GeoSensor: On-line Scalable Change and Event Detection over Big Data

Giorgos Argyriou[1], George Papadakis[1], George Stamoulis[1], Efi Karra Taniskidou[1], Nikiforos Pittaras[2], George Giannakopoulos[2], Sergio Albani[3], Michele Lazzarini[3], Emanuele Angiuli[3], Anca Popescu[3], Argyros Argyridis[4], Manolis Koubarakis[1]

[1] National and Kapodistrian University of Athens, Greece {gioargyr, gpapadis, gstam, efikarra, koubarak}@di.uoa.gr, [4]arargyridis@gmail.com

[2] NCSR Demokritos, Greece {pittarasnikif, ggianna}@iit.demokritos.gr

[3] European Union Satellite Centre, Spain {sergio.albani, michele.lazzarini, emanuele.angiuli, anca.popescu}@satcen.europa.eu

ABSTRACT

GeoSensor is a novel system that enriches change detection over satellite images with event detection over news items and social media content. GeoSensor faces the major challenges of Big Data: volume (a single satellite image may be a few GBs), variety (its data sources include two different types of satellite images and various types of user-generated content) and veracity, as the accuracy of the end result is crucial for the usefulness of our system. To overcome these three challenges, while offering on-line functionality, GeoSensor comprises a complex architecture that is based on the open-source platform developed in the H2020 project Big Data Europe. Through the presented demonstration, both the effectiveness and the efficiency of GeoSensor's functionalities are highlighted.

KEYWORDS

Big Data; Satellite image processing; Event Detection; Change Detection; Semantic Web

1 INTRODUCTION

The growing digitization of our society has a large influence on all aspects of everyday life. Huge amounts of data are being produced and have the potential to create new knowledge and intelligent solutions for economy and society, when properly analyzed and interlinked.

The research area of Big Data can make important contributions to technical progress in key societal sectors and help shape business. What is needed are innovative technologies, strategies and competencies for the beneficial use of big data to address societal needs.

The Big Data Europe[1] project is an H2020 action which developed a Big Data Infrastructure *(BDI)* and applied it to the seven societal challenges of the H2020 Programme: Health, Food and Agriculture, Energy, Transport, Social Sciences and Security.

The system presented in this paper, called *GeoSensor*, has been developed in the context of the Security Challenge[2]. It integrates remote sensing information with social sensing sources. Remote sensing techniques are used to compare two or more satellite images that depict the same area of interest on the Earth surface taken at different times in order to identify areas with changes in land cover or land use (e.g. formerly forested area that is now occupied by buildings). This process is called *change detection* [2] and it is a well-studied problem in the area of Remote Sensing. Social Sensing, on the other hand, is the process of clustering together text sources such as news items and social media posts that pertain to the same real-world event. This process is called *event detection* [1], [6]. In this way, GeoSensor is able to provide insights into the changes extracted from satellite images by associating them with relevant news and social media content for the same location.

To the best of our knowledge, GeoSensor is the only existing system to offer this functionality. Another advantage of GeoSensor is that change detection and event detection are carried out efficiently by exploiting the massive parallelization that is offered by the Apache big data system, Spark. GeoSensor has been developed using the BDI platform [5] and all of its components are provided as Docker images[3]. The whole system can be launched with a single docker-compose file and all the components will run as Docker containers within Docker Swarm[4].

2 GEOSENSOR ARCHITECTURE

The architecture of GeoSensor is depicted in Figure 1. It consists of 11 components and 4 external sources that give rise to 3 workflows, which correspond to the 3 horizontal layers of the architecture. The components at the top (namely News Crawler, Cassandra, Event Detector, Lookup Service and Entity Extractor) implement the *event detection* workflow, while the ones at the bottom (i.e. Image Aggregator, HDFS and Change Detector) form the *change detection* workflow. The components of the middle layer (namely Sextant, SemaGrow, Strabon and Geotriples) form the *activation* workflow that supports the other two.

[1]https://www.big-data-europe.eu

[2]https://www.big-data-europe.eu/security

[3]https://hub.docker.com/u/bde2020/dashboard

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186984

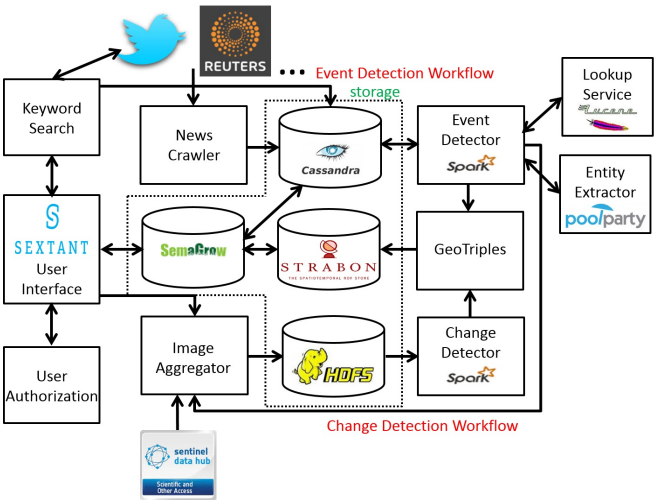

Figure 1: The system architecture of GeoSensor

2.1 System Components

The News Crawler component runs periodically, crawling through social media sources and news articles from specific news agencies. For the time being, these sources include RSS feeds from Reuters, as well as several selected accounts from Twitter. The crawler structure is extensible, supports multiple operation modes and additional information sources can be integrated as needed.

Apache Cassandra[4] is an efficient and scalable distributed database. All produced data from the News Crawler and the Event Detector module are stored in the Cassandra database.

The Event Detector[5] component receives the latest news and social media items stored in Cassandra and groups them into events, using a modified version of the NewSum[7] algorithm[6].

The algorithm of the Event Detector consists of two steps:

(1) A coarse-grained identification of events is performed by comparing distinct pairs of news items, using ngram-graph representations and graph-based textual similarity measures. Appropriate thresholding is applied to identify related pairs, followed by a clustering process to form pools of such related news article pairs. Clusters having too low support are discarded; the rest are considered events and are submitted to a summarization process.

(2) The set of events is subsequently enriched with meaningful metadata. Each identified event is augmented with relevant temporal, geolocation, named-entity and image elements, extracted from its content directly or through RESTful-based external or internal services. In addition, social media items are attached to events through a similarity-based classification process.

Each final event structure is stored back into the Cassandra database, and the Strabon component is notified of the new entries. The Event Detector component can operate on large collections of news articles and posts via distributed execution on Apache Spark.

Lookup Service[7] is a component that accepts location names from input text and finds their geo-coordinates. It is used in the location extraction process of the Event Detector. Lookup Service uses Apache Lucene[8] to index a large set of location names from the Global Administrative Areas[9] (GADM) dataset that covers all administrative areas on all countries worldwide. It then performs fuzzy keyword queries in order to identify the most similar administrative area to every location name.

Image Aggregator[10] is a RESTful Web Service that downloads the most suitable satellite images from the Copernicus Open Access Hub[11]. It searches for Sentinel-1 and Sentinel-2 image pairs that match user-defined spatial and temporal acquisition criteria: a certain area of interest and two different acquisition dates. These criteria constitute a set of arguments that are received by the Image Aggregator as geo-coordinates and sensing dates for the images. The match is done by identifying the images with the largest overlap with these criteria and launches a download process. The downloaded images are stored in the HDFS file system and then, the change detector module is involved as a Spark job, having Sentinel-1 images as input. The Sentinel-2 ones are used for visual confirmation of the results through Sextant.

Apache Hadoop[12] is used for its Distributed File System, known as HDFS. Apart from being scalable and fault-tolerant, HDFS's main advantage is that as soon as the images are stored they become available to every Spark node, so the image processing Spark code can be executed in parallel for each image.

Change Detector[13] is the main component of the change detection workflow. It receives as input two GRD (Ground Range Detected) Sentinel-1 images, performs data pre-processing and calibrations, runs a change detection algorithm and generates as output a change map. The Change Detector is based on a parallelized version of several operators (Calibrate, CreateStack, GroundControlPoints, Warp, Subset, TerrainCorrection, ChangeDetection) of the SNAP[14] toolkit for Sentinel images. The output of the Change Detector is an image file in which the detected changes are indicated by specific pixel values. As a final step, a parallel version of the DB-Scan algorithm implemented in Spark is used to reduce false alarms and group neighboring pixels into clusters. The generated output is then visualized as areas of changes through Sextant.

Geotriples [9] is a tool created by the National and Kapodistrian University of Athens for transforming geospatial data from their original formats into RDF. For the developed use case, it receives descriptions of areas (the output of Change Detector) or summaries of events (the output of Event Detector) in JSON format and converts them into RDF. The output of Geotriples is then stored into Strabon.

[4] http://cassandra.apache.org
[5] https://github.com/big-data-europe/docker-event-detection
[6] https://github.com/scify
[7] https://github.com/big-data-europe/pilot-sc7-lookup-service
[8] https://lucene.apache.org
[9] http://www.gadm.org
[10] https://github.com/big-data-europe/pilot-sc7-image-aggregator
[11] https://scihub.copernicus.eu/
[12] http://hadoop.apache.org
[13] https://github.com/big-data-europe/pilot-sc7-change-detector
[14] http://step.esa.int/main/toolboxes/snap

Strabon [8] is a spatio-temporal RDF store created by the National and Kapodistrian University of Athens that efficiently executes GeoSPARQL and stSPARQL queries. Strabon supports spatial datatypes enabling the serialization of geometric objects in OGC standards WKT and GML. Strabon is used to store triples representing descriptions of areas and summaries of events.

SemaGrow [3] is such a federated query processing system that provides a single SPARQL endpoint that federates multiple remote SPARQL endpoints, transparently optimizing queries and dynamically integrating heterogeneous data models by applying the appropriate vocabulary transformations. SemaGrow hides schema heterogeneity and also applies methods from databases and Semantic Web research that take into account the content of data sources to optimize querying plans. In the described use case, SemaGrow federates Cassandra and Strabon and offers a unified SPARQL endpoint for both of them.

Sextant [10] is the basic component of the activation workflow and the entry point for GeoSensor. It has been extended for GeoSensor's needs and provides a graphical interface for the user to perform event detection or change detection by launching the corresponding workflow.

2.2 Data Sources

The primary data used in GeoSensor are Earth observation data, news articles and social media posts.

Earth observation is the use of remote sensing technologies to monitor land, marine and atmosphere. Satellite-based Earth observation relies on the use of satellite-mounted payloads to gather imaging data about the Earth's characteristics. We can distinguish two kinds of remote sensing. *Passive remote sensing* is when the satellite instruments monitor the energy received from the Earth due to the reflection and re-emission of the Sun's energy by the Earth's surface or atmosphere. Optical or thermal sensors are commonly-used passive sensors. *Active remote sensing* is when the satellite is sending energy to Earth and monitoring the energy received back from the Earth's surface or atmosphere, enabling day and night monitoring during all weather conditions. Radar and lasers are commonly used active sensors. Synthetic Aperture Radar (SAR) data is the type of radar data used in the change detection workflow of GeoSensor.

Copernicus[15], the European program for monitoring the Earth, is currently the world's biggest Earth observation program. It consists of a set of complex systems that collect data from satellites and in-situ sensors, process this data and provide users with reliable and up-to-date information on a range of environmental and security issues. The Earth observation satellites that provide the data exploited by the Copernicus services are the *Sentinels*, which are currently developed for the specific needs of the Copernicus program, and the contributing missions, which are operated by national, European or international organizations. The access and use of Copernicus Sentinel data is regulated by EU law; the free, full and open data policy adopted for the Copernicus program foresees access available to all users for the Sentinel data products, via a simple pre-registration on the Copernicus Open Access Hub. GeoSensor uses SAR and optical data from the Copernicus Open Access Hub. The SAR data come from the satellite Sentinel 1 while the optical data come from Sentinel 2.

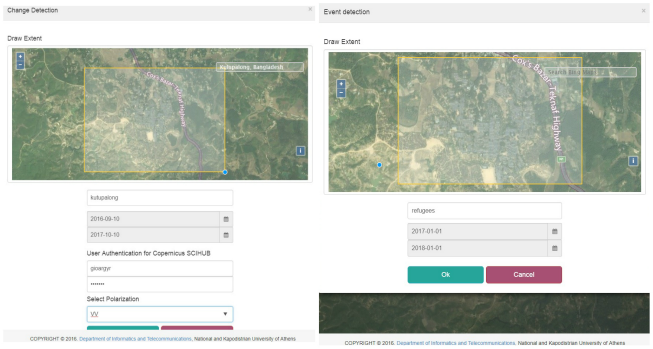

Figure 2: Change Detection and Event Detection initiations

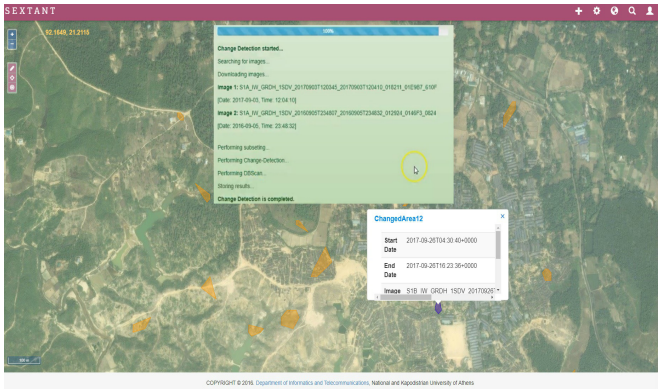

Figure 3: Change Detection results and workflow progress

Regarding the social sensing part, GeoSensor relies on two sources for the time being. News articles are taken from the international news agency, Reuters, through RSS feed. Also, user-generated post from several Twitter accounts are used. Being extensible, our system can retrieve data from more sources if needed.

There are three more data sources that support the event enrichment procedure. The GADM dataset, that contains about 180,000 location names, is used by the Lookup Service making it possible to match location names with their geo-coordinates. The Pool Party platform[16] provides the entity thesaurus and RESTful services required for entity extraction. Lastly, the images that are geolocated with the same area as an event can be retrieved from Flickr.

3 DEMONSTRATION SCENARIO

In the demonstration scenario we will show how GeoSensor can be used for searching for detected changes in land cover or land use within a user-defined area through its change detection workflow. GeoSensor's capability to detect events that are either related to the area where changes have been detected or satisfy other criteria defined by the user will also be shown.

[15] http://www.copernicus.eu

[16] https://www.poolparty.biz

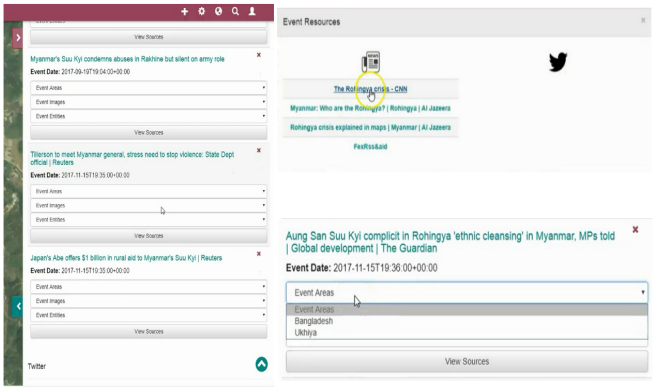

Figure 4: GeoSensor's interface depicting the detected events, the corresponding sources as well as the set of related areas

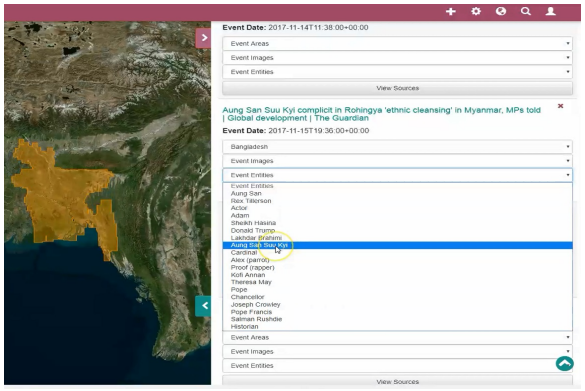

Figure 5: GeoSensor's interfance depicting event-related entities and the view of a selected related area (Bangladesh)

In Figure 2, the user initiates the change detection workflow by providing the following criteria: an area of interest (which is easily drawn on GeoSensor's Earth map), two dates (the reference and the target one), the session name (free choice) as well as the credentials for ESA's Copernicus Open Access Hub (this is necessary for enabling the Image Aggregator to connect with ESA's repository and download the images). In the demonstration scenario we will be targeting the detection of changes in the area of the Kutupalong refugee camp in Bangladesh, which was created as a result of the Rohingya humanitarian crisis. After initiating the change detection workflow, GeoSensor continuously reports the status of its progress (Figure 3) together with the identifying file names of the satellite images that have been selected according to defined criteria. Once the workflow is completed, the areas, where changes occurred between the considered dates, are visualized for the user (yellow areas in Figure 3).

In a similar way, the event detection workflow can be initiated by giving three optional criteria (see Figure 2): an area of interest, a time window defined by two dates, or keywords that pertain to the events of interest. Alternatively, the Event Detection can be launched by selecting an the area of interest defined in the results of change detection. In the returned results, events (Figure 4) are displayed and for each one of them the visualization of its related information is shown. The news sources (Figure 4), the related locations (Figure 4), the relevant images from Flickr, which have been retrieved through geo-tags, and the relevant entities that have been retrieved from Pool Party (Figure 5) are all the different kinds of information that are provided for each event. GeoSensor also supports live, direct search in Twitter using keywords through its integrated search API.

4 CONCLUSIONS

Geosensor is a system developed as contribution to the H2020 BigDataEurope project. GeoSensor is demonstrating a use case relevant for the Security domain that integrates Remote Sensing information with Social Sensing sources. It is supported by the BDI platform and all of its components are provided as Docker images. The proposed system could be adapted for operational scenarios, that are of interest to the European Union Satellite Centre.

5 ACKNOWLEDGEMENT

This work was supported by grant from the European Union's Horizon 2020 research Europe flag and innovation program for the project Big Data Europe (GA no. 644564).

REFERENCES

[1] Farzindar Atefeh and Wael Khreich. 2015. A Survey of Techniques for Event Detection in Twitter. *Comput. Intell.* 31, 1 (Feb. 2015), 132–164.

[2] Francesca Bovolo, Carlo Marin, and Lorenzo Bruzzone. 2012. A novel hierarchical approach to change detection with very high resolution SAR images for surveillance applications. In *2012 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2012, Munich, Germany, July 22-27, 2012*. 1992–1995.

[3] Angelos Charalambidis, Antonis Troumpoukis, and Stasinos Konstantopoulos. 2015. SemaGrow: optimizing federated SPARQL queries. In *Proceedings of the 11th International Conference on Semantic Systems, SEMANTICS 2015, Vienna, Austria, September 15-17, 2015*. 121–128.

[4] Ivan Ermilov, Axel-Cyrille Ngonga Ngomo, Aad Versteden, Hajira Jabeen, Gezim Sejdiu, Giorgos Argyriou, Luigi Selmi, Jürgen Jakobitsch, and Jens Lehmann. 2017. Managing Lifecycle of Big Data Applications. In *Knowledge Engineering and Semantic Web - 8th International Conference, KESW 2017, Szczecin, Poland, November 8-10, 2017, Proceedings*. 263–276.

[5] Sören Auer et al. 2017. The BigDataEurope Platform - Supporting the Variety Dimension of Big Data. In *Web Engineering - 17th International Conference, ICWE 2017, Rome, Italy, June 5-8, 2017, Proceedings*. 41–59.

[6] Jonathan G Fiscus and George R Doddington. 2002. Topic detection and tracking evaluation overview. In *Topic detection and tracking*. Springer, 17–31.

[7] George Giannakopoulos, George Kiomourtzis, and Vangelis Karkaletsis. 2014. NewSum:"N-Gram Graph"-Based Summarization in the Real World.. In *Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding, ed. Alessandro Fiori, 205-230 (2014), accessed June 01, 2017*. 205–230.

[8] Kostis Kyzirakos, Manos Karpathiotakis, and Manolis Koubarakis. 2012. Strabon: A Semantic Geospatial DBMS. In *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I*. 295–311.

[9] Kostis Kyzirakos, Ioannis Vlachopoulos, Dimitrianos Savva, Stefan Manegold, and Manolis Koubarakis. 2014. GeoTriples: a Tool for Publishing Geospatial Data as RDF Graphs Using R2RML Mappings. In *Joint Proceedings of the 6th International Workshop on the Foundations, Technologies and Applications of the Geospatial Web, TC 2014, and 7th International Workshop on Semantic Sensor Networks, SSN 2014, co-located with 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Trentino, Italy, October 20, 2014*. 33–44. http://ceur-ws.org/Vol-1401/paper-03.pdf

[10] Charalampos Nikolaou, Kallirroi Dogani, Konstantina Bereta, George Garbis, Manos Karpathiotakis, Kostis Kyzirakos, and Manolis Koubarakis. 2015. Sextant: Visualizing time-evolving linked geospatial data. *J. Web Sem.* 35 (2015), 35–52.