Detecting Online Conversations Going Viral

A Master Thesis Plan

Matt Chapman
matthew.chapman@student.uva.nl

Spring 2017, 11 pages

Supervisor: Evangelos Kanoulas
Host organisation: Buzzcapture, http://buzzcapture.com

Universiteit van Amsterdam
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Master Software Engineering
http://www.software-engineering-amsterdam.nl

Contents

Chapter 1

Introduction

1.1 General Project Information

1.1.1 Project Title

"Detection of online discussions going viral"

1.1.2 Student Details

Name: Matthew Chapman
UvA ID: 11403772
E-Mail: matthew.chapman@student.uva.nl
Course: Master Software Engineering

1.1.3 Host Organisation

Buzzcapture
Overtoom 197
1054 HT Amsterdam
The Netherlands
http://www.buzzcapture.com

1.1.4 Contact Person

Name: Wouter Koot
Title: Lead Developer
E-Mail: wouter@buzzcapture.com
Tel: +31 (0)20 320 0377
Mob: +31 (0)63 012 0616

1.2 Project Summary

Buzzcapture provides a tool called BrandMonitor that allows clients to watch in real-time how their brand is being perceived and talked about by the internet-using public. It makes use of scraping various popular social media sites for posts (as well as handling various sources of print media) and carrying out operations such as volume calculations and sentiment analysis. This information is then presented to the client through the BrandMonitor interface.

At this time there is a rudimentary system in place for informing clients that conversations regarding their particular brand are going viral. The current system uses volume analysis over a configurable time parameter, and informs the client if the conversations concerning their brand increase in volume by $\%n$ over the given time. While this approach is functional, Buzzcapture desires to have a *smarter* way of handling this functionality - preferably through predicting the increase in media volume in advance of the volume actually increasing. This will allow for clients to be made aware of these volume increases occurring before they present a problem for their brand.

Approaches have been explored in the past by researchers such as Buntain, Natoli, and Zivkovic - who have applied change detection algorithms to various data sets such as bridge sensors (to identify cracks in the structure before they became visible) and BitCoin market data to predict currency valuations[3]. There is a wealth of mature research available on the subject of change detection algorithms, but they have not (as far as could be found at the time of writing) been applied to data sets specific to social media monitoring and viral conversation detection. This is where I intend to conduct novel and relevant research - investigating if and how these change detection algorithms can be useful for the purposes of anomaly detection in social media analysis.

Chapter 2

Project Planning

2.1 Problem Analysis

As stated in the Project Summary section of this document, the project is being carried out at a company called Buzzcapture, which provides various Software as a Service (SAAS) products to clients for the purposes of monitoring their brand in both online (social) media and in print media. The main project is called BrandMonitor, which serves to provide a dashboard to clients wishing to monitor their media presence and understand how their brand is performing in relation to others.

One of the uses of this product is to allow clients to see when their brand is the subject of a growing conversation or topic. At the time of writing, clients can activate a function in BrandMonitor that notifies them (either via e-mail or SMS) when the volume metric breaks a certain threshold in a certain timescale. For example, a client can opt to be notified when the volume of mentions of their brand doubles over the course of an hour. Buzzcapture considers this functionality to be too basic, and wishes to implement a system that utilises a method of detecting sudden increases in traffic and notifying clients as soon as possible once this occurs.

The problems occur when considering existing change detection algorithms. The first issue here is that these algorithms detect changes, as opposed to predict them. It will be necessary to adapt the results such that the algorithms produce results that are useful and meaningful to the clients of the host organisation. Results that are computed an hour (for example) after the spike are not useful - the clients of my host company require that notifications are sent as soon as possible. The second issue is that at the time of writing, I was unable to locate research to indicate how the various change prediction algorithms perform with regards to social media data. Because post volume data for social networking sites is a simple time-indexed value, it stands to reason that the existing algorithms may well work perfectly fine when applied to this data. However, this cannot be assumed, and it is a research question that I aim to answer in this thesis project.

Data accquired from statistical analysis of social trends and volume is both multi-variate, and parametric. Multi-variate data is data that has several underlying variables that influence how it acts. In the example of social media data, this means that the volume of conversations on a specific topic may be influenced by several factors such as time of day or the availability of information on the topic. Parametric data is data that follows a particular distribution of probability, which is based on a fixed parameter set. Being that social media conversation volume can be considered to have these properties, it can be said that one of several existing parametric change detection algorithms will be adequate for the needs of the host organisation.

The existing research into change detection algorithms provides considerable insight into the situations in which they are best suited for use. For example, Buntain, Natoli, and Zivkovic ran tests on three different change detection algorithms: Galeano and Peña's Likelihood Ratio and CUSUM (Cumulative Sum) tests[6], and Desobry, Davy, and Doncarli's Kernel Change Detection test[4]. These algorithms were applied to various data sets to show, for example, detection of musical note segmentation and denial-of-service detection. While this is an important proof of validity for these algorithms, as stated earlier - it cannot be assumed that these algorithms will adapt well to social media data.

Work has also been carried out by Kifer, Ben-David, and Gehrke on the subject of applying change

detecton algorithms to *data streams* as opposed to static data[8]. Their work confirmed an existing hypothesis in the statistical community: that there is no single approach that could be considered best in all situations. Importantly, they also included in their test harness for the algorithms, a measure for late detections - wherein a change is detected that does not exist within the bounds of the moving window in which algorithms are operating. This is of particular value to me due to the necessity that I am able to detect changes in the data stream swiftly, with a minimum of lag between the event occuring and it being detected.

Detection of abrupt changes: theory and application is a book written by Basseville, Nikiforov, et al., specifically (as the title may imply) on the subject of change detection algorithms. It has proved an invaluable source of understanding so far on the subject, and provides a considerable amount of background on the subject. As it was published in 1993, it is clearly before the time of social media, but nevertheless is still very relevant. It provides an excellent summary of a possible framework for evaluating change detection algorithms[1]:

1. mean time between false alarms;
2. probability of false detections;
3. mean delay for detection;
4. probability of nondetection;
5. accuracy of the change time and magnitude estimates

This will likely prove invaluable in the formulation of a strong test framework in which I can evaluate approaches for change detection in social media data.

Buzzcaptures product is also built on the Elasticsearch platform, which does provide support for calculating moving averages of data, as well as statistical anomaly detection[14]. However, while this may prove an effective solution, it does not provide enough of a basis for a research topic, so will likely be touched upon only in passing, in the final research report.

The other concern in conducting this research is scalability of the various algorithms that I could implement and test. Buzzcapture operates an online SAAS product as their main source of income, and it is important that whatever algorithm performs best on their dataset does not impact the availability of this product to their clients. As such, the algorithm must not only perfom well in terms of accurately determining change points as soon as possible when they occur, but perform the computations quickly. Algorithm performance on data streams (as opposed to static data) will play a very important part in the evaluation of said algorithms.

2.2 Research Questions

- How can one detect spikes or statistical anomalies in social media data?
- How could one detect such spikes or statistical anomalies soon enough for such a detection to be useful?

2.3 Research Method

Buzzcapture have considerable amounts of past data available for analysis. Because old data is available, its behaviour is already known and documented: peaks and troughs in conversation volume exist at known points.

I believe that the best method to test and evaluate the various existing change detection algorithms will be to apply them to this data and see how well large increases in volume are detected, and how soon after a peak starts to build it can be detected.

I will apply change detection algorithms to data that abruptly ends before a peak in volume occurs, and see how soon before this peak the change will be detected by the algorithm. In this way I will

be able to compare the algorithms performance to the existing solution - which is a simple volume comparation over a variable time slice.

I will follow the suggestion from Basseville, Nikiforov, et al. on a test harness, and focus on the following attributes:

1. mean delay for detection;
2. probability of nondetection;
3. algorithm performance when applied to live streaming data in production

The reader will note that there are no metrics in place for the analysis of false detections. I believe that to cover every criteria suggested by Basseville, Nikiforov, et al. would result in a project that would have considerable creep in scope. As the timeline for completion of this research project is fairly short and rigid, it is necessary for me to define a clear focus on attributes that will allow me to complete the research within the time given. After consulting with Buzzcapture, I have deemed the evaluation of algorithms based on metrics concerning false detections to be superfluous - as detection of a change event in the data only results in a notification being sent to a client, there are no negative consequences of which to speak, of a false positive detection taking place.

2.4 Expected Results

My hypothesis is as follows:

No single method will prove more effective than the others tested, given the test harness that I detailed in my research method.

I expect that there will be some deviation in detection delay, probability of non-detection and algorithm performance when I evaluate the approaches to this problem, but I do not believe that there will be significant enough deviation between the algorithms to be able to declare a single one better than all of the others.

2.5 Required Expertise

The following expertise is required for the successful completion of this project:

- Python 2.7
- ElasticSearch
- Java
- MySQL
- Data Science:
  - Change detection/prediction algorithms
  - Statistical analysis of data (and associated mathematical notation)
  - Expression of statistical operations in Python
  - Plotting results using matplotlib
- Understanding of the Buzzcapture toolchain, workflow and products

2.6 Timeline

The following is a time-line for project completion:

- **Begin:** 1st April 2017
- *End of full-time work:* 30th June 2017
- *Final work deadline:* 15th August 2017
- **Final possible defence date:** 31st August 2017

Assuming a 12 week project (full time), I propose that the work is split in the following way:

- **Week 1:** Initial settle-in at Buzzcapture
- **Week 2:** Collation of data for testing
- **Week 3:** Performance of necessary transformations of data
- **Week 4-8:** First implementation of algorithm implementations
- **Week 9:** Development and testing of test harness
- **Week 10:** Generation of first results
- **Week 11-12:** Writeup of results and implementation of algorithm in production

It is assumed that throughout the process detailed above, the final thesis document will be treated as an ongoing task, with various sub-reports and drafts completed during the project.

2.7 Risks

There are a number of risks associated with undertaking this project:

- This area of research requires some understanding of data science - which is a subject not covered in the formal education I have received thus far. To mitigate this I will ensure that I study the required material to gain such an understanding, as well as consulting with researchers or staff (both UvA and Buzzcapture) that have worked in the field.
- The area of research is fairly broad, with a wealth of approaches available to me. To mitigate this, I will need to ensure that my project remains focussed on a well defined subset of the field, with as well constructed timeline for success.
- The host company (Buzzcapture) where this project will be undertaken, works primarily to develop web applications, within the team that I will be working. I will need to ensure that I take sufficient time outside of working hours to cultivate a thorough understanding of the technologies in use to create and maintain the BrandMonitor product.

Chapter 3

Literature Survey

[3]: **Cody Buntain, Christopher Natoli, and Miroslav Zivkovic. "Comparing Algorithms for Detecting Abrupt Change Points in Data". In: *Supercomputing* (2014)**

One of the first papers on the subject I read, this paper discusses and compares three different types of change detection algorithm: the *Likelihood Ratio Test* (LRT), the *Cumulative Sum* (CUSUM) test, and the *Kernel-based Change Detection* (KCD) algorithm. I have also been in touch with one of the authors of this paper, who has kindly provided me with the source code implementations of these algorithms. The algorithms were applied to several data sets, such as historical Bitcoin valuations and data from structural stress sensors.

[7]: **Yoshinobu Kawahara and Masashi Sugiyama. "Change-point detection in time-series data by direct density-ratio estimation". In: *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM. 2009, pp. 389–400**

This is a paper proposing what the authors call a "novel non-parametric change detection algorithm"[7]. It is applied alongside four other approaches, against three different data sets generated by models borrowed from other research papers on time-series data change detection. The approaches are evaluated according to accuracy rate and degree, though not for performance or other relevant metrics.

[8]: **Daniel Kifer, Shai Ben-David, and Johannes Gehrke. "Detecting change in data streams". In: *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment. 2004, pp. 180–191**

This paper, much like the one written by Kawahara and Sugiyama presents a novel change detection algorithm for evaluation. However, this particular paper differs in that the presented algorithm is also useful for estimation of the detected change. They also discuss the use of a 'two window' approach to applying change detection algorithms to data streams, so as to limit memory usage in practice. This part in particular seems relevant to the research I will be undertaking.

[6]: **Pedro Galeano and Daniel Peña. "Covariance changes detection in multivariate time series". In: *Journal of Statistical Planning and Inference* 137.1 (2007), pp. 194–211**

This paper served as the basis of two of the algorithms tested in the paper written by Buntain, Natoli, and Zivkovic This paper also provides its own evaluations of the approaches contained therein.

[4]: **Frédéric Desobry, Manuel Davy, and Christian Doncarli. "An online kernel change detection algorithm". In: *IEEE Transactions on Signal Processing* 53.8 (2005), pp. 2961–2974**

Desobry, Davy, and Doncarli write concerning an implementation of a Kernel Change Detection (KCD) algorithm that is considered *online* - that is, applicable to moving and updating data sets such as those that I shall be working with at Buzzcapture. This is particularly important, as the approach they take to change detection is considered against data sources that could be considered similar to those that I will be working with.

[13]: **Alexander G Tartakovsky et al. "Detection of intrusions in information systems by sequential change-point methods". In: *Statistical methodology* 3.3 (2006), pp. 252–293**

This paper discusses applying a CUSUM (Cumulative Sum) approach for detection of network intrusions. The approach taken by Tartakovsky et al. was intended to research the possibility of change detection while maintaining a low rate of false alarms. The idea was to apply an algorithm for this purpose that would not need to take into account pre-change and post-change models in order to be effective.

[12]: **Alexander G Tartakovsky, Boris L Rozovskii, and Khushboo Shah. "A nonparametric multichart CUSUM test for rapid intrusion detection". In: *Proceedings of Joint Statistical Meetings, Minneapolis, MN*. Citeseer. 2005**

A collection of papers and discussions by the same authors as the above papers, expanding somewhat on the problem they tackled and the solutions found.

[10]: **David S Matteson and Nicholas A James. "A nonparametric approach for multiple change point analysis of multivariate data". In: *Journal of the American Statistical Association* 109.505 (2014), pp. 334–345**

Discussion of an 'offline' (that is, applying an algorithm to a fixed data-set as opposed to processing moving 'live' data) approach to change detection in multi-variate data. The authors carry out a simulation study to compare various approaches to this problem and present their results.

[11]: **David Siegmund and ES Venkatraman. "Using the generalized likelihood ratio statistic for sequential detection of a change-point". In: *The Annals of Statistics* (1995), pp. 255–271**

A paper on using a *Generalized Likelihood Ratio* test for the detection of change points in a data set. An old piece of text that pre-dates social media, yet is still useful in explaining how the GLR test can be used for the detection of change points. The GLR approach is compared against standard CUSUM tests.

[15]: **Alan Willsky and H Jones. "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems". In: *IEEE Transactions on Automatic control* 21.1 (1976), pp. 108–112**

Much like the paper written by Siegmund and Venkatraman, this paper pre-dates social media as a data source to which change detection algorithms could be applied. It is however, a useful look into how change detection algorithms have progressed and improved over the years. This paper is primarily focussed on uses of change detection in automated controls (being that it was published in the IEEE Transactions on Automatic Control), but is still a useful resource to understand how change detection algorithms can be evaluated.

[9]: **Tze Leung Lai and Jerry Z Shan. "Efficient recursive algorithms for detection of abrupt changes in signals and control systems". In: *IEEE Transactions on Automatic Control* 44.5 (1999), pp. 952–966**

Another paper concerning change point detection in control systems, Lai and Shan write primarily regarding Generalised Likelihood Ratio tests, and how these can be configured, specifically with regards to window size.

[2]: **Sotiris Bersimis, Stelios Psarakis, and John Panaretos. "Multivariate statistical process control charts: an overview". In: *Quality and Reliability engineering international* 23.5 (2007), pp. 517–543**

Control charts are a mechanism used in various industries to decide whether a given process is 'in control' or 'out of control'. In this way, control charts are used to discover outlying or anomalous results in a given data set, or inform operators of a sudden change in the data. This makes control charts particularly relevant to my research. In fact, one source I have mentioned earlier in this document, [14], talks specifically about implementing an effective control chart natively in ElasticSearch.

This particular paper discusses various approaches to control charts, and the methods behind their operation.

[5]: **Allen B Downey. "A novel changepoint detection algorithm". In: *arXiv preprint arXiv:0812.1237* (2008)**

Downey discusses his creation of a 'novel change detection algorithm' that can also be used for "...predicting the distribution of the next point in the series."[5]

He discusses in some detail the difference between online and offline change detection algorithms, and compares his implementation of a new algorithm with implementations of existing and established algorithms.

Bibliography

[1] Michèle Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*. Vol. 104. Prentice Hall Englewood Cliffs, 1993.

[2] Sotiris Bersimis, Stelios Psarakis, and John Panaretos. "Multivariate statistical process control charts: an overview". In: *Quality and Reliability engineering international* 23.5 (2007), pp. 517–543.

[3] Cody Buntain, Christopher Natoli, and Miroslav Zivkovic. "Comparing Algorithms for Detecting Abrupt Change Points in Data". In: *Supercomputing* (2014).

[4] Frédéric Desobry, Manuel Davy, and Christian Doncarli. "An online kernel change detection algorithm". In: *IEEE Transactions on Signal Processing* 53.8 (2005), pp. 2961–2974.

[5] Allen B Downey. "A novel changepoint detection algorithm". In: *arXiv preprint arXiv:0812.1237* (2008).

[6] Pedro Galeano and Daniel Peña. "Covariance changes detection in multivariate time series". In: *Journal of Statistical Planning and Inference* 137.1 (2007), pp. 194–211.

[7] Yoshinobu Kawahara and Masashi Sugiyama. "Change-point detection in time-series data by direct density-ratio estimation". In: *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM. 2009, pp. 389–400.

[8] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. "Detecting change in data streams". In: *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment. 2004, pp. 180–191.

[9] Tze Leung Lai and Jerry Z Shan. "Efficient recursive algorithms for detection of abrupt changes in signals and control systems". In: *IEEE Transactions on Automatic Control* 44.5 (1999), pp. 952–966.

[10] David S Matteson and Nicholas A James. "A nonparametric approach for multiple change point analysis of multivariate data". In: *Journal of the American Statistical Association* 109.505 (2014), pp. 334–345.

[11] David Siegmund and ES Venkatraman. "Using the generalized likelihood ratio statistic for sequential detection of a change-point". In: *The Annals of Statistics* (1995), pp. 255–271.

[12] Alexander G Tartakovsky, Boris L Rozovskii, and Khushboo Shah. "A nonparametric multichart CUSUM test for rapid intrusion detection". In: *Proceedings of Joint Statistical Meetings, Minneapolis, MN*. Citeseer. 2005.

[13] Alexander G Tartakovsky et al. "Detection of intrusions in information systems by sequential change-point methods". In: *Statistical methodology* 3.3 (2006), pp. 252–293.

[14] Zachary Tong. *Staying in control with moving averages, part 1*. Online. Aug. 2015. URL: https://www.elastic.co/blog/staying-in-control-with-moving-averages-part-1.

[15] Alan Willsky and H Jones. "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems". In: *IEEE Transactions on Automatic control* 21.1 (1976), pp. 108–112.