Anomaly detection for internet surveillance

Henri Bouma *, Stephan Raaijmakers, Arvid Halma, Harry Wedemeijer

TNO, PO Box 96864, 2509 JG The Hague, The Netherlands.

ABSTRACT

Many threats in the real world can be related to activity of persons on the internet. Internet surveillance aims to predict and prevent attacks and to assist in finding suspects based on information from the web. However, the amount of data on the internet rapidly increases and it is time consuming to monitor many websites. In this paper, we present a novel method to automatically monitor trends and find anomalies on the internet. The system was tested on Twitter data. The results showed that it can successfully recognize abnormal changes in activity or emotion.

Keywords: Anomaly detection, internet surveillance, data mining, forensic, cybercrime, pattern recognition.

1. INTRODUCTION

Many threats to people or infrastructures in the real world can be related to the activity of persons on social media, blogs and forums. While the exact causal relationship is yet largely unknown, internet surveillance aims to prevent possible attacks and assist in profiling suspects based on information from the cyber space. The information is gathered to improve situation awareness for the protection of citizens and critical infrastructures. However, the amount of data on the internet rapidly increases and it is time consuming to monitor the continuous flow of tweets, posts and announcements on websites.

In this paper, we present a novel method to automatically monitor trends and detect abnormal behavior on Twitter or other social media. Specifically, we propose a profiling system based on the correlation analysis of a number of variables, such as sentiment, retweets, post frequencies and properties such as senders, medium and receiver. Correlation patterns between these variables are used for the profiling of network snapshots and individuals.

In addition, we analyze pictures and combine the information from text and images to improve the overall system. The system was tested on Twitter data. The results show that our system can successfully recognize abnormal changes in behavior, activity or emotion.

2. BACKGROUND

How to find a needle in a haystack if you do not know what the needle looks like? This is the problem of several governmental and commercial stakeholders in the security domain that are confronted with huge amounts of data. There is so much data that the useful information (‘something’ deviant) is not or hard to find. Today the challenge is no longer accessing open source information but in the analysis of the huge amount of data [3].

* henri.bouma@tno.nl; phone +31.88.866.4054; http://www.tno.nl

Henri Bouma, Stephan Raaijmakers, Arvid Halma and Harry Wedemeijer, “Anomaly detection for internet surveillance”, Proc. SPIE, vol. 8408, 840807 (2012); http://dx.doi.org/10.1117/12.918573

The needs within the field of internet surveillance and forensics is to find individuals and groups efficiently in large amounts of data based on features, to separate characteristic features from other features, to visualize related matters, to detect abnormal behavior and to monitor trends and changes. TNO develops tailored technology to support surveillance tasks as good as possible.

There is special attention for the following criteria: impact, reusability and multi-modality. We are looking for solutions that supports the workflow of users, fulfilling the needs mentioned, and proofing the concept with a demonstrator. We choose to develop techniques that are reusable now and in the future for multiple stakeholders to satisfy their needs. Our aim is to show that the combination of information from different modalities can bring additional knowledge, by not only using text but also image-based content.

The internet-surveillance technology can be used to detect criminal activities and terrorism, but there are privacy concerns [10]. The privacy and proportionality aspects should not be forgotten in this process [5], to avoid undesired situations for security or economic purposes [11].

The main research question for this project is: in what way can developments in the field of information analysis techniques (e.g., video content analysis, data mining, etc.) contribute to the improvement of the information process for forensics and law enforcement.

As data source we have chosen (a few sites on) the internet for the following reasons:

1. The internet is a free and publically accessible open source.
2. It is an enormous multi-modal database that contains a huge amount of data (a haystack).
3. To avoid database specific implementation and development problems and focus on the development of reusable modules.
4. Because several stakeholders are interested in techniques for internet surveillance and open source intelligence (OSINT).

To be more specific, we have developed tools to analyze messages on Twitter: tweets.

For a number of years, authorities have been using Twitter and other social media to observer and intervene in crisis situations (such as uproars, or major incidents). The social media, such as Twitter, appear to respond more rapidly to problems than more traditional sources [19] and can be used for early detection of diseases and threats [8]. In 2009, the British West-Midlands Police used both Twitter and YouTube during a demonstration by the far right England and Wales Defence League in Birmingham, informing demonstrators about eventual illegal activities during the demonstration. Similarly, the Cambria police in Pennsylvania USA have been known to regularly use Twitter to inform the public about criminal events. Yet, fundamental insight into effective intervention strategies in online media that exceed plain information campaigns appears yet to be developed. On the positive side, a lot of research has been devoted in the last years to the analysis of information flows on social media, motivated amongst others by a desire for effective content spreading mechanisms, e.g. for online campaigns marketing purposes. While the domain differs from crisis intervention, some of the mechanisms for launching effective marketing campaigns would seem relevant for crisis management and intervention as well. Dabeer et al. [9] analyze the effectiveness of information campaigns on Twitter by taking into account the retweet behavior of active followers of an information source. They are able to derive a probalistic response model by computing optimal time slots for campaign tweets. Information diffusion on Twitter is addressed in Yang and Counts [28], who find that the mention rate of a person (the times people respond to a certain user) on Twitter is a stronger predictor of the spread of information pertaining to that user than message-internal properties, such as the timing of the message with respect to an emerging topic, or the presence of a hyperlink in the message. This implies that active participation in discussions may be a fruitful strategy to get new messages noticed. Schultz et al. [23] have investigated the role of social media for crisis communication, addressing blogs, twitter and traditional media such as newspapers. They find that the medium matters more than the message, and specifically that Twitter users are more likely to share crisis information, whilst often making reference to offline traditional media with a high perceived authority, such as national newspapers.

The general processing chain we developed to handle Twitter information consists of multiple layers: source data, features, analysis and knowledge (see Figure 1). The data on the internet comes in several forms: such as text, meta data

and images. Features can be computed on each of these forms. Information analysis includes both the computation of these (low-level) features and the (higher-level) combination and analysis to obtain knowledge.

Information analysis on the internet can focusses on changes in sentiment, radicalization, threads, abnormal changes (anomalies), trends [17][20], what is 'hot', track developments and predict the near future, relationships and authorship.

The central working hypothesis of our work is that automatic information analysis techniques improve the efficiency of the forensics and law enforcement process. We have focused on two aspects: anomaly detection and image analysis. Anomaly detection focusses on the detection of abnormal behavior on the internet (Sec. 3.1). Image analysis aims to show that the combination of text and image-content results in additional knowledge (Sec. 3.2).

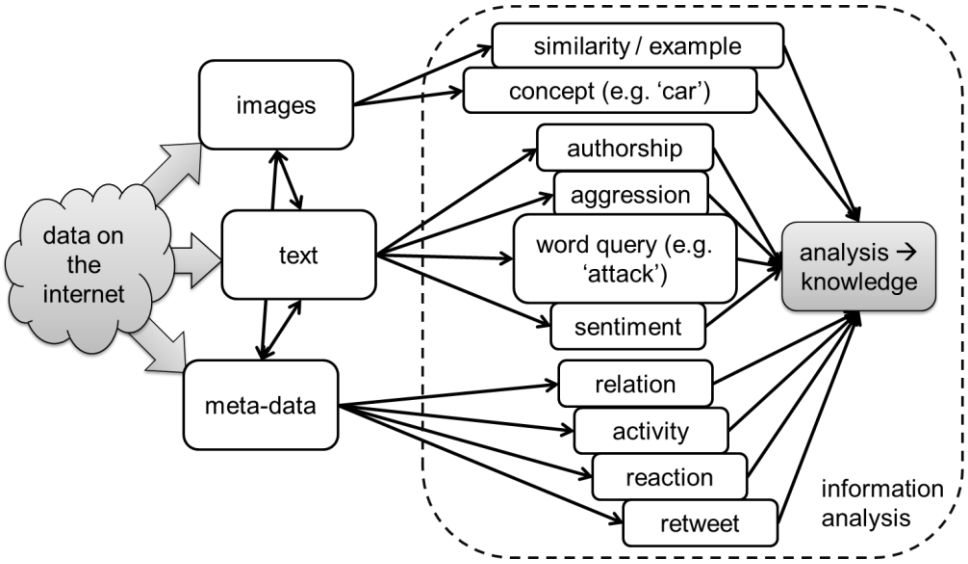

Figure 1. Information analysis on the internet is based on multi-modal sources (e.g. text and images) and it includes the analysis of features in order to obtain knowledge.

3. ANOMALY DETECTION

3.1 Anomaly-detection method

The term 'anomaly detection' usually describes the automated detection of sudden, unexpected changes in data, often consisting of temporally ordered observations. This can focus on identifying anomalous sequences with respect to a database of normal sequences, identifying an anomalous subsequence within a long sequence, or identifying a pattern in a sequence whose frequency of occurrence is anomalous [7]. Applications of anomaly detection are widespread and vary from intrusion detection in computer networks [31], credit card fraud detection [30], medical applications such as EEG analysis [29], to forensic applications such as the detection of abnormal behavior in surveillance videos [2]. Major challenges consist of sieving out the usually infrequently occurring anomalies from the total (often massive) data, and handling the problem of adaptation: certain (but not all) events that primarily occur as anomalies become accepted as 'normal' events, and should not be detected over and over again. Yet, the self-calibration of the system should not be fully autonomous and unsupervised, as this would clearly lead to security breaches in a number of application domains. Preferably, therefore, a human is in the loop that labels the signals produced by the system as true positives or false positives, identifies false negatives, and labels true positives and false negatives as persistent anomalies (not leading to calibration of the system) or non-persistent. The various approaches to anomaly detection can roughly be partitioned according to the schema in Figure 2, which is based on the work of Chandola et al. [6].

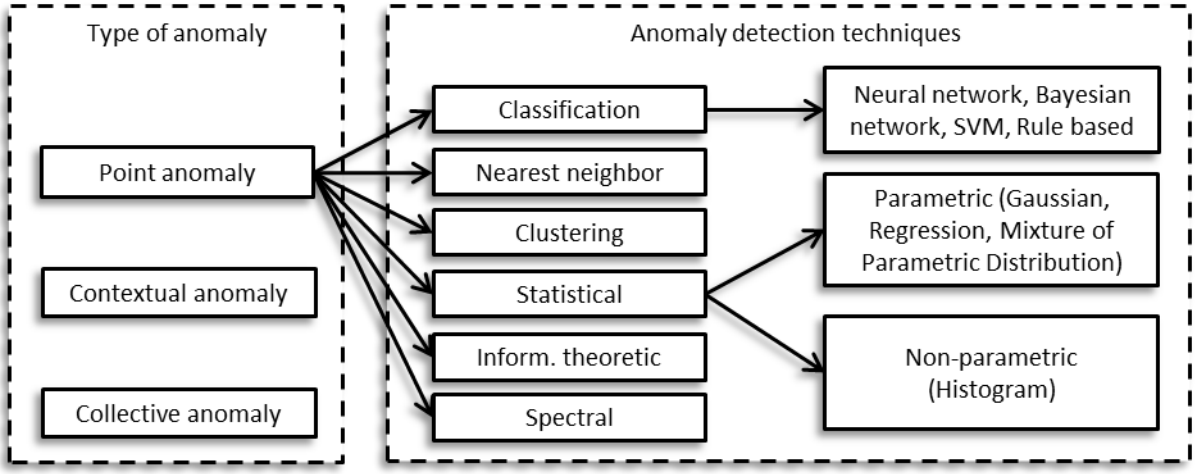

Figure 2: A schema for several types of anomaly detection (based on the work of Chandola et al. [6]).

According to this schema, anomaly detection either concentrates on single data points (single observations, irrespective of context) or contextual data (sequences, time series), and can proceed in an either online fashion (adapting the notion of 'normality' on the fly) or as a distributed system, combining several domain-specific detectors. The schema does not express that both contextual/collective and point anomaly detection can proceed in online and distributed manners.

A complicating factor in measuring anomalies in any kind of data is that it is not trivial to define the concept of anomaly. Usually, data constitutes a complex event space, with many interacting trends and concepts. For instance, measuring anomalies in Twitter data can be performed by looking at post behavior, changes in network structure (follower-followee relations, suspension or (re)activation of users, new channels, new reply-behavior, textual content, textual style, sentiment, temporal issues, etc. It would seem natural to decompose this type of entanglement into different feature spaces, and to detect anomalies separately, combining them afterwards to an aggregated anomaly score. This process of aggregation is by no means clear: certain anomalies will probably be of lower significance than others. The weighting of anomalies can however only be achieved in supervised settings, where humans label certain events as being anomalous or not.

For anomaly detection, a processing chain has been created that consists of data collection, data storage, data enrichment, anomaly detection and visualization (Figure 3).

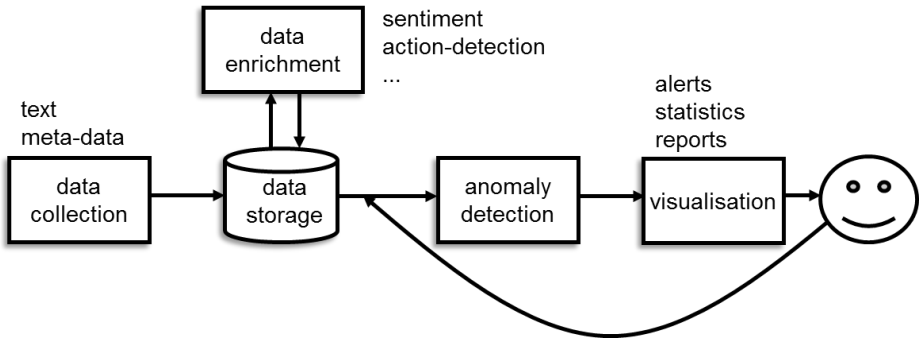

Figure 3. Processing chain for anomaly detection.

To develop and validate on relevant data, a tool was made that collects both the past and coming period of Twitter-data. This allow the analysis of an incident after it occurred and helps to see whether it could have been predicted.

The detection of anomalies is challenging for several reasons: anomalies are rare, some anomalies become normal after a while, the border between normal and deviant is sometimes thin, what is normal in one context may be abnormal in

another context, negation and irony, and the data can be viewed from multiple perspectives (what is the best context). Therefore we have chosen to design a system that is not a 'black box', but allows the analyst to define the measured features, after which the system studies the interaction between the variables [15]. We propose a anomaly-detection system based on non-negative matrix factorization and the correlation analysis of a number of variables, such as sentiment of tweet and retweet, number of reactions or retweets, number of followers, post frequencies and topological properties of senders, medium and receiver and an estimation of violence. Correlation patterns between these variables are used for the profiling of network snapshots and individuals.

3.2 Experiments and results

The data exists of text and meta data, such as activities and relations. Furthermore, features are added based on text interpretations, such as sentiment analysis (Figure 1). Using non-negative matrix factorization and a Dutch sentiment (subjectivity) lexicon [16], clusters partitioning the tweets in clearly subjective or more objective groups can be computed. An anomaly is recognized as an abrupt change of the distribution of the two clusters. Figure 4 shows a sudden change (mainly consisting of an outburst of positive sentiment) corresponding to Feyenoord scoring during a soccer match.

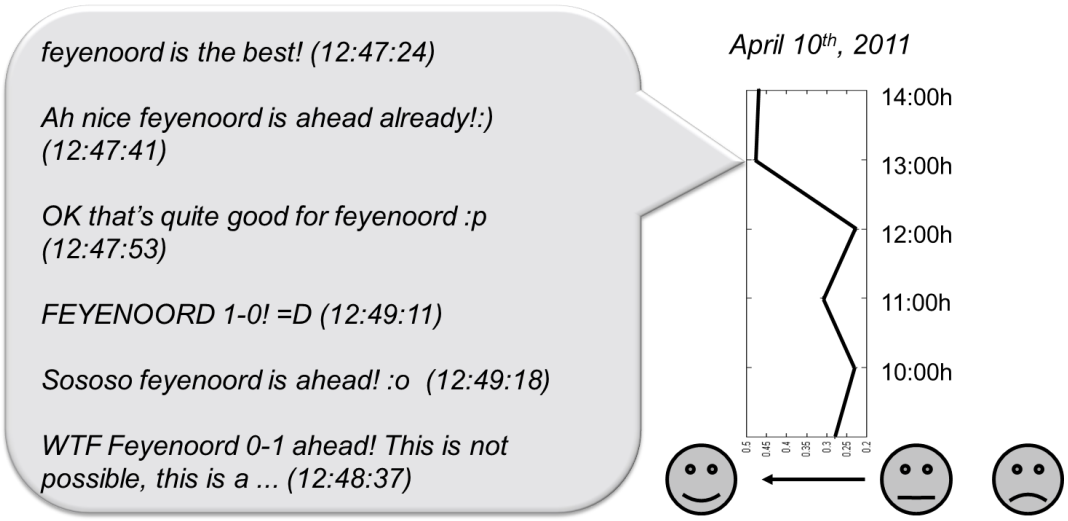

Figure 4. Example of sentiment analysis. 'Feyenoord' is the name of a soccer team in Rotterdam, The Netherlands. Tweets have been translated from Dutch to English for the illustration.

Another example is a riot between Turkish and Kurdish people in Amsterdam in October 2011. The tweets show that there is a large increase in activity (e.g., by retweets and reactions ('mentions')) in this context.

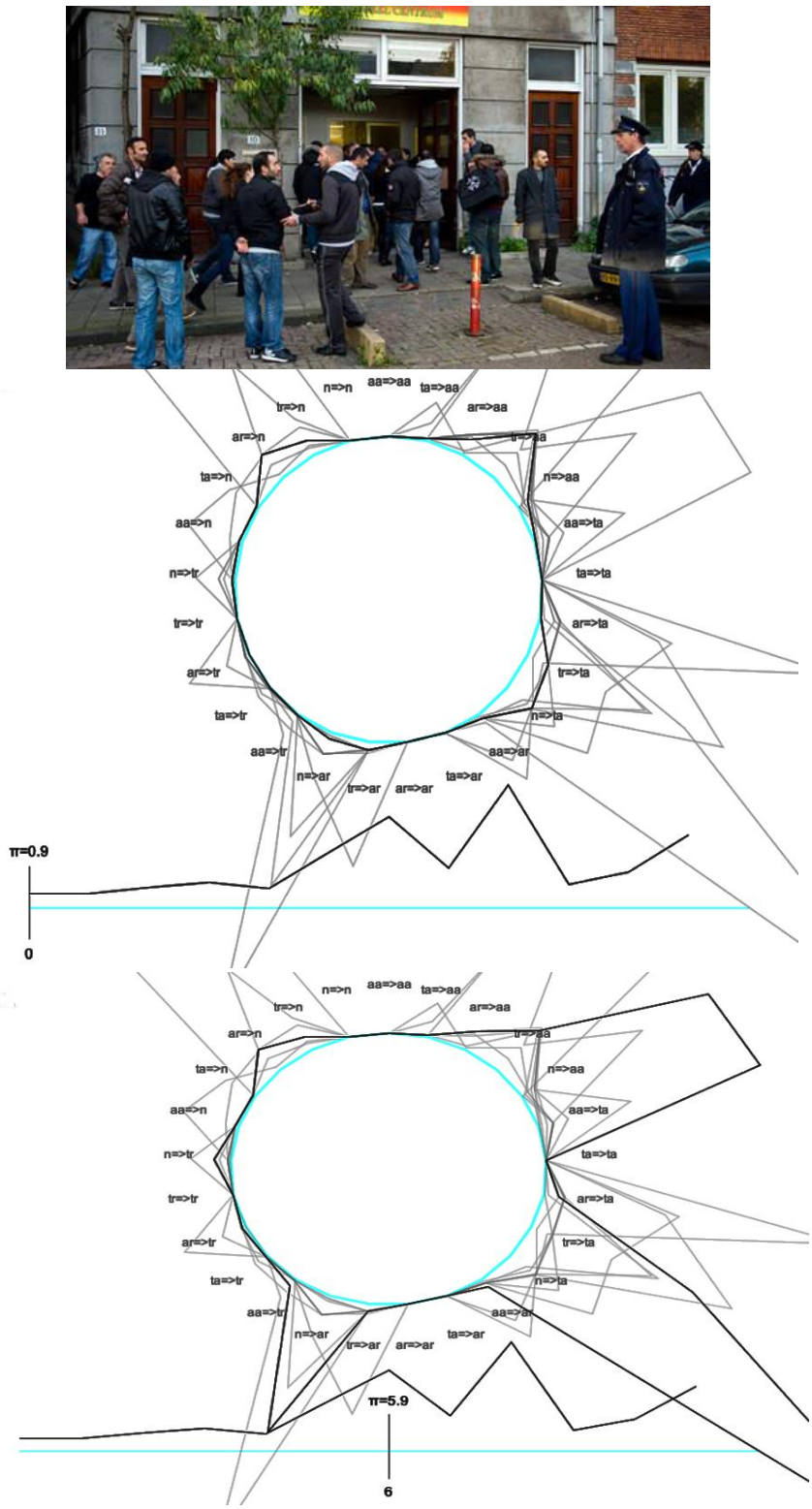

Figure 5: Visualization of the system response during a riot between Turkish and Kurdish people in Amsterdam in October 2011 at two moments in time. The deviation of the response is shown per variable in a spider diagram, where the circle is normal (n=total tweets, ta=total mentions, tr=total retweets, aa=ta/n, ar=tr/n). Below each spider diagram, the anomaly curve is shown, which combines the responses to a single score for each moment in time.

For this data, we took a multiple variable analysis approach. We devised a system that computes correlation matrices for a number of variables over time. These variables represent the outcomes of data analysis (e.g. the number of positive tweets, the amount of foul or aggressive language, the number of replies, etc.) and are typically based on input by analysts, who already may have good intuitions about which variables, given a certain data set, are interesting The correlation computation deploys the DirectLiNGAM algorithm of Shimizu et al. [24]. In our system, the normal (mean) values and standard deviations of the various correlation pairs are computed over the entire range of measured frames. In addition, the *z-score* (Eq. 1) is computed for every correlation pair.

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

The aggregated *z*-scores of the various correlation pairs for a certain interval are represented by a time series, plotted as a curve that describes for every interval the difference of its mean with the mean of the entire time series. Peaks in this curve represent 'anomalies'. In addition, our system displays a feature polygon (connecting all variables as points), with an inner polygon representing the normal (mean) values of the various variables, and an outer polygon representing (per time tick) the eventual deviations of certain variables. This adds an extra explanatory facility that may help analysts to gain insight into which variables are responsible for a certain peak in the anomaly curve. In Figure 5, the output of the system is shown for the Turkish-Kurdish clash in Amsterdam.
The left sub-figure shows the response before the riot and the right shows the response when the riot started, with a latency of 30 minutes.

Changes and deviations can be analyzed with this system. Also changes from an individual, group or subject can be observed over time and relative to others in the same context can be judged.

4. IMAGE ANALYSIS

The analyzed messages have a multi-modal character; besides text there can be images and movies with sound present. The complete message forms an expression of which we want to analyze the meaning. As an extension of the text-based analysis, we investigated whether additional information from attached images can be extracted. In this research, we assumed that the text of a message does not have to be descriptive, but it can refer to a meaningful attached image.

To automatically find interesting cases in a large number of messages, it is important to know what happens in the image. Cases that desire special attention are (company) logos, letters or numbers in images, person detection and recognition, naked persons, shirts of a team (hooligans), avatars, and also abstract issues such as aggression.

Two approaches are shown that are important for the image-analysis methods to come to an understanding of the meaning.

- Is it new or reused image material?
- Is it about the search for an abstract concept or a specific instance?

4.1 Search for a reused specific instance

In the following fictive message, the text says "We will do this again" together with a picture of the Twin Towers. The text lacks information for the text-based sentiment analysis or anomaly detection; it is merely a reference to an image. However, the image does contain meaning that may be relevant in a certain context (Figure 6).

Figure 6. The text "we will do this again" is meaningless without the image of the twin towers.

In the case of the Twin Towers, existing photo material was reused and published. By using a commercial service, named TinEye [14], republished material can be found on the internet, if it is present in their database. By searching with this service for similar images, pages are found where (a modified form of) the image is used. The text on this page can now be extracted and with existing text-mining techniques keywords are selected that correspond to these images (Figure 7). The keywords of these pages are used to improve the text-based analysis and anomaly detection.

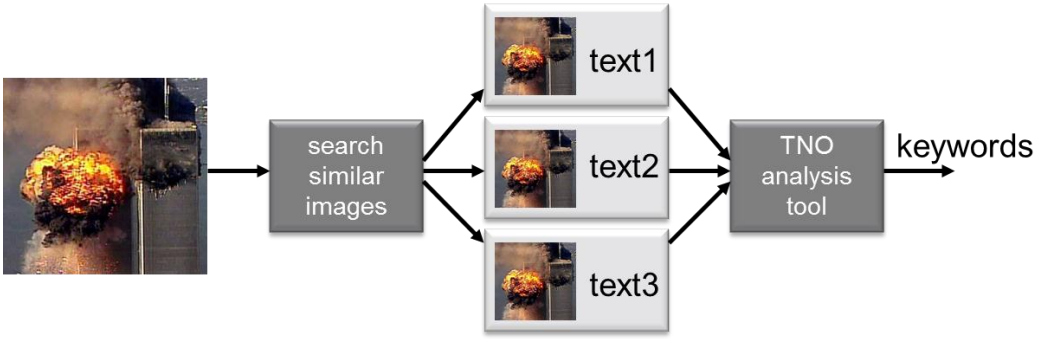

Figure 7. Analysis process for extraction of keywords.

The text analysis extracts all words without tags from the pages, removes irrelevant stop-words (the, is, at, on), reduces the words to canonical form (with Porter stemming), and makes a histogram of all words. Words that are present at multiple pages are probably related to the image. In the case of the twin-tower image, the keywords appeared to be the following:

- muslim (415), iran (322), islam (294), hood (266), fort (256), obama (176), terror (141), saudi (136), kill (136).

The textual description of the message can be extended with these keywords that are related to the image. Actual anomaly detection is then done on the extended text as described before.

4.2 Search for an abstract concept in new material

Another approach is based on the detection of abstract concepts in an image.

The following is based on messages of twitter with the text "accident" (they were selected with the Dutch word for accident: 'ongeluk'). One shows a photo of a traffic accident, and the other a photo of a table with fallen dinnerware

(Figure 8). The first is for our stakeholders of more interest to retrieve witnesses (the sender of the message may know more), while only few are interested in the message with dinnerware.

Figure 8. Accident with dinnerware (left) and an accident with a police car (right).

New or unique material cannot be analyzed by this or similar services, because they have not been indexed yet. With the help of a trained concept detector, it can be determined to which category the image belongs. The used techniques for this approach are reported by Schavermaker [21][22].

Figure 9. Examples of a search for the concept ‘police car’ and ‘demonstration / crowd’.

To select a set of relevant accidents a police car detector was made [21][22]. When a new image contains (a part of) a police car, it is more likely to be of interest for the “accident” class (Figure 9).

5. CONCLUSION

In this paper, we presented a method to automatically monitor trends and detect behavior of anomalies on Twitter or other social media. Specifically, we proposed a profiling system based on non-negative matrix factorization and the correlation analysis of a number of variables, such as sentiment, retweets, post frequencies and topological properties of senders, medium and receiver. Correlation patterns between these variables are used for the profiling of network snapshots and individuals. We analyzed pictures and generated keywords to improve the overall system. The system was tested on Twitter data. The results showed that our system can successfully recognize abnormal changes in behavior, activity or emotion.

The two systems that can analyze the content of reused or new images each have their strengths and weaknesses. The advantage of the first approach is that this can be successful without knowing the number of categories. In this case it only works for known images, not for recent or unique photos. An advantage of the second approach is that unique and new material can be analyzed. Then it only works for specific predetermined concepts.

Our first tests showed that most of the photographic material on Twitter is not a duplicate of an existing image on a website. In our database are hundreds of images that were uniquely created. The question is how important it will be to find reused material when it appears, but it seems better to focus on the classification of predefined concepts.

On this moment, it is possible to look back to see whether deviations in text or images in the social media could have indicated the upcoming commotion. The potential of automatic anomaly detection is that in a large amount of data abnormal behavior can alert an analyst in an early stage. The analyst can interpret the signals for improved usage during the occurring situation.

6. ACKNOWLEDGEMENT

The work for this paper was supported by the Dutch Ministry of Security and Justice in the program for safety and security research: "Veilige maatschappij" (project "Slimmer omgaan met grote hoeveelheden informative in de veiligheidsketen"). The authors thank the National Police Intelligence Service (IPOL) in The Netherlands for their help and advice.

REFERENCES

[1] Androulidakis, G., Chatzigiannakis, V., Papavassiliou, S., "Network anomaly detection and classification via opportunistic sampling", IEEE Network 23(1), 6-12 (2009).
[2] Au, C, Skaff, S, Clark, J., "Anomaly detection for video surveillance applications", Proc. Int. Conf. Pattern Recognition 4, 888-891 (2006).
[3] Best, C., "Challenges in open source intelligence", IEEE Intelligence and security informatics conference, 58-62 (2011).
[4] Betzwieser, J., Mason, W., Redmann, R., Taylor, Z., Tsao, S., Brown, D., Conklin, J., "Systems methodology to characterizing the threat posed by anonymous systems on the internet", Systems and Information Engineering Design Symposium, 159 - 164 (2009).
[5] Brown, I., Korff, D., "Terrorism and the proportionality of internet surveillance", European Journal of Criminology 6(2), 119-134 (2009).
[6] Chandola, V., Banerjee, A., Kumar, V., "Anomaly detection: a survey", ACM Comp. Surv. 41(3), (2009).
[7] Chandola, V., Banerjee, A., Kumar, V., "Anomaly detection for discrete sequences: A survey", IEEE Trans. Knowledge and Data Engineering, (2010).
[8] Corley, C., Cook, D., Mikler, A., Singh, K., "Text and structural data mining of influenza mentions in web and social media", Int. Journal of Environ. Research and Public Health 7(2), 596-615 (2010).
[9] Dabeer, O., Mehendale, P., Karnik, A., Saroop, A., "Timing tweets to increase effectiveness of information campaigns", Proc. ICWSM, (2011).
[10] Dinev, T., Hart, P., Mullen, M., "Internet privacy concerns and beliefs about government surveillance - an empirical investigation", Journal of Strategic Information Systems 17(3), 214-233 (2008).
[11] Fuchs, C., "New media, web 2.0 and surveillance", Sociology compass 5(2), 134-147 (2011).
[12] Gaans, M.M., [Wat nou social media], Coosto, Eindhoven The Netherlands, (2011).
[13] Heide, L. van der, "Individual terrorism; indicators of lone operators", Master Thesis, Utrecht The Netherlands, (2011).
[14] Hua, G, Tian, Q., "What can visual content analysis do for text based image search?", IEEE Int. Conf. Multimedia and Expo, 1480-1483 (2009).
[15] Inazumi, T., Shimizu, S., Washio, T., "Use of prior knowledge in a non-Gaussian method for learning linear structural equation models", Proc. Int. Conf. Latent Variable Analysis and Signal Separation, 221-228 (2010).
[16] Jijkoun, V. and Hofmann, K, "Generating a Non-English Subjectivity Lexicon: Relations That Matter", Proc. Conf. Eur. Chapter Assoc. Computational Linguistics, (2009).

[17] Kasiviswanathan, S., Melville, P., Banerjee, A., Sindhwani, V., "Emerging topic detection using dictionary learning", ACM Proc. Conf. Information and Knowledge Management CIKM, (2011).
[18] Meij, E., Trieschnigg, D., Rijke, M. de, Kraaij, W., "Conceptual language models for domain-specific retrieval", Information Processing & Management 46(4), 448-469, (2010).
[19] Qui, T., Feng, J., Ge, Z., Wang, J., Xu, J., Yates, J., "Listen to me if you can: tracking user experience of mobile network on social media", ACM Proc. Internet Measurement Conf. IMC, (2010).
[20] Saha, A., Sindhwani, V., "Learning evolving and emerging topics in social media: a dynamic NMF approach with temporal regularization", ACM Web search and data mining WSDM, (2012).
[21] Schavemaker, J, Doets, P., Kraaij, W., Raaijmakers, S., Staalduinen, M., "Notebook paper: TNO instance search submission 2010", Proc. TRECVID, (2010).
[22] Schavemaker, J, Eendebak, P., Staalduinen, M., Kraaij, W., "Notebook paper: TNO instance search submission 2011", Proc. TRECVID, (2011).
[23] Schultz, F., Utz, S., Göritz, A., "Is the medium the message? Perceptions of and reactions to crisis communication via twitter, blogs and traditional media", Public Relations Review 37(1), 20-27 (2011).
[24] Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P.O., and Bollen, K., "DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model", J. Mach. Learn. Res. 12, 1225-1248 (2011).
[25] Snoek, C.G.M., Smeulders, A.W.M., "Visual-concept search solved?", IEEE Computer, vol. 43(6), 76-78 (2010).
[26] Raaijmakers, S., Kraaij, W., "Classifier calibration for multi-domain sentiment classification", Int. Conf. Weblogs and Social Media, 311-314 (2010).
[27] Uijlings, J., Smeulders, A., Scha, R., "Real-time visual concept classification", IEEE Trans. Multimedia 12(7), 665 – 681 (2010).
[28] Yang, J., Counts, S., "Predicting the speed, scale, and range of information diffusion in Twitter", Proc. ICWSM, (2010).
[29] Wulsin, D., Blanco, J., Mani, R., Litt, B., "Semi-supervised anomaly detection for EEG waveforms using deep belief nets", Proc. Int. Conf. Machine Learning and Applications, (2010).
[30] Wu, S., Banzhaf, W., "Combatting financial fraud: a coevolutionary anomaly detection approach", Proc. Conf. Genetic and evolutionary computation GECCO, 1673-1680 (2008).
[31] Xu, X., Wang, X., "An adaptive network intrusion detection method based on PCA and support vector machines", Advanced Data Mining and Applications LNCS 3584, 696-703 (2005).