OTMedia - The French Transmedia News Observatory

Nicolas HERVÉ – Institut National de l'Audiovisuel (Ina)

FIAT/IFTA Media Management Seminar 2019
Stockholm – Garnisonen Conference Centre – May 23+24th

The information industry is going through the digital revolution. Its practices of production, diffusion, consumption and its economic models are upset, bringing new possibilities, new constraints, but also blurring the roles of the actors.

How does news spread today? What is the Internet's or Twitter's role compared with that of traditional players such as television, radio and the press? Who starts a media "buzz"? Who are the players involved? What place do images occupy in the media? Which is the most frequently broadcast image of the week?

All these questions have arisen with the numerous developments in the news and information sector in recent years and the Transmedia Observatory has attempted to answer them by inviting researchers in IT and social sciences to work together.

OTMedia is a software platform dedicated to research projects that can analyze vast quantities of diverse, multimodal, transmedia data (television, radio, Web, press agencies, Twitter feeds) linked to French and French-language news. The main purpose is to achieve quantitative transdisciplinary research between Human and Social Science and Computer Science (Hervé et al. 2017).

OTMedia permanently collects, processes and indexes over thousands of streams from television, radio, the Web, the press, news agencies and Twitter. The volume and diversity of its collection and performance of its modules make OTMedia a unique platform that incorporates transcription, visual, text and linguistic analysis and cutting-edge data mining software components in order to quantify a number of phenomena such as the spread of information, the importance of copy-pasting in the media or the links between traditional media and social networks.

Although it has already been the subject of very interesting studies in the digital humanities and social sciences, its potential for exploitation has barely begun.

Objectives

The OTMedia project started in 2010. One of its main challenges was to start from the analysis needs expressed by Human Sciences researchers and information stakeholders, and to collaborate throughout the developments to create new concepts, models and tools dedicated to the analysis of the information landscape. This collaboration allowed the creation of a first prototype after two years. The discussions focused on the collection scope, the definition of analysis criteria, use tests (evaluations of results and ergonomics) and the analysis of system biases.

The technological challenge of the project lies in the volume but also in the diversity of the sources of information taken into consideration. In this sense, media corpora have interesting properties for automatic analysis approaches. It was necessary to be able to identify for each piece of information, whatever its original format (image, sound, text), descriptive data, making it possible to identify its properties, its source, but also its coverage in the form of copies (integral or partial) or strong proximity from one medium to another (same subject, same theme...). It is then necessary to carry out different phases of data mining on these automatically enriched, potentially noisy and incomplete data. One of the major concerns of the project (and one of the most complex ones!) lies in the estimation of the cumulative biases produced by all the operations of a task.

The preliminary steps: data enrichment and indexing

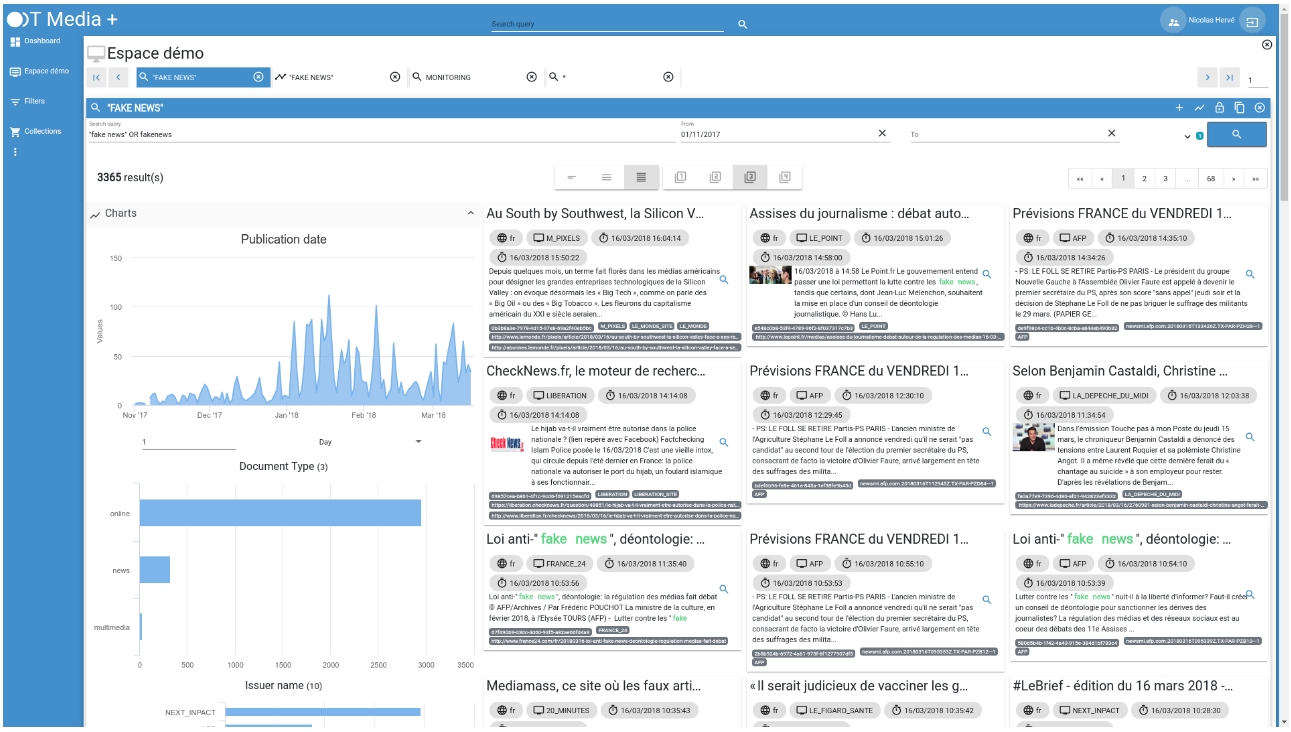
We have implemented several algorithms for automatic data enrichment. Some are generic and well established, others are more specific and developed as needs arise for some studies. The degree of integration of these algorithms into the platform depends mainly on their maturity and the need to be able to benefit from the results in real time. The complexity of the approaches and the computational capacity of the servers also play a role in determining whether an analysis is automatically performed on all the content captured or simply on some corpora for specific studies.

The main tools available are as follows. For audio data (from television, radio and online videos) a transcription is made. We have two speech-to-text software available, allowing us to quantify potential biases in analyses that are related to transcription errors. For images (still images from online sites, social networks or extracted from videos) we use several approaches to index and make queries by similarity. Our indexing engine, developed in-house, allows us to efficiently manage several million images without any problems and thus allows us to search corpora on the scale of all the images produced by the media ecosystem. Finally, many natural language processing methods are used: extraction of named entities, categorization, salient word extraction, quote detection, plagiarism. More specific algorithms are also implemented such as the detection of referencing of a media as a source of information.

Towards interactive data mining

Data mining, or knowledge extraction from data, aims to extract knowledge from large amounts of data, by automatic or semi-automatic methods. In OTMedia, data mining is used to bring out groups, trends, structures or movements from the mass of information available.

The prototype is a dedicated search engine: the user types a simple query, using the standard search bar, or complex, using the widgets to help formulate queries and filters. The documents are displayed in a list, each item in the list being clickable to access the source or its enriched metadata, so that the user can validate its content. In addition to the chronological distribution, summary tables of the results allow the user to have a feedback on the results of the query: number of results per medium, or for the 20 most represented media, number of occurrences of the most salient words, personalities, acronyms or places. These interactive tables allow you to read the entire result, detect anomalies or refine queries. Functionalities for creating, deleting, merging, intersecting and visualizing user corpora are also available.



Search query: "fake news" OR fakenews

From: 01/11/2017

3365 result(s)

Au South by Southwest, la Silicon V...
fr · M_PIXELS · 16/03/2018 16:04:14 · 16/03/2018 15:50:22
Depuis quelques mois, un terme fait florès dans les médias américains pour désigner les grandes entreprises technologiques de la Silicon Valley : on évoque désormais les « Big Tech », comme on parle des « Big Oil » ou des « Big Tobacco ». Les fleurons du capitalisme américain du XXI e siècle seraien...
0b3b8a3e-7978-4d15-97e8-69a2f40eb5bc · M_PIXELS · LE_MONDE_SITE · LE_MONDE
http://www.lemonde.fr/pixels/article/2018/03/16/au-south-by-southwest-la-silicon-valley-face-a-ses-re...
http://abonnes.lemonde.fr/pixels/article/2018/03/16/au-south-by-southwest-la-silicon-valley-face-a-se...

Assises du journalisme : débat auto...
fr · LE_POINT · 16/03/2018 15:01:26 · 16/03/2018 14:58:00
16/03/2018 à 14:58 Le Point.fr Le gouvernement entend passer une loi permettant la lutte contre les fake news, tandis que certains, dont Jean-Luc Mélenchon, souhaitent la mise en place d'un conseil de déontologie journalistique. © Hans Lu...
e548c0b8-53f4-4789-96f2-8f037317c7b3 · LE_POINT
http://www.lepoint.fr/medias/assises-du-journalisme-debat-autour-de-la-regulation-des-medias-16-03-...

Prévisions FRANCE du VENDREDI 1...
fr · AFP · 16/03/2018 14:35:10 · 16/03/2018 14:34:26
- PS: LE FOLL SE RETIRE Partis-PS PARIS - Le président du groupe Nouvelle Gauche à l'Assemblée Olivier Faure est appelé à devenir le premier secrétaire du PS, après son score "sans appel" jeudi soir et la décision de Stéphane Le Foll de ne pas briguer le suffrage des militants le 29 mars. (PAPIER GE...
de9f98c4-cc1b-4b0c-8cba-a844eb490b32 · newsml.afp.com.20180316T133426Z.TX-PAR-PZH28--1 · AFP

CheckNews.fr, le moteur de recherc...
fr · LIBERATION · 16/03/2018 14:14:08 · 16/03/2018 14:14:08
Le hijab va-t-il vraiment être autorisé dans la police nationale ? (lien repéré avec Facebook) Factchecking Islam Police posée le 16/03/2018 C'est une vieille intox, qui circule depuis l'été dernier en France: la police nationale va autoriser le port du hijab, un foulard islamique à ses fonctionnair...
69857cea-b861-4f1c-9cd6-f691215eacfd · LIBERATION · LIBERATION_SITE
https://liberation.checknews.fr/question/48851/le-hijab-va-t-il-vraiment-etre-autorise-dans-la-police-nat...
http://www.liberation.fr/checknews/2018/03/16/le-hijab-va-t-il-vraiment-etre-autorise-dans-la-police-na...

Prévisions FRANCE du VENDREDI 1...
fr · AFP · 16/03/2018 12:30:10 · 16/03/2018 12:29:45
- PS: LE FOLL SE RETIRE Partis-PS PARIS - L'ancien ministre de l'Agriculture Stéphane Le Foll a annoncé vendredi qu'il ne serait "pas candidat" au second tour de l'élection du premier secrétaire du PS, consacrant de facto la victoire d'Olivier Faure, arrivé largement en tête des suffrages des milita...
bdef6b96-fe8e-461a-845a-1ef38fe9b45d · newsml.afp.com.20180316T112945Z.TX-PAR-PZD84--1 · AFP

Selon Benjamin Castaldi, Christine ...
fr · LA_DEPECHE_DU_MIDI · 16/03/2018 12:03:38 · 16/03/2018 11:34:54
Dans l'émission Touche pas à mon Poste du jeudi 15 mars, le chroniqueur Benjamin Castaldi a dénoncé des tensions entre Laurent Ruquier et sa polémiste Christine Angot. Il a même révélé que cette dernière ferait du « chantage au suicide » à son employeur pour rester. D'après les révélations de Benjam...
fa0a77e9-7396-4d80-afd1-942823ef3332 · LA_DEPECHE_DU_MIDI
https://www.ladepeche.fr/article/2018/03/16/2760981-selon-benjamin-castaldi-christine-angot-ferait-...

Loi anti-" fake news ", déontologie: ...
fr · FRANCE_24 · 16/03/2018 11:35:40 · 16/03/2018 10:53:56
Loi anti-" fake news ", déontologie: la régulation des médias fait débat © AFP/Archives / Par Frédéric POUCHOT La ministre de la culture, en février 2018, à l'Elysée TOURS (AFP) - Lutter contre les " fake
67f490b9-d3dc-4d60-93f5-a82ae66fd4a9 · FRANCE_24
http://www.france24.com/fr/20180316-loi-anti-fake-news-deontologie-regulation-medias-fait-debat

Prévisions FRANCE du VENDREDI 1...
fr · AFP · 16/03/2018 10:55:10 · 16/03/2018 10:53:53
- PS: LE FOLL SE RETIRE Partis-PS PARIS - L'ancien ministre de l'Agriculture Stéphane Le Foll a annoncé vendredi qu'il ne serait "pas candidat" au second tour de l'élection du premier secrétaire du PS, consacrant de facto la victoire d'Olivier Faure, arrivé largement en tête des suffrages des milita...
2b8b924b-6972-4a61-975f-6f1277907df5 · newsml.afp.com.20180316T095353Z.TX-PAR-PZB12--1 · AFP

Loi anti-" fake news ", déontologie: ...
fr · AFP · 16/03/2018 10:54:10 · 16/03/2018 10:53:39
Lutter contre les " fake news " nuit-il à la liberté d'informer? Faut-il créer un conseil de déontologie pour sanctionner les dérives des journalistes? La régulation des médias et des réseaux sociaux est au coeur des débats des 11e Assises ...
580d5b4b-1f42-4a43-915e-384d1bf783c4 · newsml.afp.com.20180316T095339Z.TX-PAR-PZB10--1 · AFP

Mediamass, ce site où les faux arti...
fr · 20_MINUTES · 16/03/2018 10:35:43

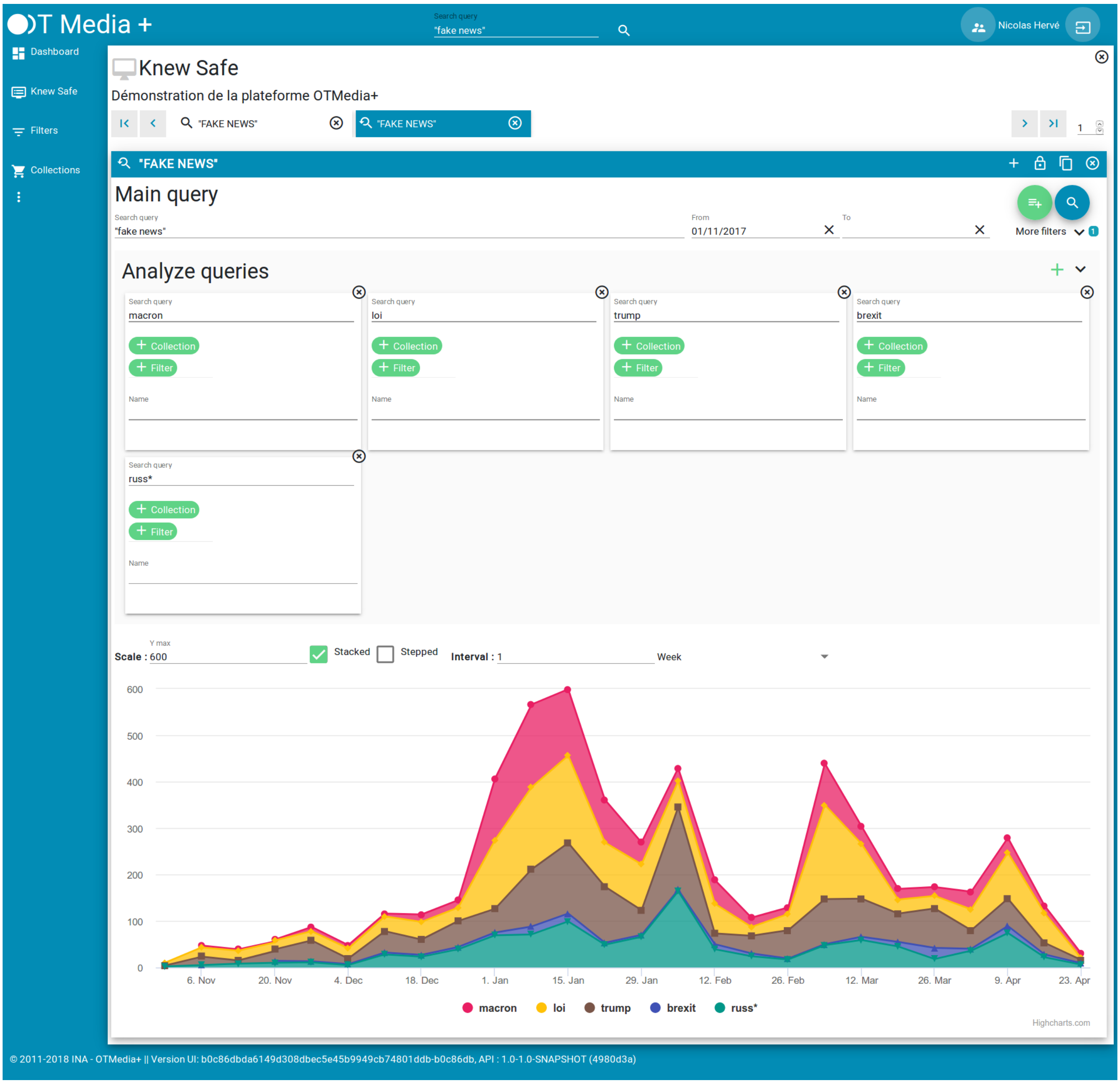
« Il serait judicieux de vacciner les g...
fr · LE_FIGARO_SANTE · 16/03/2018 10:35:42

#LeBrief - édition du 16 mars 2018 -...
fr · NEXT_INPACT · 16/03/2018 10:28:30

OTMedia search engine

Visual searches are performed via an interface activated when an image is selected. An enlarged version of the image is presented to the user, who requests all or part of the selected image with the mouse. The engine finds copies of images or partially similar images. The association of a set of visual copies with the source documents from which they originate makes it possible to study the distribution of a specific image in the media: when and by whom was it used? The selection of small objects such as logos, allows us to group all the images of the same cultural or sporting event with a "covering" (such as the Venice Mostra, the US Open...).

For each query, based on terms, people, media, themes or structuring fields, the interface produces comparative chronologies of occurrences in all the results relating to the global query. For example, the global education query, coupled with the names of presidential candidates, will display the number of times each candidate's name is associated with the term education during the campaign. If candidates are replaced by the media, then the question can potentially be answered: does a media actor more often than others cite the issue of education ?



Search query: "fake news"

Knew Safe

Démonstration de la plateforme OTMedia+

Main query

Search query: "fake news"

From: 01/11/2017

To:

Search query: macron

Search query: loi

Search query: trump

Search query: brexit

Search query: russ*

Y max

Scale : 600

☑ Stacked ☐ Stepped

Interval : 1 Week

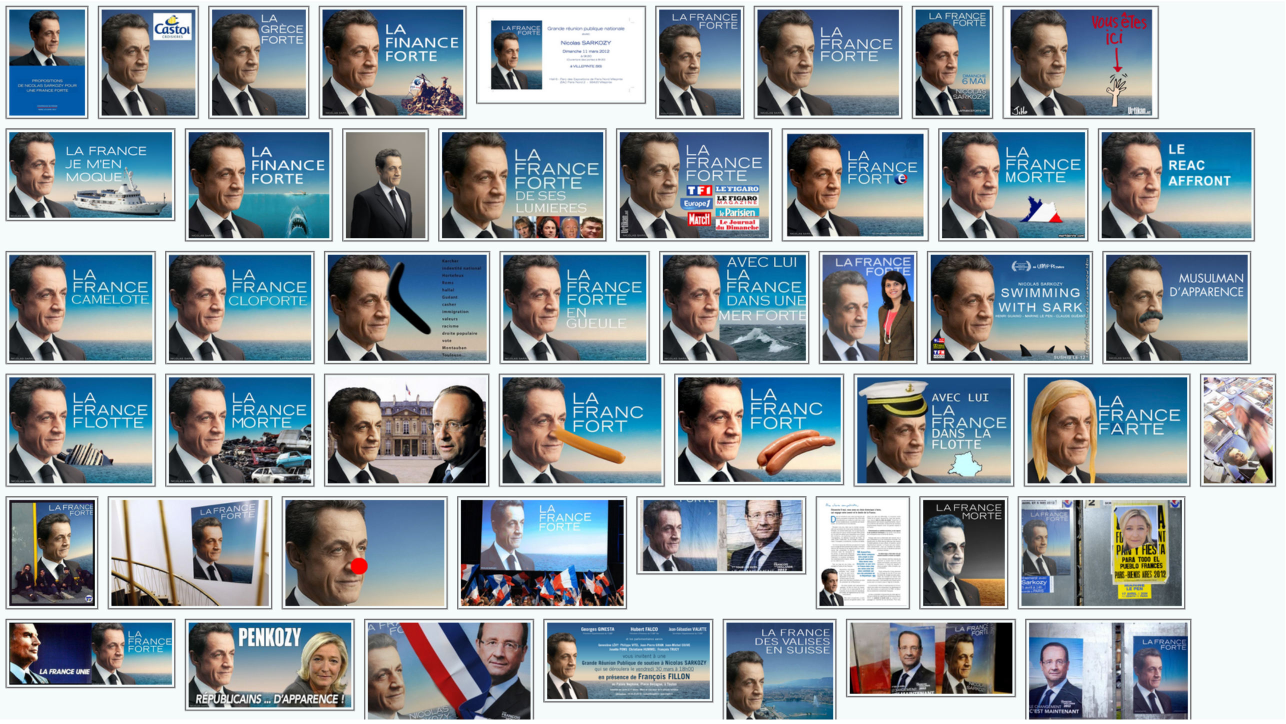
© 2011-2018 INA - OTMedia+ || Version UI: b0c86dbda6149d308dbec5e45b9949cb74801ddb-b0c86db, API : 1.0-1.0-SNAPSHOT (4980d3a)

Temporal distribution of documents

Object emergence: detection of textual and visual events

One of the fundamental tasks of data mining is to group and label sets of "close" items in order to bring out "higher level" objects. The underlying idea is to reduce the volume of data by classifying and prioritizing it in an attempt to perceive its content. Indeed, if the search engine answers the question, "does the database contain elements related to my request ?", the data mining attempts to answer the question: "What is in this database ?". In the context of the project, we studied two types of specific aggregates corresponding mainly to groups of objects with similarities in textual and visual modalities.

To address visual data mining problems, it is important to rely on a powerful visual engine. Research in this area aims to establish strategies to minimize as much as possible the number of requests necessary to create meaningful proximity links between image parts. The strategy is to randomly select a part of an image and request it from the database in order to find the parts of similar images, and to repeat the process in order to cover a "fairly large" part of the visual content of all the images in the database. The tool is mainly used in three different ways. As an interactive visual query engine, it enable the end used to navigate in the huge dataset. As a clustering tool, it automatically groups similar images together. We can, for example, determine the images most diffused by the media (television and web) over a given period of time or follow the propagation of an image on social networks with its multiple modifications (Internet meme). These are “visual events”.


Castor
LA GRÈCE FORTE
LA FINANCE FORTE
LA FRANCE FORTE
Grande réunion publique nationale
Nicolas SARKOZY
LA FRANCE FORTE 6 MAI NICOLAS SARKOZY
Vous êtes ici
LA FRANCE JE M'EN MOQUE
LA FRANCE FORTE DE SES LUMIERES
TF1 LE FIGARO Europe1 LE FIGARO MAGAZINE le Parisien Match Le Journal du Dimanche
LA FRANCE MORTE
LE REAC AFFRONT
LA FRANCE CAMELOTE
LA FRANCE CLOPORTE
LA FRANCE FORTE EN GUEULE
AVEC LUI LA FRANCE DANS UNE MER FORTE
SWIMMING WITH SARK
MUSULMAN D'APPARENCE
LA FRANCE FLOTTE
LA FRANC FORT
AVEC LUI LA FRANCE DANS LA FLOTTE
LA FRANCE FARTE
LA FRANCE UNIE
PENKOZY
RÉPUBLICAINS ... D'APPARENCE !
LA FRANCE DES VALISES EN SUISSE
LE CHANGEMENT C'EST MAINTENANT

Visual query example

The detection of media events from textual documents is carried out in several phases. The most important one is to assess the semantic similarity between documents. From an analysis of the salient words in the documents, we create aggregates that are then merged over time. We also take into account the disparity of the textual elements at our disposal: press articles, audio transcripts, teletext, documentary notes and tweets. The results have been thoroughly evaluated as the “media event” is the central object of most of your further studies.

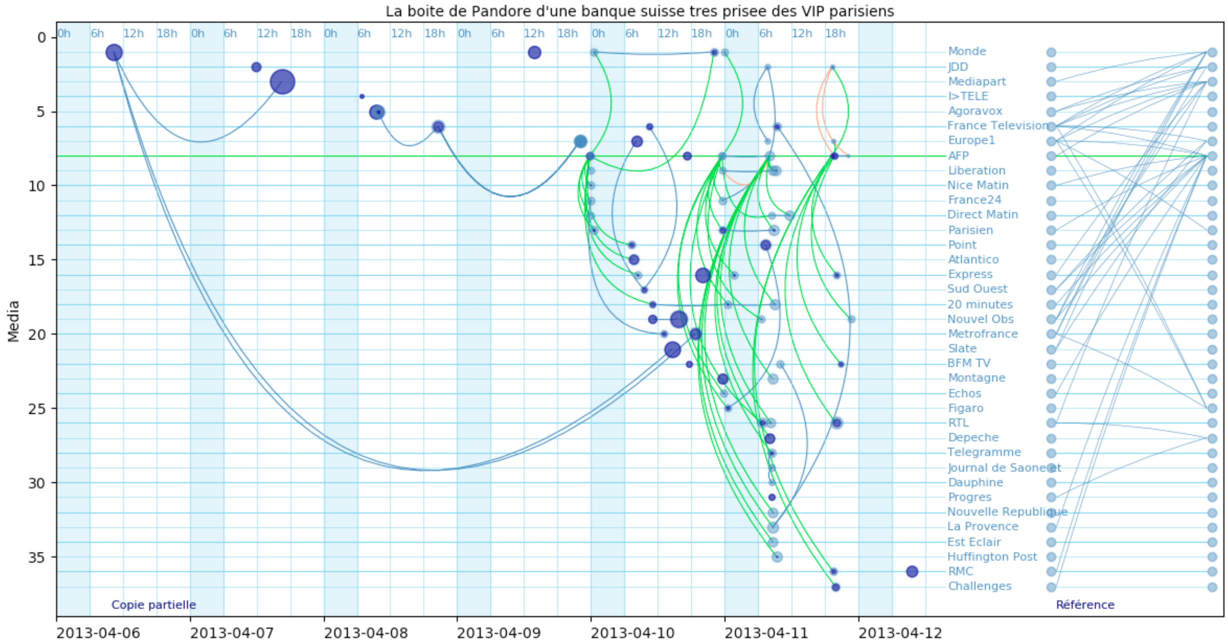
Economic study of the French news ecosystem

Using OTMedia, we try to answer a quite simple question from an economical point of view: “Is there still a ‘commercial value’ of news in this online world ?” In a recent working paper, we document the extent of copying online and estimate the returns to originality in online news production (Cagé et al. 2017).

We use a unique dataset covering the entire news content provided online by the universe of French news media during an entire year (2013). Our dataset covers 87 general information media outlets in France: 2 news agency, 59 newspapers, 10 pure online media, 9 television channels, and 7 radio stations. We track every piece of content these outlets produced online in 2013. Our dataset contains 2.5 million documents.

Using the content produced by news media, we perform a topic detection algorithm to construct the set of news stories. Each document is placed within the most appropriate cluster, i.e. the one that discusses the same event-based story. We obtain a total number of 25,000 stories, comprised of 850,000 documents (about 35 documents per news story).

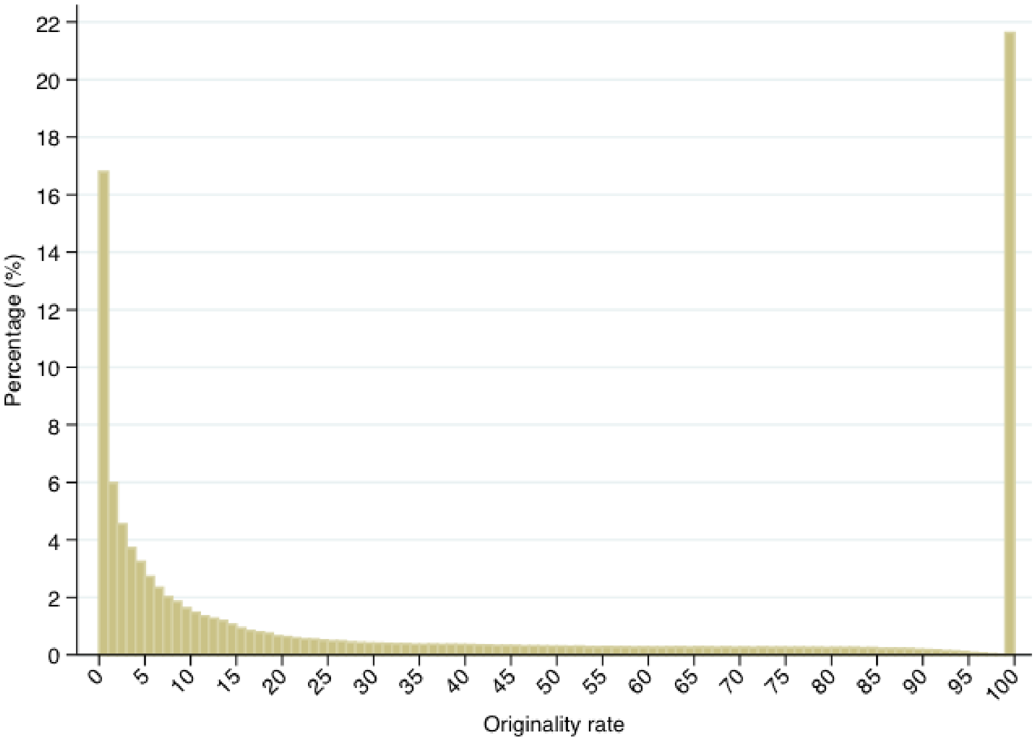
Nearly one third of the news events are about politics, 30% about the economy and less than one quarter about crime, law, and justice. We then study the timeline of each story. In particular, for each story, we determine first the media outlet that breaks the story, and then analyse the propagation of the story, second by second. We investigate the speed of news dissemination and the length of the stories, depending on the topic and other story characteristics.



A single media event with all its documents, plagiarism and source mentions

We show that, on average, news is delivered to readers of different media outlets 172 minutes after having been published first on the website of the news breaker – but in less than 224 seconds in 25% of the cases. The reaction time is shortest when the news breaker is a news agency, and longest when it is a pure online media, most likely because of the need for verification.

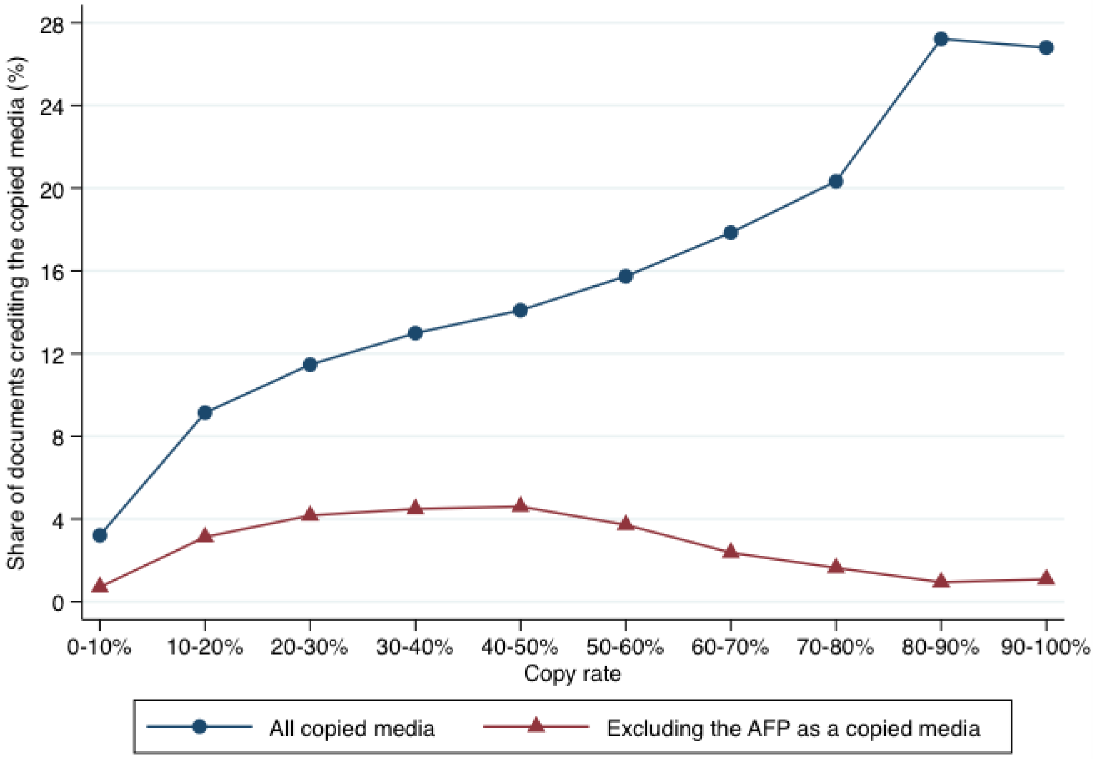
High reactivity comes with verbatim copying. We develop a state-of-the-art plagiarism detection algorithm and find that only 32.6% of the online content is original. The distribution is bimodal, with one peak for the article with less than 1% original content (nearly 17% of the documents) and one peak for the 100% original articles (nearly 22% of the documents). The median is 14%. In other words, with the exception of the documents that are entirely original, the articles published within events consist mainly of verbatim copying – more than 55% of the articles classified in events have less than 20% originality.



Distribution of the originality rate

Obviously, copy can take different forms. First of all, we distinguish external (copying from another media outlets) from internal (copying from a previous article you published) copy. Second, we distinguish content copied from the news agencies and content copied from other media outlets. All the media outlets that are clients of a news agency are indeed allowed to reproduce its content in its entirety, and the business model of the news agency is based on the reproduction of its content by other media outlets.

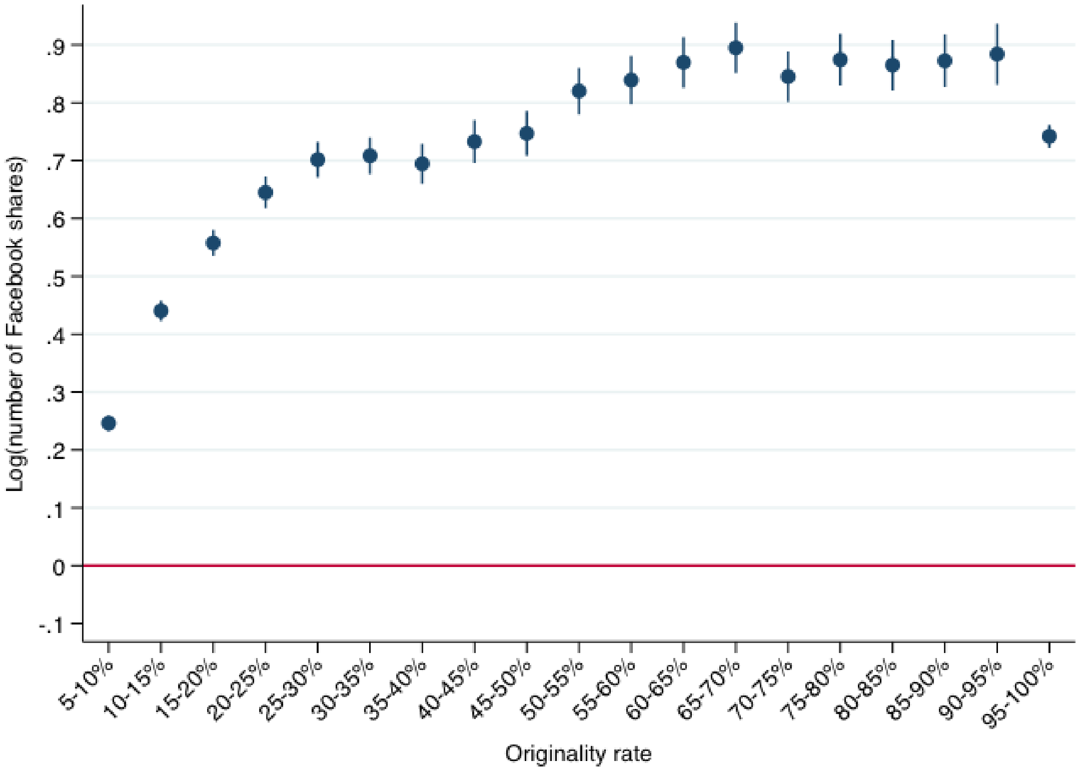
But in effect, every time an original piece of content is published on the Internet, it is actually published three times – once by the original producer, and twice by media outlets who simply copy-and-paste this original content. (Obviously, in practice, we often observe large numbers of media outlets copying part of the content of an original article. But in terms of numbers of original characters copied, this is equivalent to a situation where each piece of original content is published three times.) Moreover, despite the substantiality of copying, media outlets hardly name the sources they copy. Once we exclude copy from the news agency, we show that only 3.5% of the documents mention competing news organisations they copy as the source of the information.



Share of documents crediting the copied media

Do original news producers nonetheless benefit from their investment in newsgathering? In instances where the online audience was distributed randomly across the different websites and regardless of the originality of the articles, our results would imply that the original news producer captures only 33% of the audience and of the economic returns to original news production (which, as a first approximation, can be assumed to be proportional to audience, for example via online advertising revenues). However, we show that reputation mechanisms and the behaviour of Internet viewers allow the mitigation of a significant part of this copyright violation problem.

First, using article-level variations (with event, day, and media fixed effects), we show that a 50% increase in the originality rate of an article leads to a 35% increase in the number of times it is shared on Facebook. This finding is illustrated in the following figure, which plots the estimates of the coefficients from the estimation of the number of times an article is shared on Facebook, as a function of the originality of the article.



Facebook shares depending on originality rate of documents

Second, by using media-level daily audience data and article-level Facebook shares, we investigate to which extent readers ‘reward’ originality. To do so, we compute audience-weighted measures of the importance of originality. As a first ‘naïve’ approach, we assume that all the articles published on the website of an outlet on a given day are ‘equally successful’. Doing so, we find that the average audience-weighted original content is above 46%. This reflects the facts that media outlets with a larger fraction of original content tend to receive more audience.

More importantly, if we weight content by media-level audience shares and article-level Facebook shares, we show that the original content represents up to 58% of online news consumption, i.e. much more than its relative production (33%). This means that within a given media outlet, the articles that get more views (as approximated by the number of Facebook shares) are those with more original content. In effect, reputation mechanisms actually appear to solve about 40% of the copyright violation problem, as long as the media outlets realise this and allocate their effort and journalist time accordingly. The observed collapse in the number of journalists in all developed countries may reflect the fact that some outlets have not.

Of course, greater intellectual property protection could also play a role in solving the copyright violation problem and raising the incentives for original news production, and we certainly do not mean to downplay the extent of this problem. Other factors may help rationalise the observed drop in the number of journalists, the decline of advertising revenues, and the increasing use of ad-blockers to begin with. However, our results suggest that in order to effectively address this issue, it is important to study reputation effects and how viewers react to the newsgathering investment strategies of media outlets.

Conclusion

The collaboration between computer researchers and Human Sciences researchers, which is at the heart of this project, is very rich, even if it sometimes leads to misunderstandings on both sides. First, complex Human Sciences concepts are rarely directly modelable by sets of criteria or measures that can be manipulated by algorithms. Thus, the concept of a "media event" remained a topic of discussion between the partners for much of the project! Nevertheless, the consideration of the multiple dimensions of analysis required in the social sciences and humanities has been productive because it has led to linguistic and visual processing sequences not foreseen at the outset of the project. The evaluation examined the utility and usability of the system: recommendations were collected from users for each version of the prototype. Several functionalities have been added such as managing corpus of results, exporting data, tracking quotations or detecting partial text copies (when a sentence or expression, for example, is transferred from one medium to another). The analysis of the validity of the results made it possible to make a qualitative improvement of some modules of the system. Finally, the users' recommendations on the handling of the prototype have changed the interface: some data have become interactive, the interfaces have been linked together to logically link the operations linked to an analysis task. The user tests made it possible to study the balance between technological automation and the control that must be left to the user.

Finally, the use of the prototype by expert users highlighted the two types of bias in the OTMedia system: bias due to technological processing and bias related to media editorial practices such as, for example, backdating sources or non-compliance with the duty to quote. One of the fundamental aspects of the project with regard to the use of data mining technologies was their validation in well-managed settings in order to measure the biases generated by the tools. Indeed, analysis systems can generate biases at all levels, from description to final visualization, and thus distort the interpretation of results. This is why, in addition to technological innovation, it is the whole methodology of use, in relation to practices, that is the subject of research and experimentation in the implementation of the second phase of the project, currently under development at Ina.

Bibliography

Cagé Julia, Hervé Nicolas and Viaud Marie-Luce (2017), “The Production of Information in an Online World: Is Copy Right?” CEPR Discussion Paper 12066.

Hervé Nicolas, Viaud Marie-Luce, Thièvre Jérôme, Saulnier Agnès, Champ Julien, Letessier Pierre, Buisson Olivier and Joly Alexis (2013), “OTMedia: The French TransMedia News Observatory” ACM Multimedia 2013, Barcelona, Spain.