OTMedia: L'Observatoire TransMedia

Nicolas Hervé[1], Marie-Luce Viaud[1], Jérôme Thièvre[1], Agnès Saulnier[1],
Julien Champ[2], Pierre Letessier[1], Olivier Buisson[1], Alexis Joly[2], Benjamin Renoust[1]

nherve@ina.fr, mlviaud @ina.fr

[1] INA, 4 avenue de l'Europe
94366 Bry Sur Marne, France

[2] INRIA Zenith, 161 rue Ada
34095 Montpellier, France

ABSTRACT

Who said What, Where and How? How are images, video and stories spreading out? Who produces the information? OTMedia addresses these questions by collecting, enriching and analysing continuously more than 1500 streams of French media from TV Radio, Web, AFP, and Twitter. Two studies on media produced by end users with the OTMedia framework are presented.

Categories and Subject Descriptors

H2.8. Data mining, H.2.4. Multimedia Databases H.3.1. Indexing methods, H.3.3.Clustering, Retrieval models

General Terms

Experimentation, Algorithms

Keywords

Multimedia, Datamining, Indexing Retrieval, natural language processing, visual search

1. INTRODUCTION

The world of media is strongly impacted by the last decade's digital evolutions: means of production, publishing, broadcasting broaden and users' practices evolve, changing on the way established economic rules. Social networks, Wikipedia, tweets may report news events before traditional actors of the domain, questioning their role in this new landscape. Media studies focus more often on a single media [1], on small data sets [2], or on a specific aspect [3]. The ambition of the TransMedia Observatory is to make the relationship between the Internet, press, radio and television broadcasts and twitter more intelligible, viewable and searchable, at a macro-level of analysis. The paper describes briefly the framework of the project and presents two studies developed by sociologists and professional users.

2. DATA COLLECT AND ENRICHMENT

The collect gathers more than 1500 streams of French media (12 TV, 8 radio, AFP contents, web RSS streams (90 press sites, 339 blogs, 62 institutional sites, 263 political party, 646 politicians, 14 radio, 17 TV, 19 ONG), that represent 4 million documents from July 2011 to December 2013. Collect, stream segmentation and archiving processes are based on the tools developed by the French Legal Deposit for Radio, TV and the Web. Each document's contents are then segmented and analyzed according to their modalities (text, image or video). Named entities, salient words and citations are extracted for textual data. Videos are segmented with image content detectors and images are described with up to 5000 SIFTS per image. Audio files are transcribed. This metadata is then stored in 3 indexes: an SQL DB, SOLR and a specific visual index [4]. Twitter is collected with a specific policy: a set of initial twitter accounts dealing with news has been selected as a kernel. The collect adapts to follow emerging tags related to news events.

3. TEXT AND VISUAL SEARCH

Documents and sources may be accessed by textual or visual requests from a dedicated web interface (cf fig 1). Partial visual queries may be performed on 6 millions images by selecting interactively a zone in the current image. Counts on specific fields are presented to users to help their understanding of the current set content. Corpuses may be generated, viewed and filtered either document by document from the sources or by filtering fields. Moreover, corpuses may be visualized in detail in a chronological timeline.

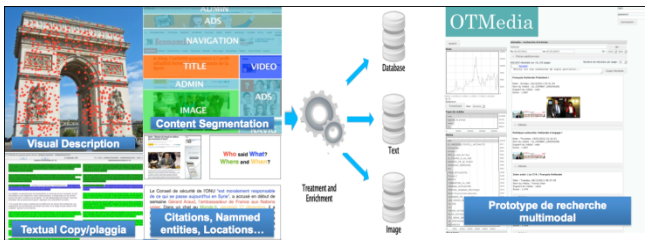

fig 1: Enrichment processes, indexing and the online multimodal search interface.

4. DATAMINING

Visual Object discovery has been performed on about 3000 hours of video and 3 million images. Based on the technology described in [6], a dedicated sampling method associated with a precise CBIR system process generates a graph of visually similar zones in the whole corpus. The graph is then clustered, allowing the emergence of sets of contents that share visual objects (cf fig 2). Finally, textual labels are computed for visual clusters with classical *tf/idf* word saliency on the associated text content. Clusters that represent a good dispersion on media and content providers reflect visual news events while other characterize ads or specific visual attributes of media channels (logo or settings).

Copy and plagiarism detection are computed to study not only citation propagation but also how much AFP provides news to the press, television and radio stations (cf fig 4).

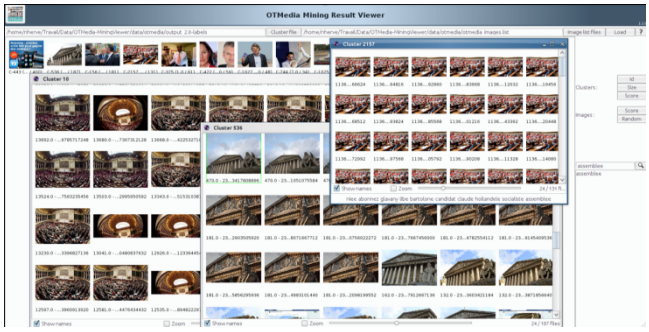

Fig2: Search on "National Assembly" gathers visual clusters illustrating articles on laws and deputy work

5. USER STUDIES

The prototype has been used for sociological and professional studies. The ONU-Alliance of Civilization was interested to know how media in France were talking about immigration in the year of the presidential campaign. The analysis of the events linked to this thematic (fig 3) and a manual tagging of the tone of the articles conclude that a) immigration is a hot topic during electoral campaigns, and b) In general journalists report on migration in a politically correct fashion (according to the Universal Declaration of Human Rights). Unlike the previous elections, the major topics were not security and terrorism but the vote for foreigners in local elections and the eventual limitation of economic immigration.

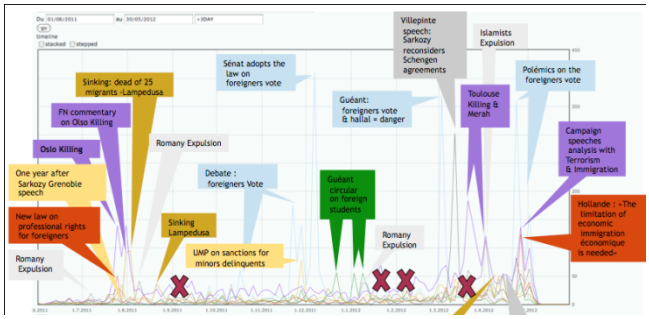

Fig3: Events linked to immigration in France in 2011/12. Notice the increase of the peaks 2 months before the presidential elections (last part of the graphic). Red crosses mark events that occur abroad and were out of the scope of the study.

The study, based on the copy rate of AFP for 5 free online web sites (fig 2), shows the content policy of these news providers. 67% of Orange content comes directly from AFP (notice how the curve is flat), while the number of articles containing AFP content on Yahoo France is very low and applies to only a very small percentage of the content (Yahoo canceled its AFP subscription in 2012). The curves from the 3 news providers show that Rue 89 produces mainly original articles (only 12% of them contain short AFP content), 20% of "20 minutes" production comes directly from AFP. The "Huffington Post" curve decreases strongly showing that most of the articles borrow some AFP content, but this content is incorporated into original longer articles.

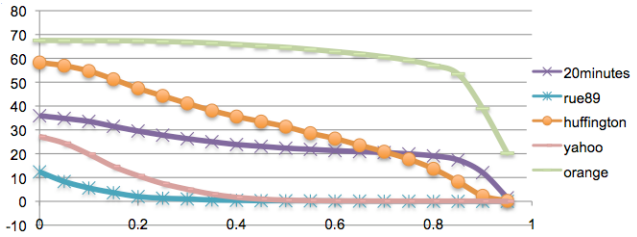

Fig4: The y axis represents the percentage of documents of the media stream, the x axis represents the copy rate. This graphic illustrates the Percentage of articles of French online free sites that borrow a part X of their content to AFP.

The visual object discovery has been used to produce the visual event of the week and to discover political mockery in campaign posters. With this prototype, sociologists are studying the evolution of vocabulary around major events such as the "Toulouse Killing" in March 2012.

More than 10 researchers or professionals are working with the prototype, 40 students are using it to explore how French media treats innovation. This user experimentation shows that the prototype and the functionalities are easy to use.

6. CONCLUSION

The strengths of this project lie in the synergy between IT and social science research, the development of a set of software modules addressing all modalities, together with resources diversity and volume. Current studies already trace the news stories and circulation but further work will be needed to answer more complex questions: is the information diversity increasing? How is news content evolving with acquisitions and editorial board concentration?

7. ACKNOWLEDGMENTS

Our thanks to the other project partners, LIA, Syllabs, AFP, CIM and LATTS, the French National Agency and Cap Digital.

8. REFERENCES

[1] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html

[2] PEW Research Center's Project for Excellence in Journalism, 2010, How News Happens: A Study of the News Ecosystem of One American City

[3] Leskovec ; J Backstrom L Kleinberg J 2009 « Meme-Tracking and the dynamics of th enews cycle » KDD'09 Paris.

[4] P. Letessier, O.Buisson, A Joly, 2012, *Scalable mining of small visual objects* In Proceedings of ACM Multimedia, Pages 599-608,.

[5] A. Joly, O Buisson, Logo retrieval with a contrario visual query expension, In Proceedings of ACM Multimedia pages 581-584, 2009.