Open Source Intelligence for Traditional- and Social Media Sources

The Sail Labs Media Mining System for OSINT

Gerhard Backfried[12], Christian Schmidt[1], Mark Pfeiffer[1], Gerald Quirchmayr[2], Johannes Göllner[3]
[1] Sail Labs Technology GmbH, Vienna, Austria,
[2]University of Vienna, Faculty of Computer Science, Multimedia Information Systems Research Group, Vienna, Austria
[3]Federal Ministry of Defence and Sports, National Defence Academy, Vienna, Austria
{gerhard,christian,mark}@sail-technology.com[1], gerald.quirchmayr@univie.ac.at[2], johannes.goellner@bmlvs.gv.at[3]

***Abstract*—In this paper we describe the Sail Labs Media Mining (MM) System, a state-of-the-art end-to-end Open-Source-Intelligence (OSINT) system capable of processing large amounts of data such as typically gathered from open sources in unstructured form. Audio, video and textual content from traditional as well as social sources are ingested, processed and made available for search and retrieval, analysis and visualization. Processing is performed by a modular set of technologies packaged as components and covering a large variety of formats and languages. The system can serve as a search platform across open, isolated or secured networks and can be used as a tool for situational awareness, information sharing or risk assessment and supports the complete range of phases in the OSINT-cycle. [1]**

Keywords— Multimedia Computing, Speech Processing, Situational Awareness, Multilinguality, Social Media, Open Source Intelligence

I. INTRODUCTION

An ever-increasing amount of information is being produced by the second, put on the internet by professional news agencies or individuals, broadcast by TV- and radio stations or chatted about on social media in a multitude of languages. An increasingly large portion of this immense pool of data is multimedia- and social media content. To tap into this constant flow of information and make these contents searchable and manageable on a large scale, the Media Mining System (MM-System) provides a framework which allows the flexible combination of a variety of components for the analysis of the different kinds of data involved. Information and clues extracted from audio- as well as video-tracks of multimedia documents are gathered and stored for further analysis. This is complemented by information extracted from textual documents of diverse qualities and formats, web-feeds, blogs and social media sources. The resulting information is combined and made available on a multi-media server which allows visualization and analysis of the data.

[1] This work was supported by the FFG project QuOIMA (836278) under the Austrian National Research Development Programme KIRAS

Whereas in the past textual content was the primary source for the extraction and gathering of intelligence in the area of situational awareness and Open Source Intelligence (OSINT), the analysis of multi-media- and social media content has been receiving increased attention over the last years. News and content presented on national as well as international sources and across a variety of media complement and extend each other. Together they form a broad basis for long-term trend analyses and ad-hoc situational awareness.

II. SYSTEM DESCRIPTION

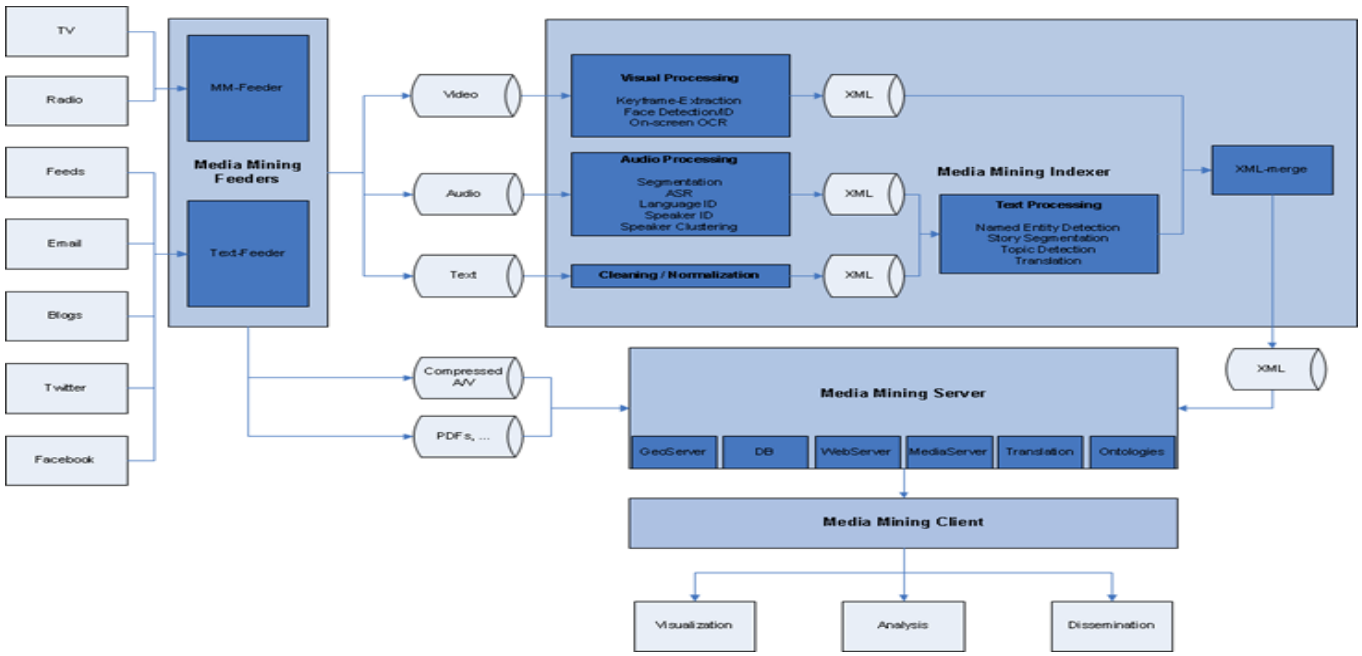
The MM-System is a modular system aiming to cover the complete OSINT workflow cycle from the requirements phase to the dissemination and feedback phases [1]. It enables OSINT professionals to quickly extract meaningful analyses from unstructured data in a variety of formats across multiple languages and sources. Analysts are supported in their work by tools to visually explore and search large volumes of data according to their mission and fields of intelligence. The MM-System consists of a set of technologies packaged into components and models, combined into a single system for end-to-end deployment. A number of toolkits allow end-users to update, extend and refine models in order to respond flexibly to a highly dynamic environment. Not all components and technologies need to be present initially and can be added flexibly over time. Several Feeders, Indexers and Servers may be combined to form a complete system. Fig. 1 below provides an overview of the components of the MM-System and their basic interaction [22]. The resulting documents of the individual processing tracks are fused at the end of processing (late-fusion). The XML-files and associated meta-data are uploaded, together with a compressed version of the original media files, onto the Media Mining Server (MMS), where they are made available for full-text search and retrieval.

A. Media Mining Feeders

Feeders represent the input interfaces of the MM-System to the outside world. For audio or mixed audio/video input a variety of formats can be ingested from external sources. To handle textual input, such as Web-Pages, press-agencies, e-mails, blogs and social media sources, separate feeders exist which extract the relevant content from these sources and pass

it on to the text processing components. The feeder determines the eventual unit of processing - the document - in the MM-System, forming the basis for all storage, visualization and retrieval activities. All documents in the MM-System are assigned a set of properties among which the point in time when the document was entered, and its source, channel and program play primary roles. Filtering of search results and visualization can be applied to these properties.

Fig. 1. Overall Architecture of the Media-Mining-System



B. *Media Mining Indexer (MMI)*

The MMI forms the core for the processing of audio and text within the MM-System. It consists of a suite of technologies and associated models, packaged as components, which perform a variety of analyses on the audio and textual content. Processing results are combined by incrementally enriching XML-structures. Facilities for processing a number of natural languages exist for the components of the MMI. Models for languages are developed by Sail Labs in cooperation with partner organizations and customers and form an active area of development.

1) Audio Processing:

After having been converted to the appropriate format by the multimedia-feeder, the audio signal is segmented for further analysis [6]. The content of each segment is analyzed with regard to the proportion of speech contained, and only segments classified as containing a sufficient amount of speech are passed on to the automatic speech recognition (ASR) component. The ASR-component is designed for large-vocabulary, speaker-independent, multi-lingual, real-time decoding of continuous speech. Decoding is performed in a multi-pass manner, each phase employing more elaborate and finer-grained models refining intermediate results, until the final recognition result is produced [7]. Subsequently, text-normalization as well as language-dependent processing is applied to yield the final decoding result in an XML format [8]. Speaker Identification (SID) is applied to the output produced by the segmentation step using a set of predefined target-speaker models. Data of the same speaker is clustered and labelled with a unique identifier [9].

2) Text Processing

All text-based technologies perform their processing either on the output of the ASR-component or on data provided by the text-normalization components. Text- processing includes the pre-processing, cleaning, normalization and tokenization steps. Named entity detection (NED) of persons, organizations or locations is performed on the pre-processed texts. The NED system is based on patterns as well as statistical models defined over words and word-features, can be extended and is run in multiple stages [10]. The topic-detection component (TD) first classifies sections of text according to a specific hierarchy of topics. Similar adjacent segments are merged. The models used for TD and story segmentation are based on support vector machines (SVM) with linear kernels [11]. Sentiment analysis is performed on text or transcription results on a document as well as a sub-document (e.g. paragraph) level [16].

3) Visual Processing

In order to complement the information extracted from the audio stream of input data, the MM-System also provides facilities to extract information from the visual signal (3rd-party components developed in co-operation with partners). In particular, faces of persons [2,3] and on-screen text are detected and identified [4,5].

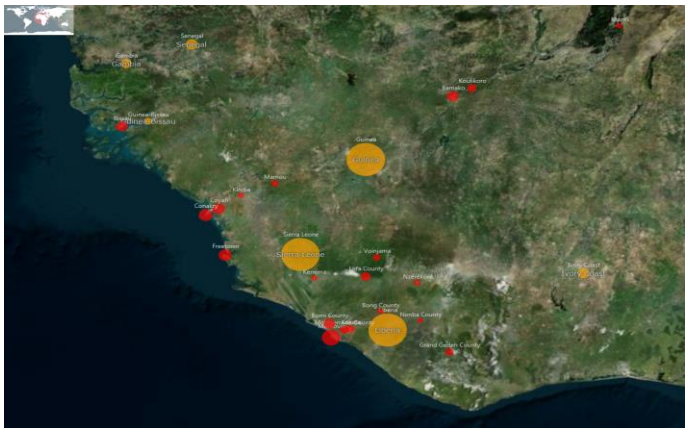
C. Media Mining Server (MMS)

The MMS comprises the actual server, used for storage of XML and media files, as well as a set of tools and interfaces used to update and query the contents of the database. The MMS is based on Oracle 11g and Apache Lucene [17] for all search and retrieval functionalities. The MMS also comprises semantic technologies forming the basis for all ontology-related operations, a geo-server and translation functionalities. Different types of and interfaces to translation facilities are offered by the MMS: parallel translations can be created for a transcript as it is uploaded to the server using 3rd-party machine translation engines, keyword translation and human translation, via an e-mail based interface, are supported. Translations are taken into account for queries and visualization. The Geo-Server provides map image rendering and projection of underlying datasources. MMS by default provides data sources obtained from NASA (BlueMarble), Unearthed Outdoors (TrueMarble) and GeoNames. These can be extended by clients if required to allow the display of proprietary imagery within the MMC.

D. Media Mining Client (MMC)

The Media Mining Client provides a set of features to let users query, interact with, visualize and update the contents of the data stored in the MMS. Users can perform queries, upload or download content, request translations or add annotations to the stored documents. All user-interaction and collaboration take place through the MMC. Documents, search results and aggregations can be displayed in a variety of manners. Users can view data and search-results from different perspectives, allowing them to focus on relevant aspects first and to iteratively circle-in on relevant issues.

Fig. 2. Locations mentioned in connection with Ebola

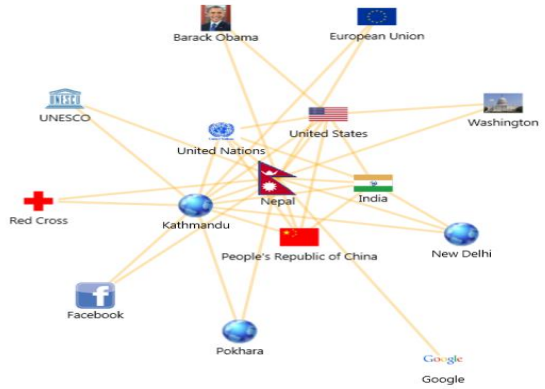


The retrieval process can thus be carried out in a drill-down manner, iteratively narrowing down the data set under inspection.

Locations detected in documents are mapped onto a 2D or 3D map via geo-coordinates. This mapping can be applied on an individual location level, or on aggregated levels such as provincial or country-levels. Fig.2 above shows a map of West Africa centered around Sierra Leone with all locations marked which had been detected in any of the input sources during the previous three days in connection with the word Ebola.

Relationships between detected entities can be visualized and explored via a relationship-graph. Entities co-occurring frequently form the basis of this graph, which can be extended in an exploratory way by interactive removal or addition of entities. Fig. 3 below displays entities mentioned in connection with the earthquake in Nepal in 2015.

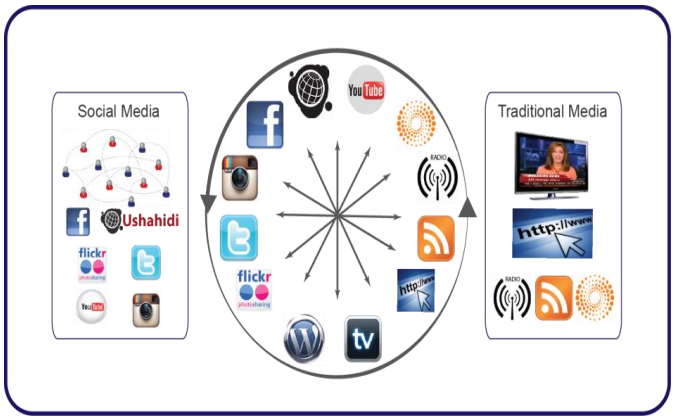
Fig. 3. Entities mentioned in connection with the Nepal earthquake, 2015



III. CROSS MEDIA

Traditional media, such as TV, radio and print-media, have a long history of providing information in times of crises and disasters. Recent years have witnessed a dramatic surge in the use of social media platforms accompanied by the extended participation of organizations and individuals. Professional users and entities increasingly exploit social media as additional channels to distribute their content, often simply re-emitting content which is available via other channels (such as feeds from the same source). Posts and comments may provide insights and further aspects of events already covered by traditional media. Including them and following the information provided by links within them allows for increased coverage and better and extended contrast across sources and media. Figure 4 below provides a schematic overview of some of the possible cross-media connections and links across platforms forming the basis for extended data-collection and –analysis (e.g. mentions of twitter-handles, accounts, hashtags, Facebook accounts, email addresses, names and times of TV broadcasts, etc.)

Fig. 4. Cross-Media Links between traditional and social media



A. *Twitter*

In our current approach for tweets we distinguish between seeders [13] and other projects called tracking-projects. Seeders are typically news-agencies, professional journalists or bloggers using Twitter as an additional distribution channel. Their tweets commonly contain a snippet of text, a headline, and a shortened link to a URL. We gather the text of the tweet itself as well as the content of the referenced web-page and the associated PDF image of the page at the time of processing and link these items together to form one document. Tracking projects are defined in terms of a set of hashtags, user-ids and/or geo-coordinates. These are combined and the Twitter streaming API [14] is used to filter and retrieve all relevant tweets. Upon retrieval, these tweets are grouped together according to their tracking-projects' properties and time-stamp. Documents created by projects can be assigned to channels or programs according to their subject, thus making it possible to combine tweets on a certain subject with feeds and TV programs on the same subject for analysis. Taking into account the information on time and date, this allows for contrasting information provided on different sources about the same event.

B. *Facebook*

Regarding Facebook, the system currently processes posts and associated comments from public pages (e.g. of politicians). These are grouped according to time and form the basis for documents for this source. Over time, further comments made to a post are added incrementally to the document. Links to web-pages made in a post are collected and added (as PDF) to the document. Documents thus eventually represent a post and all associated comments (up to a given point in time). A Facebook-project in the MM-System constitutes one public page and an associated set of parameters. These parameters determine, e.g. the frequency by which comments to posts will be retrieved and until what time a post will be monitored. Dynamic adjustment to the retrieval frequency depending on the amount of activity is made. The same possibility regarding assignment to channels or programs applies as for tweets. Likewise, the parallel monitoring of events as they evolve is possible.

IV. OSINT

Typical intelligence cycles comprise the stages of requirements gathering, planning and direction, collection, processing and exploitation, analysis and production and dissemination [1, 15]. Depending on the organization, mission and array of sources these cycles vary slightly in their exact nomenclature and detail; however, all of these cycles have certain common elements. Depending on the operational background, goals are laid out as to what information or sources might be of interest. The actual collection or harvesting and storing of the data follow. Processing/enrichment then transcodes or normalizes the data and stores them according to the particular system requirements. Other elements of processing may include the insertion of metadata indicating dates, locations, translations, etc. to the sourced information.

In the analysis phase, the information is screened and rated by the analysts, who are tasked with judging the material by several parameters. This step requires a high degree of contextual, linguistic, cultural and mission-related knowledge. The result of the analysis phase is a collection of pieces of intelligence that have been rated, verified or falsified (as this also is an important aspect especially in OSINT), categorized (sources may have extreme value in certain domains, but are used to proliferate false information on other topics) and classified. Finally, the results are fused and combined into reports that, depending on request, are made available in ad-hoc, daily- , 3 day-, or weekly briefings and long term analyses. The difficulty in any operation is not only to bridge the media gap but also linguistic- and cultural gaps.

The MM-System allows working across these constraints by providing sufficient flexibility in these aspects and offering additional tools and interfaces (e.g. translation engines) in order to best facilitate the demands. Users are thus enabled to focus on the analytical aspects of the mission rather than being constrained and bogged-down by technical aspects. From an operational point of view, the enormous amount of data harvested from multi-media, web media and social media in particular pose a unique challenge which was previously only known in the SIGINT part of an operation (in the case of massive passive surveillance). In a harvesting approach which is too broad the amount of chaff collected outweighs by far the useful information in a ratio which can only be compared to the famous needle in a haystack. Guidance and mission oriented collection and harvesting, focused visualization and incremental exploration are essential elements for any operation to have even a remote chance of success. On the operations side itself it is necessary to complement the technical advances by building up broader skilled teams that are able to collaborate efficiently. Diversity in knowledge, background and opinion is necessary in order to achieve the goals set. The MM-System aims at supporting the complete OSINT-cycle: Feeders support the collection phase, the backend MMS and visualization and search facilities in the MMC support the analyst during the analysis phase. Analysts cooperate using the MMC and jointly produce reports which can subsequently be disseminated in a variety of ways, among them via hand-held devices.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

CURRENT STATUS AND OUTLOOK

The MM-System presents a state-of-the-art end-to-end Open-Source-Intelligence (OSINT) system. The described components of the MM-System are already operational and have been deployed at numerous organizations. Their role is to aid analysts and experts by supporting data-gathering and –analysis tasks, automating processes wherever possible. Users still have to fulfill certain tasks, e.g. the selection of sources, attribution of credibility or, more generally, putting aggregated information into proper context. However, they are freed from routine tasks and supported when dealing with heterogeneous and large amounts of information, allowing them to concentrate on more complex and demanding tasks. The system provides various mechanisms to explore data-sets in an interactive way and to collaboratively create reports and intelligence.

MM-Feeders are used for monitoring news on up to 500 TV- and radio-stations, in 14 languages, on a 24x7 level to provide a constant flow of input to the MMI, which continuously processes the incoming data stream and sends its output to the MMS. In addition, several thousand feeds and dozens of social-media projects are processed by the system. All processing is targeted to take place in real-time and with minimal latency. Results of all processing are made available on the MMS and can be exported to existing infrastructure. Via the MMC, dozens concurrently access the system and produce mission-critical data. Further components are currently under development or in the stage of research prototypes. Particularly in the area of social media, further platforms (photo- and video-sharing, networking,..) are being integrated. Likewise we are planning to exploit further features available from the visual signal, such as for example the detection of logos. In parallel, the existing technologies are constantly being improved and new languages added to the portfolio of models. Installations at end user's sites are exploited for rapid feedback during development and for evaluation purposes. New features and technologies are phased in as they become available.

REFERENCES

1. http://www.fbi.gov/about-us/intelligence/intelligence-cycle, 04/21/2015
2. P. Viola, M.J. Jones, "Robust Real-Time face detection", Int'l Journal of Computer Vision, 57(2):137–154, 2004
3. S. Kim, S. Chung, S. Jung, S. Jeon, J. Kim, S. Cho, "Robust face recognition using AAM and gabor features", In Proc. World Academy of Science, Engineering and Technology, 2007
4. G. Dedeoglu, T. Kanade, and S. Baker, "The asymmetry of image registration and its application to face tracking", IEEE Trans. on Pattern Analysis and Machine Intelligence, 29(5):807–823, 2007
5. http://code.google.com/p/tesseract-ocr, 04/21/2014
6. D.Liu, F.Kubala, "Fast Speaker Change Detection for Broadcast-News Transcription and Indexing", Eurospeech 1999
7. R.Hecht,J.Riedler,G.Backfried, „Fitting German into N-gram Language Models", TSD 2002
8. D.Liu, F.Kubala, "Online Speaker Clustering", ICASSP 2003
9. R.Schwartz, L.Nguyen,J.Makhoul, "Multiple-Pass Search Strategies", Automatic Speech and Speaker Recognition, 1996
10. D.Bikel, S.Miller, R.Schwartz, R.Weischedel, „Nymble: High-Performance Learning Name-Finder", Conference on Applied Natural Language Processing, 1997
11. T.Joachims, "Text Categorization with Support-Vector-Machines: Learning with Many Relevant Features", ECML, 1998
12. A. Mitchell, T. Rosenstiel, "Overview of the State of the News Media 2012", http://stateofthemedia.org/2012/overview-4/, 03/29/2012
13. J.Sankaranarayanan, et al. „TwitterStand: News in Tweets", GIS'09, 2009
14. https://dev.twitter.com/, 04/21/2015
15. Open Source Intelligence, US Army, available at http://www.fas.org/irp/doddir/army/fmi2-22-9.pdf, 04/21/2015
16. G. Shalunts, G. Backfried, K. Prinz, "Sentiment Analysis of German Social Media Data for Natural Disasters", 11th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2014, University Park, PA, USA
17. Apache Lucene, http://lucene.apache.org/ on 04/21/2015
18. S. Petrovic, et al , "Can Twitter replace Newswire for breaking news?", ICWSM 2013
19. G. Backfried, et al. "Cross-Media Communication during Crises and Disasters", EMCSR 2014
20. L. Kwang-Hoong, and L. Mei-Li, "The Fukushima nuclear crisis reemphasizes the need for improved risk communication and better use of social media", Health Physics 103 (3): 307-310, 2012
21. Y. Tyshchuk, et al. "Social Media & Warning Response Impacts in Extreme Events: Results from a Naturally Occuring Experiment", 45th International Conference on System Science (HICSS), 2012
22. G.Backfried, et al. "Open Source Intelligence in Disaster Management", EISIC 2012, Odense, Denmark