The *RaiNewsbook*: Browsing Worldwide Multimodal News Stories by Facts, Entities and Dates

Maurizio Montagnuolo
RAI Radiotelevisione Italiana
Centre for Research and Technological Innovation
Corso Giambone 68, I-10135 Turin, Italy
maurizio.montagnuolo@rai.it

Alberto Messina
RAI Radiotelevisione Italiana
Centre for Research and Technological Innovation
Corso Giambone 68, I-10135 Turin, Italy
a.messina@rai.it

ABSTRACT

This paper presents a novel framework for multimodal news data aggregation, retrieval and browsing. News aggregations are contextualised within automatically extracted information such as entities (i.e. persons, places and organisations), temporal span, categorical topics, social networks popularity and audience scores. Further resources coming from professional repositories, and related to the aggregation topics, can be accessed as well. The system is accessible through a Web interface supporting interactive navigation and exploration of large-scale collections of news stories at the topic and context levels. Users can select news topics and subtopics interactively, building their personal paths towards worldwide events, main characters, dates and contents.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications—*Information browsers*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*Navigation*

General Terms

Management

Keywords

Tv and Web convergence, News Retrieval, Named Entities Tagging, Topic Tracking

1. BACKGROUND

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.

There have been a number of systems supporting access to news content over the years. Among them, the Google News aggregator[1], the Informedia News-on-Demand library [1] and the Físchlár News delivery system [2]. However, most of the current technologies lack in efficient data organisation and presentation, so that accessing and browsing the desired content can be a time-consuming and frustrating task. Also, only basic facilities for data analysis, filtering and management (e.g. filter by date) are provided.

In this paper, information streams from both television, the Internet and others digital libraries (i.e. a collection of audiovisual assets from the RAI's archive, and a collection of international news from the Eurovision archive[2]) are automatically acquired, aggregated in topics and indexed to provide more integrated access to these otherwise disconnected data sources. Data mining techniques are used to contextualise the news topics w.r.t. involved entities (i.e. persons, geographic locations and organisations) and impact factors (i.e. TV and Internet audience). The system runs 24 hours per day and 365 days per year. About 550 RSS feeds from 70 Web providers are registered and managed by the system. Users can at any time add new feeds. Digital television streams (DTTs) are acquired from the daily programming of 9 national channels, resulting in 21 single newscast programmes per day. The audiovisual assets search index and the international news search index are daily updated and currently, they consist in about 125,000 and 60,000 documents, respectively. To our best knowledge, this is the first attempt to address this problem in a real-world large-scale scenario. The remain of the paper is organised as follows. Section 2 overviews the main technologies of the system. Section 3 describes the interfaces accessible by users. Section 4 provides final remarks and draws conclusions.

2. MAIN TECHNOLOGIES

The *RAI interactive Newsbook* (RaiNewsbook) brings together a range of automatic technologies in a single system that provides many integrated services for multimodal and personalised fruition of informative news content. The proposed platform is modelled as a Multiple-Inputs Single-Output information processing machine. The input data streams are composed of TV video clips from news and entertainment programmes and Internet contents from RSS news feeds and user blogs. The television streams are at first partitioned into programmes. Video elements indicating

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[1] `http://news.google.com/`
[2] Eurovision News Exchanges (`http://www.eurovision.net/news/exchanges.php`)

starting and ending of newscasts are used as reference prototypes to be searched through the acquired video streams. Though this can be performed with less computationally expensive and more general methods, e.g. by electronic programme guides (EPGs), we believe that the visual approach better balances between accuracy and robustness required by an automated system for this purpose. For example, EPGs are not always available, nor they provide accurate timing information. Once a newscast has been detected, automatic segmentation into elementary stories is performed. Then, the audio track of each story is transcribed by a multilingual speech-to-text engine and classified according to a taxonomy of journalistic categories. A complete description of the news story segmentation task and of the validation criteria applied is described in [3]. The Internet contents are extracted from the HTML pages linked by the RSS feeds registered in the system. First, an RSS feed is parsed to identify the embedded items and extract the corresponding information, i.e. title, description, publication date, link and enclosure elements. Next, natural language processing (NLP) is performed on each individual item's title and description. This includes HTML tag cleaning, sentence boundary detection, stop-word removal and part-of-speech (POS) tagging. This information is then used by a hybrid clustering algorithm to link the RSS items to the news story speech transcriptions, thus providing aggregations of different information assets around the same topic [4]. The RSS processing pipeline was developed using open source tools and utilities[3] in order to provide a low cost and flexible solution to the problem.

The output data stream is a set of *multimedia dossiers*, i.e. aggregations of text Web articles and multimedia TV assets around the same topic. Each dossier is supplied with additional information that is automatically extracted from its content, e.g. the title reflecting the semantics of the event and sub-events described by the dossier and the related entities. Furthermore, temporal information is also included, in order to provide the users the ability to browse spatial-temporal paths among real-world events. Named entities are identified through lexical-syntactic patterns (e.g. a continuous succession of proper nouns or proper nouns spaced out by prepositions) combined with external knowledge repositories such as Wikipedia and other ground-truth data sources (e.g. legacy or geospatial databases). This allows us to recognise as many entities as possible (thus augmenting recall) while minimising the number of false positive results (and thus augmenting precision).

The information collected about each dossier is indexed by the Apache Solr engine[4] so that the search engine could offer unified search and browse services for any Web user. These services are described in the next section.

3. SYSTEM INTERFACES

This section introduces the main services delivered by the *RaiNewsbook* system that are accessible to the users. These include the following: (i) Named-entities oriented navigation and multimodal topic content browsing; (ii) Graph-based topic representation; (iii) Cross-domain data warehousing.

3.1 Named-entities Oriented Navigation and Multimodal Topic Content Browsing

The system provides a Web interface for searching and retrieving news topics and events by named entities, as shown in Figure 1. The list of available entities (i.e. those that have been extracted by the named entities recognition engine) is shown on the left panel. Selecting one or more entities the system shows (on the right panel) the list of retrieved topics. Each topic is browsable by title (automatically set by the aggregation algorithm among those of the included RSS feed items), last update timestamp, and contents (i.e. the list of TV news stories and RSS feed items equipped with extracted texts and tag cloud, speech transcriptions and images). The topics can be multilingual (e.g. English and Italian), provided that the speech-to-text tools and the POS tagging tools are available for the languages of interest (see Figure 3). Also the named entities panel is automatically updated in order to produce a tag cloud of new entities that have a relation with the selected ones. For example, selecting the entity "`person:Angela Merkel`" will result in a list of topics (i.e. multimedia dossiers) in which Angela Merkel is involved (e.g. financial crisis, G20 meetings, stock market news) and a list of entities related to her (e.g. Berlin, Frankfurt and Wall Street for locations). The interface was developed using the AJAX Solr javascript library[5] to completely inherit the facilities of Apache Solr into our system application. Different Solr indexes can be accessed at the same time, thus providing an integrated access interface to multiple data sources. In particular, users can select items from the RAI multimedia catalogue, i.e. a legacy digital library of the RAI's audiovisual archives, from the Eurovision news archive and from Google.

3.2 Graph-based Topic Representation

The system offers a visual navigation mode based on the representation of the multimedia dossiers in the form of graphs. Each node of the graph corresponds to a Web article included in the dossier. The edge among a couple of nodes represents the degree of affinity (i.e. topic similarity) between the two articles [6]. This affinity is evaluated using an asymmetric function that derives from a generalisation of the Cosine similarity measure. Further details on the analytical properties of this function are provided in [5]. Figure 4 shows an example. The general topic is about the bad weather in Italy during the first days of November 2011. The graph shows two sub-topics: on the left side there are the news focussing on the disaster in Genoa, while on the right side there are the news on the civil defence operations. Clicking on one node shows the article content and the related TV clips.

3.3 Cross-domain Data Warehousing

Cross-domain data warehousing and reporting provides a set of tools for data management and monitoring. News topics are supplied with statistics on the TV and Internet coverage, interest and pleasantness. Cross-provider statistics, e.g. TV channels correlation, are also provided. Different

[3]HTML cleaning is performed using a customised version of PotaModule (`http://sslmitdev-online.sslmit.unibo.it/`). Sentence detection and POS tagging is performed using OpenNLP (`http://incubator.apache.org/opennlp/`) trained on an in-house data set.

[4]`http://lucene.apache.org/solr/`

[5]A JavaScript framework for creating user interfaces to Solr (`https://github.com/evolvingweb/ajax-solr`).

Figure 1: Example of named entities topic search and retrieval. The left side of the interface shows the selected entities (e.g. "persons:Angela Merkel") and the list of entities to which they are related. The left side of the interface shows all the retrieved news topics and events (i.e. multimedia dossiers) in which the selected entities are involved. For each topic a user can browse both the list of included TV news stories and the list of included newspaper articles.

colours/markers are used to differentiate situations in which statistics are out of range. All these statistics are accessible through a Web-based dashboard showing different bar charts, pie charts or tables, as exemplified in Figure 2. In the following we briefly describe each of them.

- **Topic rating on television channels.** These tools provide evidence of the topic interest by television viewers. These include: (i) average audience by day parts; (ii) total number of news stories by day parts; (iii) average audience versus total number of news stories by day parts; (iv) average duration of the news stories versus the average duration of the whole newscasts by day parts; (v) average audience by channels and day parts; (vi) total number of news stories by channels and day parts; (vii) share by channel.

- **Topic rating on the Internet.** These tools provide evidence of the topic interest by the Internet users. These include: (viii) Facebook statistic (i.e. RSS items like, share and comment counters), popularity (i.e. day-bay-day relevance and growth of the Facebook statistics for a topic).

- **News providers correlations and trends.** These tools provide statistical correlation between the news information. The main target groups for these indicators are project managers, programme managers and monitoring committees.

4. CONCLUSIONS

This paper presents the latest technological innovations in multimodal news gathering and publication offered by the RAI interactive Newsbook. The system provides large-scale indexing, aggregation and browsing of news content coming from different distribution channels, such as television, the Internet and other digital libraries. All the provided services are accessible by Web interfaces, thus making the whole system usable across multiple platforms and devices. Innovative characteristics of the system position itself at the frontier of the technological innovation in automated information extraction and presentation, advancing the state of the art of all available commercial solutions in the news domain. Several test analysis were conducted in order to prove both the effectiveness of the system algorithms in a large scale production environment as well as the friendliness of the user-interfaces, as detailed in [4, 6]

5. REFERENCES

[1] A. G. Hauptmann and M. J. Witbrock. Intelligent multimedia information retrieval. chapter Informedia:

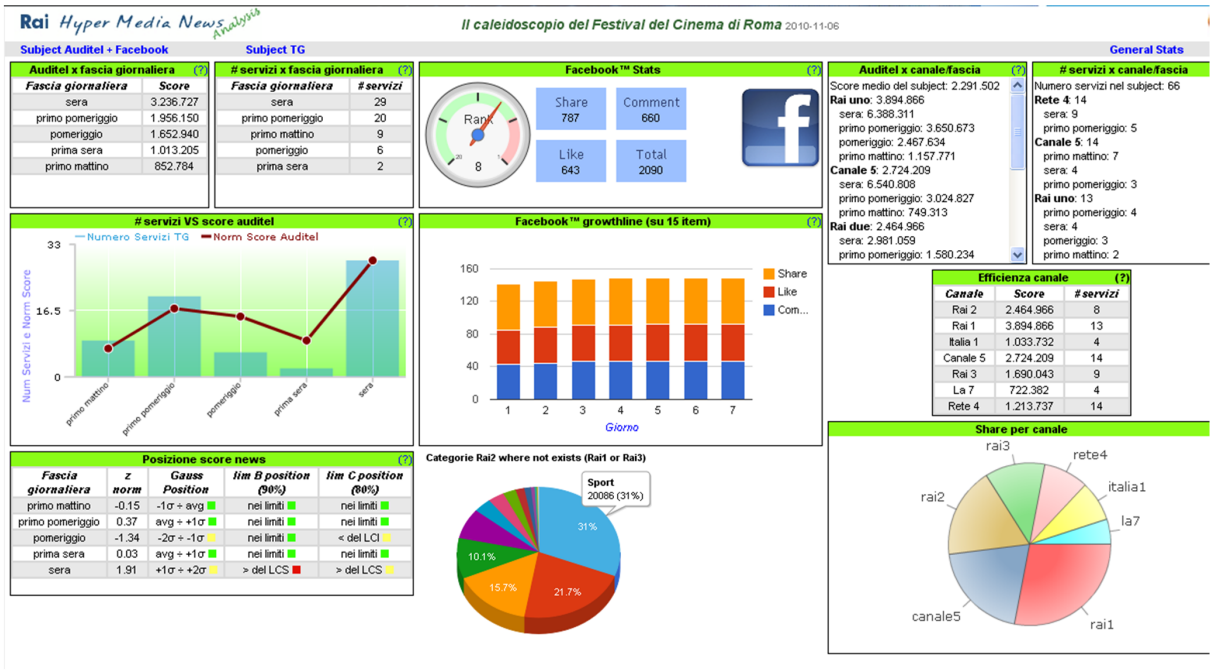

Figure 2: Example of cross-domain data warehousing and reporting. Several statistics about the topic on both television channels and Internet information channels are collected, analysed and presented, in order to produce browsable reports about trends, popularity and user interest of a topic over time.

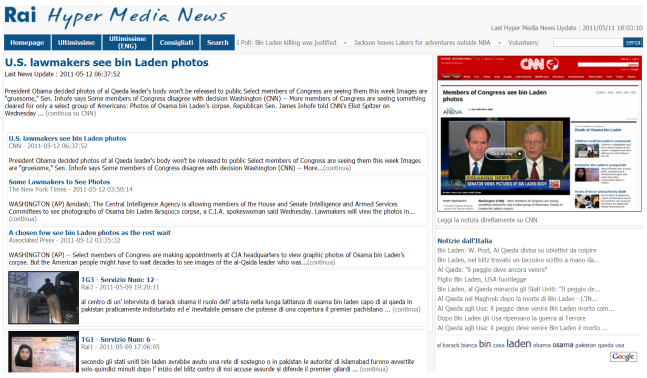

Figure 3: Example of topic content browsing. English language newspaper articles are merged with Italian language newspaper articles and Italian newscast stories.

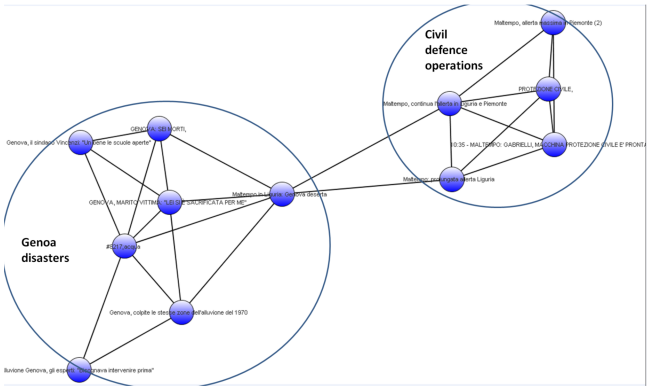

Figure 4: Example of graph-based topic visualisation. The general topic is about the bad weather in Italy. The left side shows the news focussing on the disaster in Genoa. The right side shows the news focussing on the civil defence operations.

news-on-demand multimedia information acquisition and retrieval, pages 215–239. MIT Press, 1997.

[2] H. Lee, A. F. Smeaton, N. E. O'connor, and B. Smyth. User evaluation of físchlár-news: An automatic broadcast news delivery system. *ACM Trans. Inf. Syst.*, 24:145–189, April 2006.

[3] A. Messina, R. Borgotallo, G. Dimino, L. Boch, and D. A. Gnota. An automatic indexing system for television newscasts. In *ICME*, pages 1595–1596, 2008.

[4] A. Messina and M. Montagnuolo. A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 321–330, 2009.

[5] A. Messina and M. Montagnuolo. Heterogeneous data co-clustering by pseudo-semantic affinity functions. In *Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop*, 2011.

[6] A. Messina, M. Montagnuolo, R. Di Massa, and R. Borgotallo. Hyper media news: a fully automated platform for large scale analysis, production and distribution of multimodal news content. *Multimedia Tools and Applications*, pages 1–34, 2011. (Online First: 10.1007/s11042-011-0859-1).