RAI ACTIVE NEWS: INTEGRATING KNOWLEDGE IN NEWSROOMS

M. Montagnuolo and A. Messina

RAI Radiotelevisione Italiana, Italy

ABSTRACT

In the modern digital age methodologies for professional Big Data analytics represent a strategic necessity to make information resources available to professional journalists and media producers in a more effective and efficient way and enabling new forms of production like data-driven journalism. The challenge lies in the ability of collecting, connecting and presenting heterogeneous content streams accessible through different sources, such as digital TV, the Internet, social networks and media archives, and published through different media modalities, such as audio, speech, text and video, in an organic and semantics-driven way. Rai Active News is a portal for professional information services that addresses these challenges with a uniform and holistic approach. At the core of the system there is a set of artificial intelligence techniques and advanced statistical tools to automate tasks such as information extraction and multimedia content analysis targeted at discovering semantic links between resources, providing users with text, graphics and video news organized according to their individual interests.

INTRODUCTION

The exponential growth of digital resources availability is enabling new forms of content creation, sharing, and delivery. Methodologies for aggregation and presentation of heterogeneous content are needed to make these resources effective and easily available to the final users. Here, the challenge lies in the ability of collecting, connecting and presenting data streams from different media sources, e.g. television, press, the Internet, and of different media types such as audio, speech, text and video.

Rai Active News is a portal for professional information services that addresses these challenges with a uniform and holistic approach. At the core of the system there is a set of artificial intelligence techniques and advanced statistical tools to automate tasks such as information extraction and multimedia content analysis targeted at discovering semantic links between resources, providing users with text, graphics and video news organized according to their individual interests. The system allows to define customized search profiles that are automatically and dynamically updated with the relevant contents found in the monitored information sources, which include Web feeds, television channels and specialized circuits such as the Eurovision News Exchange Network (EVN), or legacy archives. The system also provides a recommendation service based on the analysis of social activities in blogs which dynamically models the user's interest and exploits it to recommend appropriate media content.

This paper presents the underlying infrastructure and technology on which Rai Active News is built. The paper is organized as follows. As a first step we overview challenges and opportunities for knowledge integration in modern news production workflows. Then we describe the architecture developed to address these needs, as well as the services built on top of it. Finally, we conclude with considerations for future work.

CHALLENGES FOR KNOWLEDGE INTEGRATION IN NEWS PRODUCTION

News production is a complex and dynamic process which requires creating content in a fast-paced way and using a wide variety of media, including written texts, images, speech and videos. Contrary to the past, when the life cycle of news items (from sourcing of content to distribution and consumption of products) was typically linear and isolated, we are nowadays dealing with more dynamic and interactive ways to produce, publish and consume news items. Thanks to the proliferation of social networks and open data sources, not only professional journalists, but also individuals are currently taking part in the so called “data journalism” phenomenon. Data journalism is the process of collecting, filtering and structuring big data for storytelling and reporting. In this context the “event”, i.e. any relevant fact happening at some time and place, and the “topic”, i.e. real word entities or things like people, organizations, places or themes, become the basic units around which contents are produced and organized. A topic can include either closed in time or still open events. As an example, the topic about a naval disaster might contain events about the shipwreck, the rescue operations or the ship demolition. The task of data journalism is to extract, track and visualize such hidden information from available data. Machine learning, data mining and semantic Web techniques can be used to gather and analyze data in an unsupervised fashion, thus greatly ensuring productivity and efficiency through all the steps of the process. As an example the News Storyline Ontology[1] is a generic model to describe and organize the stories told by news organizations.

News aggregation aims at structuring content from newswires and broadcasts into topics. Publicly accessible aggregators, such as Google News or Yahoo! News, are becoming popular because of their capability of presenting most recent news automatically aggregated and organized by e.g. number of sources and categories, thus making users save the time they would spend in manually finding information. A study of the impact of news aggregators on Internet news consumption is presented by Athey and Mobius (1).

On the other hand, a critical problem with news aggregators is that the volumes of generated information can be overwhelming to the users. The challenge is to help users in selecting and tracking not everything that is produced but just what they really need. To this purpose, recommendation engines aim to suggest to the users potentially interesting contents based on their historical activity and implicit or explicit feedback. Liu et al (2) present a personalized news recommendation system based on profiles learned from user activity in Google News. Similarly, Li et al (3) describe a system that selects articles from Yahoo! News. Although the cited works constitute an interesting attempt to realize the conception of intelligent and personalized news fruition, they still present drawbacks that hinder their application to massive production environments. In particular, there is a general lack of temporal depth and semantic organization of data in most of the approaches presented so far, thus it is difficult for the users effectively consuming information according to some preferred criteria such as temporal spanning, information

[1] http://www.bbc.co.uk/ontologies/storyline

correlation and trends. To overtake this drawback, Lebal et al (4) describe a system to identify world events from multilingual news articles. Detected events are searchable by means of dates, entities and graphs. However, only textual resources from the Web are considered. As a consequence, processing is based on unimodal and single-domain resources only, i.e. further modalities such as speech or visual content and different domains such as TV broadcasts or digital libraries are not considered. On the other hand, the opportunity of logically integrating contents from multiple modalities and delivered by different service providers can enhance the user satisfaction during content fruition.

ACTIVE NEWS SYSTEM ARCHITECTURE

The Rai Active News architecture is shown in Figure 1. The pipeline is made of four modules: data ingestion, which is responsible of collecting news items from the registered data sources; data pre-processing, where several multimodal content analysis techniques are used to extract valuable information from the input streams; data clustering and warehousing, where news items about the same event are aggregated and semantically annotated; data index and access, where the news events and related metadata are made available for browsing, searching, retrieving and exportation. The overall architecture was developed with open-source tools and integrated with in-house software.

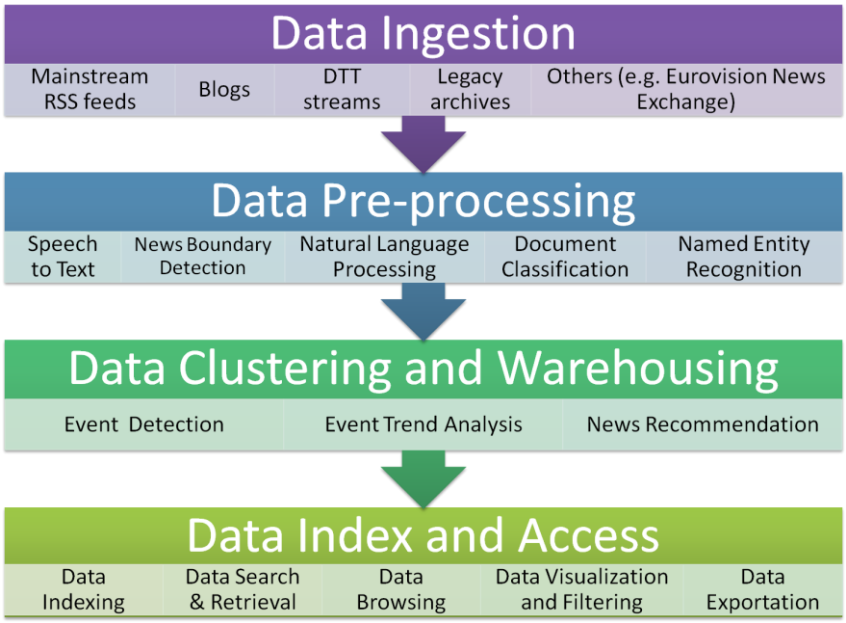

Figure 1 - Pipeline for the Active News system.

Data Ingestion

Input data are ingested from RSS feeds, users' blogs, DTT channels and other services such as the Rai archives and the EVN platform. The complexity introduced here is primarily represented by the necessity of automatically detecting and processing heterogeneous information in real-time, i.e. 24 hours/day and 365 days/year, and fulfilling the users' expectations in terms of functionality and quality. The RSS feeds input pipeline collects Web articles from about 60 national press agencies and online newspapers, processing more than 2,000 articles per day. At the time of writing, about 4,500,000 articles were processed and stored in a PostgreSQL database running on a Linux server.

An RSS feed contains a list of news article URLs and some associated metadata, such as title, author, description, publication date, enclosed elements, etc. Articles that either are new or have a more recent publication date (if already present in the database) are added to a list of article URLs to be downloaded. Title, enclosed images and publication date are also stored in the database, if included in the corresponding RSS item. Blog posts are

collected by registered users submitting their preferred blog feeds. The daily programming of the major national TV channels is acquired from the DTT broadcasts and monitored to detect newscast programmes. Video elements indicating starting and ending points of newscasts are used as visual prototypes to be automatically searched through the acquired TV streams. Metadata about the detected newscasts, e.g. broadcaster name, channel, start date/time, end date/time and title, are also stored in the database. The legacy archive input stream is connected to the Rai internal multimedia database management system, to collect metadata about historical Rai television and radio programmes. The last input source in the current version of the Active News is the EVN, the European Broadcasting Union (EBU) distribution platform to share audiovisual news footage and metadata between its member organizations.

Data Pre-processing

Ingested contents are processed by content analysis techniques in order to analytically characterize their textual and audiovisual features. The web articles linked by the tracked RSS item URLs are periodically fetched and downloaded. The process is governed by a statistical priority queue scheduler in order to load balancing the number of items per feed at each processing cycle. The complete HTML pages pass through a cleaning step that removes undesired content such as scripts, footers, styles, advertisements, etc. The cleaning process is based on the jusText[2] tool whose outputs are plain text files containing the main content of the news articles. The Apache OpenNLP[3] toolkit is then used to perform sentence detection, tokenization, part-of-speech (POS) tagging and named entity detection from the extracted texts. Automatic classification is applied to categorize the articles according to the journalistic taxonomy used by Rai archivists. All of such information is stored in the database as part of the RSS item metadata. Similarly, blog posts by users are converted to an RSS format and treated accordingly.

Newscast programmes are automatically segmented into stories. Story boundary detection is based on the probabilistic cluster model originally proposed by Di Iulio and Messina (5), because of its demonstrated robustness against different newscast editorial formats. The core algorithm adopts a multimodal approach based on merging information from (audio) speaker and (visual) shot clustering. Detected stories are transcribed by a multilingual speech recognition software. A timestamp is associated to each transcribed word, so that it is possible to find the exact position of each spoken word within the story timeline. TV news stories are classified using the same taxonomy as for the web articles, thus providing additional semantic information that can be used for further filtering and data mining.

Data Clustering and Warehousing

The data clustering module implements a bottom-up approach to identify events and topics. First, RSS items and TV news stories about the same event are grouped together. Since news about the same event are typically produced and published/broadcasted in close time proximity, an interval time window is used to limit the overall amount of processed data. At the core of the method is the hierarchical, asymmetric co-clustering algorithm described by Messina and Montagnuolo (6). Each RSS item is represented by the set of nouns and proper nouns extracted by the POS tagging process from its title and

[2] https://github.com/miso-belica/jusText

[3] http://opennlp.apache.org/

body content. This word set is matched with the speech transcriptions of the TV news stories, thus resulting in a vector in which each element is the relatedness of a TV news story to the RSS item. Based on the computed vector a pseudo-semantic affinity (6) is calculated to cluster RSS items. Because of the asymmetric nature of the algorithm, cause and effect relations between the clustered elements can be defined. As a result, an event is modeled as a directed graph; nodes represent RSS items labeled by the titles of the corresponding Web articles, and edges are relations such as entailments (i.e. inbound arrows), and equivalences (i.e. bi-directional arrows) between them. Node sizes and edges are determined based on (6). Particularly of interest is the in-degree, which denotes the representativeness of the RSS item w.r.t. the event. This criterion is based on the observation that the higher is the in-degree of a node, the higher is the number of nodes that are semantically entailed by it, so that the corresponding article content is expected to be the most complete w.r.t. the semantics of the cluster. Event title and description are determined according to this criterion. As an example, the Expo Milano 2015 inauguration concert took place in Piazza del Duomo the night of 30th April. This was one of the events scheduled to celebrate the opening of the Universal Exposition. Figure 2 shows how the event is represented in the Active News system. The body of the representative article (i.e. the green node in the graph) and some of the included TV stories are also shown. Information about detected events, e.g. included RSS items and news stories, overall categories and entities, is stored in the database as hypermedia dossiers.

In order to group events related to the same topic we perform a second level clustering based on the nouns and named entities in the event clusters. The method is based on the assumption that the more nouns, persons, organizations and locations are shared among events, the higher the probability that they will refer to the same topic. Figure 3 shows an example. The overall topic is the opening of Expo Milano 2015. Nodes in the graph are related events labeled with their representative title. In particular, three main groups of events can be identified, namely the opening ceremony show, the speech of the Italian Prime Minister, and the riots occurred in the city. Among the opening ceremony nodes, the yellow node is related to the concert, as already detailed in Figure 2. Co-occurring nouns and entities are also shown. The clustering process runs every two hours, thus resulting in more than 10 daily updates. This service level agreement was judged to be acceptable by a panel of Rai's journalists and editors.

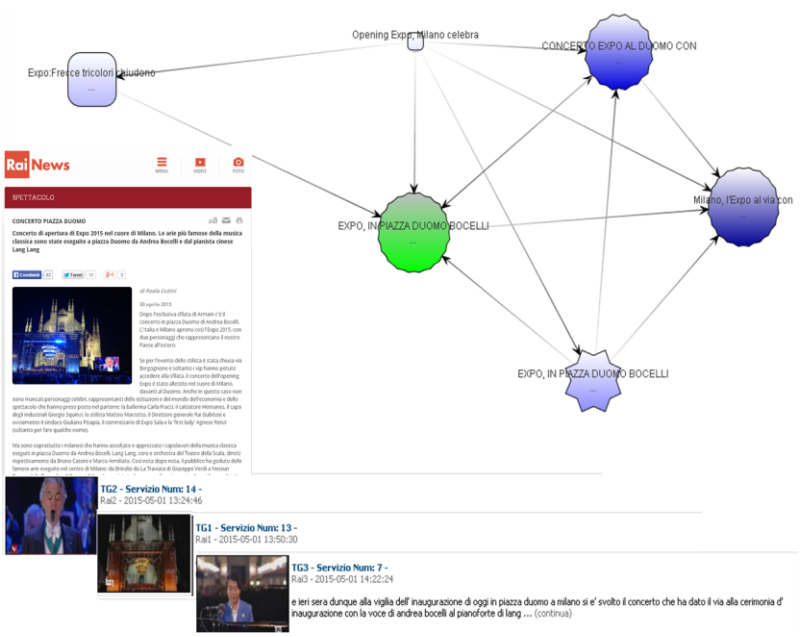

Figure 2 – Graph-based representation of an event. Nodes are Web articles. Edges are relations between them weighted based on their shared TV news.

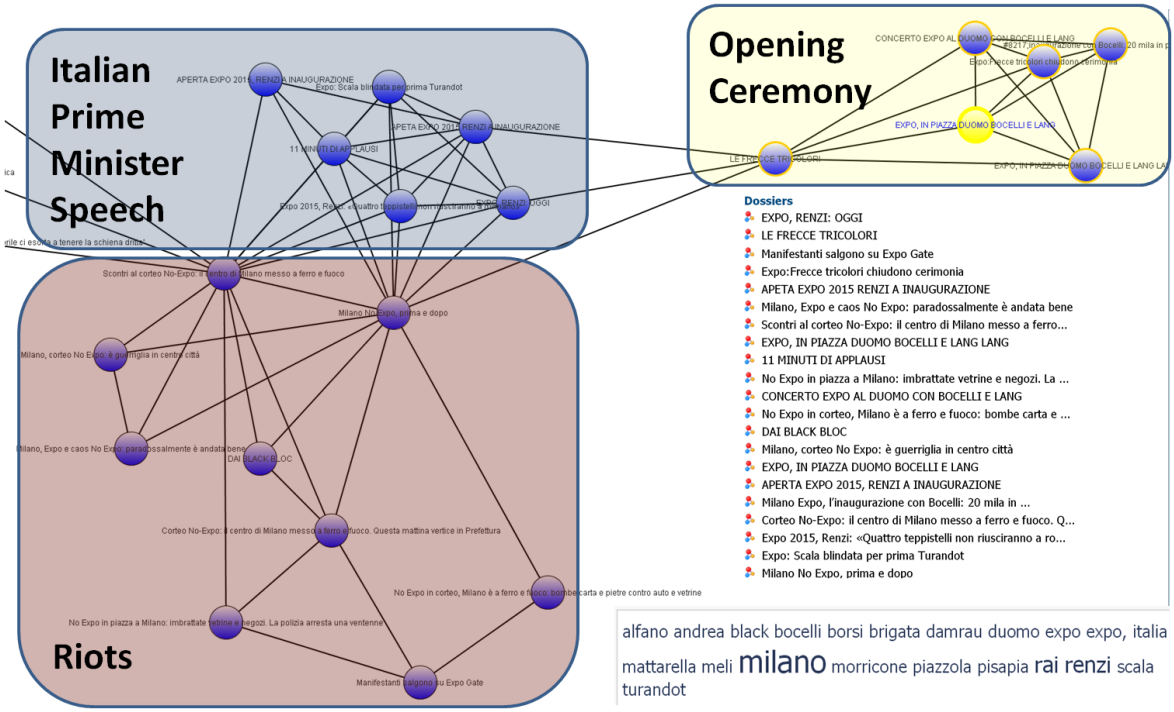

Figure 3 - Example of a topic and related events

The data warehouse (DW) module allows users to create customized statistical and analytical reports about the detected events. The system was designed basing on the Dimensional Fact Model (DFM) principle, and integrating data available from the internal database with external data, such as TV audience scores/categories and social network ratings, e.g. likes, dislikes, shares, etc. Cross-provider statistics, e.g. TV channels and RSS providers correlations, are also identified. These statistics are stored in a PostgreSQL database and are accessible through a Web-based dashboard.

Personalized multimedia news recommendations are generated based on the users' interests. Personal interests are inferred by latent semantic analysis of the users' blog posts, assuming that what they wrote can be interpreted as a reasonably good approximation of their interests. Further details on the core algorithm are given in Di Massa et al (7). Generated recommendations are distributed according to the MPEG-21 User Description[4] (MPEG-21 UD) standard, thus ensuring interoperability and enabling the seamless integration of heterogeneous data sources including user generated contents, i.e. blogs, and professional information items, i.e. online newspapers, press services and TV channels. Experiments with a panel of users demonstrated the quality of the method, getting a precision of about 91% for the proposed recommendations.

Data Index and Access

Data processed and generated by the Active News system are maintained in two DBMS, i.e. the core database and the DW database. Different full text[5] search indexes are constantly updated to provide a scalable, flexible and fast search and retrieval interface to

[4] http://mpeg.chiariglione.org/standards/mpeg-21/user-description

[5] http://lucene.apache.org/solr/

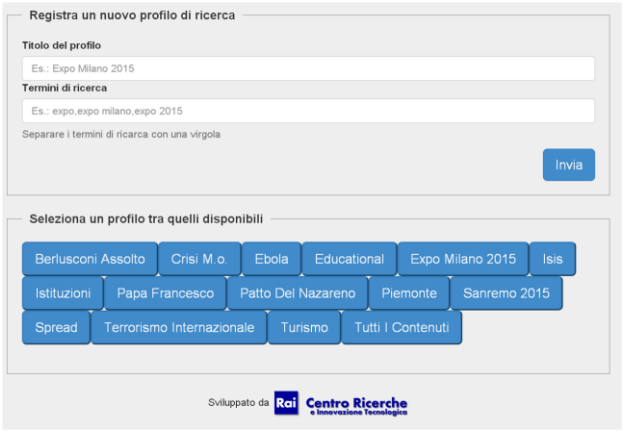

Figure 5 – Active News home page

input/output data and metadata. RESTful APIs and an HTML5 Web portal were developed to support search, filtering, browsing and export of data. The search interface enables users to look for information covering international, national or local news. Listed on the home page is a selection of search profiles users can subscribe (see Figure 5). New profiles can be created entering a name for the new profile, and a list of concepts (entities, keywords or phrases). Once a profile is selected, or a new one is created, the interface shows the list of search results organized in different tabs, each corresponding to one search index. The list of results can be filtered by date and sorted by date or relevance to the profile. Figure 4 shows the TV news search results for the profile 'Expo Milano 2015' in the last year. In order to get the global picture of the context of the profile different panels are displayed. The panels are interactive, so that, when selected, data in the visualization is filtered and updated accordingly. Information in each panel can be exported in CSV, XML or JSON format, enabling integration with the most common content management systems. The timeline shows the distribution of the news over time. Peaks of information occur in correspondence with the existence of some related event (see e.g. the peak around May 1st, 2015 opening day of the exposition). The visualization of categories shows the categorization of the profile news w.r.t. the adopted taxonomy.

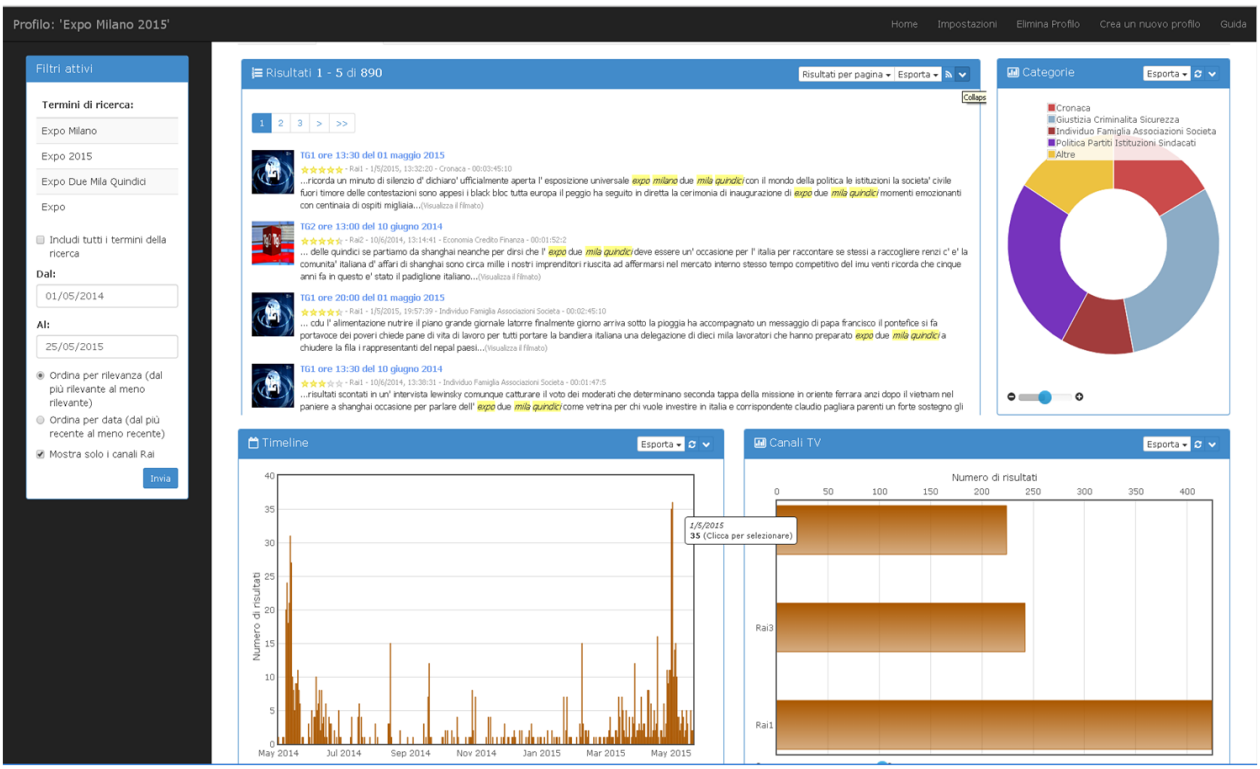

Figure 4 – Example of search results for the profile 'Expo Milano 2015'

Figure 6 – Locations and organizations for the profile 'Expo Milano 2015'

Most important entities are shown accordingly with their relevance w.r.t. the profile (see Figure 6). The geolocation panel shows the coverage of the profile news around the world (see Figure 7).

CONCLUSIONS

This paper described a novel architecture called Active News. The system uses multimodal content sources to aggregate, link and organize data by news topics. Semantic and statistical information about aggregated contents is provided, thus going towards the requirements of modern data-driven journalism applications. The use of open source tools and the complete automation of the processing pipeline minimized the overall costs of the system, while meeting the requested service level requirements. Research plans for the future include the integration of multilingual functionalities, the creation of a knowledge base to represent and share topics' information and relations, and further retrieval facilities, such as visual search.

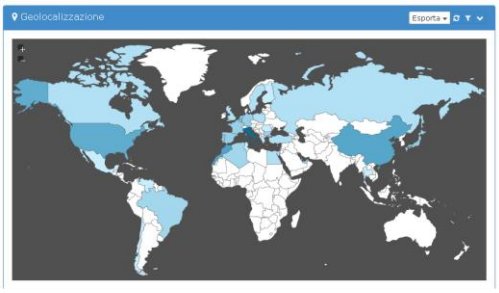

Figure 7 – World coverage of the profile news

REFERENCES

1. Athey, S. and Mobius, M., 2012. The impact of News Aggregators on Internet News Consumption: the Case of Localization. Working paper, Harward University.
2. Liu, J., Dolan, P. and Pedersen, E. R., 2010. Personalized news recommendation based on click behaviour. IUI '10: Proc. of the 14th Intl. Conf. on Intelligent user interfaces. pp. 31 to 40.
3. Li, L., Chu, W., Langford, J., and Schapire, R.E., 2010. A contextual-bandit approach to personalized news article recommendation. Proc. of the 19th Intl. Conf. on World wide web (WWW '10), pp. 661 to 670.
4. Leban, G., Fortuna, B., Brank, J., and Grobelnik, M., 2014. Event registry: learning about world events from news. Proc. of the companion publication of the 23rd Intl. Conf. on World wide web companion (WWW Companion '14), pp. 107 to 110.
5. Di Iulio, M. and Messina, A., 2008. Use of Probabilistic Clusters Supports for Broadcast News Segmentation. DEXA Workshops, pp. 600 to 604.
6. Messina A, and Montagnuolo M., 2011. Heterogeneous data co-clustering by pseudo-semantic affinity functions. Proc. of the 2nd Italian information retrieval workshop.
7. Di Massa, R., Montagnuolo, M. and Messina, A. 2010. Implicit news recommendation based on user interest models and multimodal content analysis. In Proc. Of the 3rd Intl. Workshop on Automated information extraction in media production. pp 33 to 38.