twitter-Stream-Data are available on the internet.

The bulk of the study is to be conducted on a non-topical general data stream, without regard to any specific topic. With an additional aspect being, narrowing down the scope and attempting at using only data relevant to a certain topic. Naturally a proposed topic would be a business-related theme, which would concentrate on subjects such as the **stock market**, **economic policy**, **innovation** et cetera.

Moreover, as a case study a certain organization could be selected and studied in all above mention regards. This includes but not restricted to: How influential is social media on the organization (measured in stock price), how credible is the information being spread, what characterizes users which influence the public view of the organization.

An additional platform to be considered, depending on ease-of-access, would be **StockTwits**. The latter borrows the concept of Twitter and implements it on a narrower niche, namely the stock market and affiliated spheres. This platform also offers an API would could be used for data queries. Additionally, using StockTwits might prove useful when concentrating on business-related and commercial queries.

References

[1] Adrien Guille, Hakim Hacid, C. Favre, Djamel Abdelkader Zighed. *Information Diffusion in Online Social Networks: A Survey.* SIGMOD record, ACM, 42 (2), pp.17-28, 2013.

[2] Benhardus, J. and Kalita, J.. *Streaming trend detection in Twitter.* Int. J. Web Based Communities, Vol. 9(1), pp.122 - 139, 2013.

[3] Go, A., Bhayani, R. and Huang, L *Twitter sentiment classification using distant supervision.*. CS224N Project Report, Stanford, 1(12), 2009.