ANALYSE ET CLASSIFICATION D'EVENEMENTS D'ACTUALITE POUR L'AMELIORATION DE METHODES DE PREVISION

Antoine JEANJEAN (*)(**), **Thibault HANIN** (***)
ajeanjean@bouygues.com, thibault@synthesio.com

(*) e-lab de Bouygues S.A., 40 rue de Washington, 75008 Paris (France),
(**) LIX, UMR CNRS 7161, École Polytechnique, F-91128 Palaiseau (France),
(***) Direction Technique de Synthesio, 8 rue Lemercier, 75017 Paris (France).

Mots clefs :
Prévision, audience, analyse, évènement, impact, classification, clusters, buzz, bruit
Keywords:
Forecasting, audience, analysis, event, impact, classification, cluster, buzz, noice

Résumé

Nous présentons ici une analyse de l'impact d'événements d'actualité sur la publication d'articles en ligne et donc indirectement sur l'audience des sites internet. Ces événements peuvent être des manifestations sportives, des élections, des catastrophes naturelles, des décès de personnalités ou encore des sorties de nouveaux albums, de films ou d'émissions télévisuelles, et plus généralement toute sortie de nouveau produit ou service. Selon leur nature et leur puissance, ces événements ont pour conséquence directe d'influencer le nombre et le type d'internautes connectés sur les sites d'actualité et donc indirectement sur internet. Nous utilisons ici des données collectées quotidiennement par les crawlers de la société Synthesio afin d'identifier le nombre et le type des publications quotidiennes (articles de presse nationale, régionale, billets de blogs, vidéos d'information, etc.). Les événements sont ensuite catégorisés et modélisés afin de fournir un support de prévision. Une bonne catégorisation des événements permet en effet de déduire le comportement probable d'un événement et d'anticiper les montées en charge. Nous présentons également ici les premiers apports qu'elle a pu avoir dans la prévision des audiences internet des sites du groupe TF1.

Abstract

We present an analysis of the impact of news events on the publication of online news articles and thus indirectly on the audience of websites. These events can be sporting events, political elections, natural disasters, deaths of key figures or outflows of new albums, movies or TV shows, or any event related to the release of new products or services in general. Depending on their kind and intensity, these events have a direct influence on the volume and type of users connecting on news websites, and thus indirectly on the internet. We use for this study the data gathered daily by the crawlers of the company Synthesio. This data provides the volume and type of daily publications (national or regional online newspapers, blog posts, videos, etc.). Events are then categorized and conceptualized in order to support forecasting. Proper categorization of events allows to infer the probable pattern of an event and anticipates scalability issues. We also present here the first contributions it may have in predicting internet audiences of the websites of the French TV channel TF1.

1 Introduction

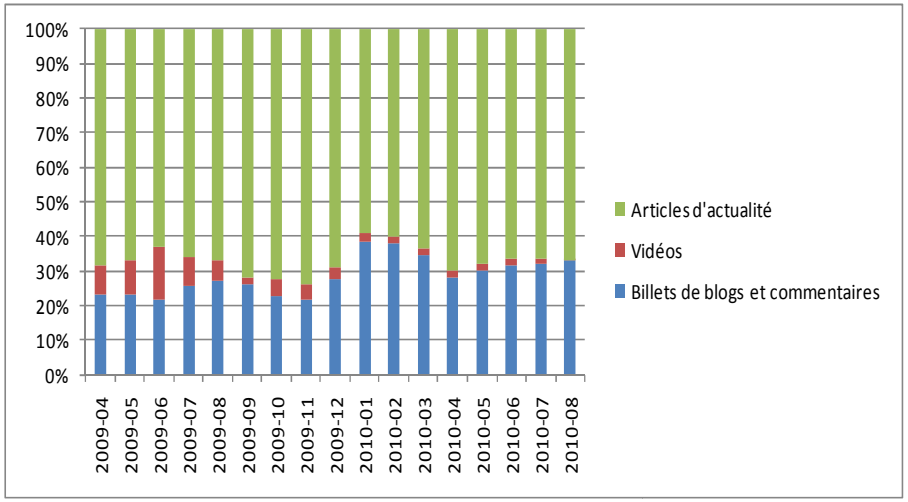
L'analyse réalisée dans le cadre de cette étude porte sur un corpus de sources d'informations Françaises publiant des informations en langue Française. Sont donc exclues du périmètre les sources d'informations francophones dont les auteurs sont situés en Belgique, au Canada, etc. Le corpus est constitué d'environ 43 000 sources d'informations influentes qui constituent un périmètre constant d'analyse tout au long de l'étude :

- **8 sites de presse nationale (exemples : Le Figaro, Le Monde)**
- **110 sites de presse régionale (exemple : Ouest France, Le Parisien)**
- **950 sites d'institutions (exemples : Assemblée Nationale, INSERM)**
- **700 sites de presse économique et professionnelle (exemples : Les Echos, Le Journal du Net)**
- **1500 sites grand public d'information généraliste (exemples : L'Equipe, Le Post)**
- **30 sites d'agences de presse (exemples : Reuters, AFP)**
- **39000 blogs de référence (exemples : Presse Citron, Le Blog Auto)**
- **70 sites de TV et radio (exemples : TF1, RTL)**
- **15 sites de partages de vidéos (exemples : DailyMotion, YouTube France)**

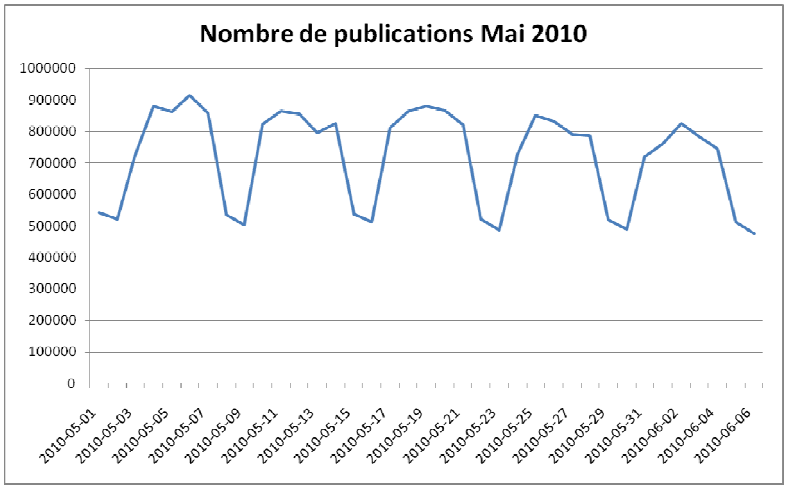
Les commentaires liés aux actualités publiées sur ces différents sites ont également été pris en compte afin de permettre une mesure plus complète du buzz lié aux différents événements étudiés. L'ensemble des informations collectées sur ces différents sites au cours de la période considérée (17 mois) représente plus de 22 millions d'articles d'actualité. La collecte des informations a été réalisée en temps quasi réel (toutes les 15 minutes) par un pool d'environ 1000 crawlers, puis stockée et indexée en intégralité afin de permettre l'exécution des équations de recherche nécessaires à l'étude. Les informations ont été collectées à l'aide de connecteurs propriétaires développés par la société Synthesio. Cette société bénéficie depuis Février 2008 du label de Jeune Entreprise Innovante (JEI) délivré par le Ministère de l'Enseignement Supérieur et de la Recherche pour ses travaux de recherche et développement en matière de collecte d'information sur Internet. Pour chaque information collectée, les données suivantes ont été stockées afin de réaliser des recherches par mots-clés : la date de publication, le titre, l'URL, le contenu intégral, le site d'origine, le type de site, le pays et la langue.

2 Nombre de publications

Les publications mensuelles sont découpées en trois sous-ensembles : les billets de blogs, les vidéos et les articles d'actualité. Nous présentons dans le ***Tableau 1*** et le G***raphique 1*** la volumétrie sur laquelle porte cette étude. On constate une composition relativement constante : le nombre de publications journalières obéit globalement à un pattern très régulier et donc facile à prévoir au fil de l'année, comme le montre le **Graphique 2**, avec une baisse du nombre de publications le week-end.



Graphique 1 : *Volumétrie de publication mensuelle*



Graphique 2 : *Volumétrie quotidienne détaillée à périmètre constant pour mai 2010 (Articles de presse uniquement).*

Mois	Billets de blogs	Vidéos	Articles d'actualité	**Total**
2009-04	237 120	84 531	703 686	1 025 337
2009-05	225 677	91 950	646 872	964 499
2009-06	252 221	173 127	724 257	1 149 605
2009-07	244 209	80 309	628 610	953 128
2009-08	209 581	47 799	515 542	772 921
2009-09	274 152	20 610	753 159	1 047 921
2009-10	289 694	64 629	927 568	1 281 891
2009-11	320 093	60 858	1 085 235	1 466 186
2009-12	369 565	43 789	910 787	1 324 141
2010-01	495 472	33 111	762 840	1 291 423
2010-02	389 566	20 764	620 337	1 030 667
2010-03	415 586	25 015	765 882	1 206 483
2010-04	547 522	37 805	1 348 811	1 934 138
2010-05	563 097	34 471	1 258 933	1 856 501
2010-06	589 587	30 850	1 232 061	1 852 497
2010-07	519 363	27 753	1 080 677	1 627 793
2010-08	476 472	2 455	964 487	1 443 413

Tableau 1 : *Volumétrie de publications mensuelles*

3 Indexation des données

Comme la majorité des systèmes destinés à réaliser une collecte à grande échelle de données issues d'Internet [PAGE98], l'architecture du moteur de recherche développé par la société Synthesio repose sur quatre éléments principaux :

- ***Le crawler,*** chargé de rapatrier les pages web et de les analyser pour en extraire les informations clé, dans notre cas les actualités.
- ***La base de données,*** qui stocke l'ensemble des documents récupérés sous un format exploitable.
- ***L'indexeur,*** qui génère un index plein texte, c'est-à-dire un répertoire des mots-clés extraits du texte intégral des documents de la base de données.
- Le ***requêteur***, ou ***runtime***, qui permet d'interroger l'index et de récupérer pour affichage les documents correspondant à la requête, extraits de la base de données des documents, en appliquant un ensemble d'opérateurs de recherche, notamment booléens.

Les algorithmes spécifiques développés par la société Synthesio se distinguent par leur capacité à extraire de manière automatisée l'intégralité du contenu pertinent des articles en ligne, en ignorant les éléments générateurs de bruit que sont publicités, menus, liens externes et autres informations sans rapport avec l'article en question. Le taux de reconnaissance automatique du contenu intégral atteint ainsi 95%, un résultat particulièrement élevé pour un système en production [LI06]. Enfin, les connecteurs propriétaires de Synthesio s'adaptent à une grande variété de types de pages (actualités avec un lien texte, avec un lien image ou non significatif, ou sans lien), permettant une collecte aussi exhaustive que possible des articles autour d'un thème d'actualité donné, ce qui est nécessaire à notre étude.

4 Analyse d'évènements réels

Nous présentons ici quelques catégories d'évènements que nous avons définies et analysées. Pour chacune d'entre elles, nous en sommes en mesure de proposer des patterns, c'est-à-dire des modélisations simplifiées, de l'impact qu'ont ces évènements sur le nombre d'articles publiés. Pour établir ces modèles, nous nous appuyons sur l'indexation des articles réalisée à partir d'algorithmes utilisant des anti-dictionnaires et permettant d'indexer uniquement les mots pertinents. Nous paramétrons ensuite une requête propre à l'évènement concerné et en déduisons le volume quotidien d'articles publiés sur le sujet.

4.1 Méthodologie

Nos travaux s'appuient sur une méthodologie commune à l'ensemble des types d'évènements analysés. En effet, quatre éléments définissent la "signature" caractéristique d'un évènement :

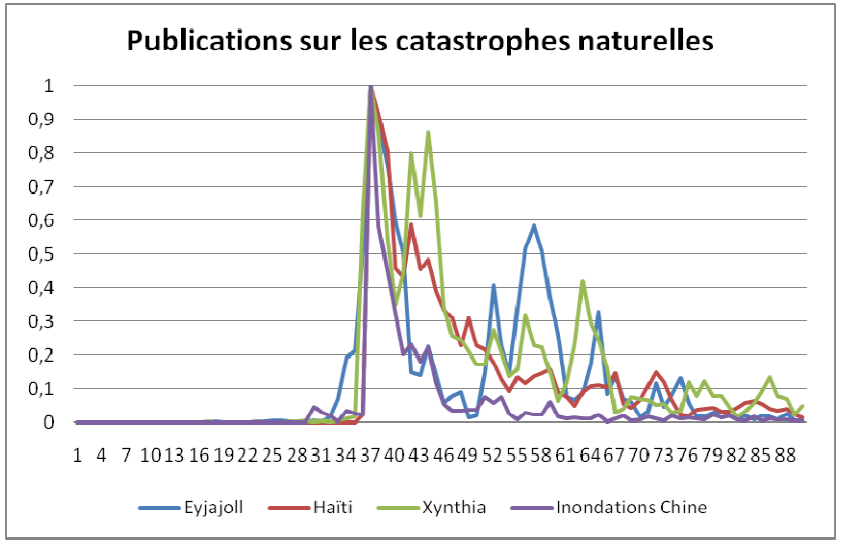
- le volume d'actualités incompressible en l'absence de tout évènement, appelé bruit de fond.
- le volume d'actualité maximal attribuable à l'occurrence de l'évènement, appelé *plateau* pour une manifestation ayant lieu sur plusieurs jours et *pic* pour un évènement ponctuel.
- la courbe de croissance, absente pour les évènements soudains (type décès ou catastrophe).
- la courbe de décroissance.

Notre analyse s'appuie sur un découpage des différents profils d'évènements permettant d'isoler ces quatre éléments. Nos analyses montrent que l'intensité du pic ne peut pas être prédite à partir d'une simple mesure du niveau du bruit de fond avant évènement. Dans le cas d'un décès par exemple, celui-ci peut en effet aussi bien concerner une personne encore active dans l'actualité (c'est-à-dire avec un fort bruit de fond, comme *Philippe Seguin*) qu'une personnalité inactive depuis longtemps (c'est-à-dire avec bruit de fond quasi nul, comme *Bruno Cremer*).

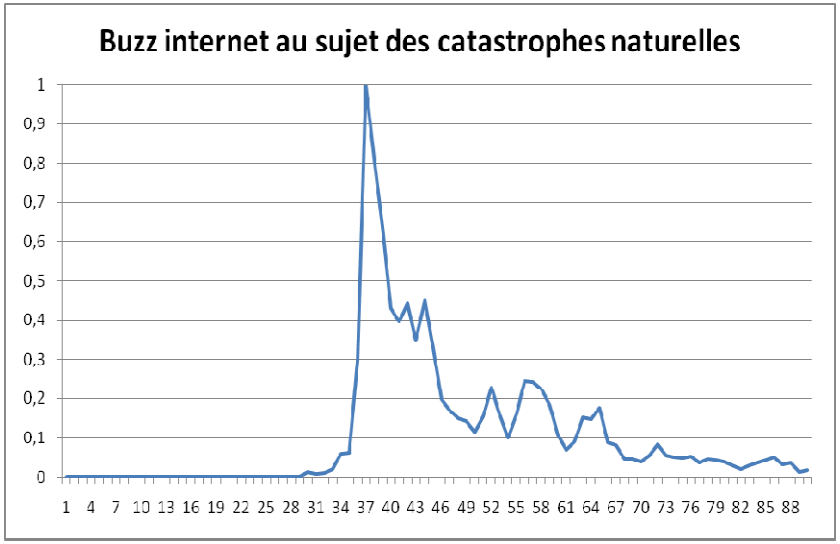
En revanche, nous avons constaté que les lois gouvernant les courbes de croissance et de décroissance des différents types d'évènement peuvent être calculées avec une précision suffisante pour permettre une application au domaine de la prévision d'audience.

Le principe de l'analyse consiste à normaliser l'ensemble des valeurs de volume sur la base de l'intensité du pic (ou du plateau), ce qui permet d'effectuer des comparaisons entre évènements dans une unité commune (max=1). Après normalisation, la moyenne du taux de croissance (respectivement de décroissance) du volume d'articles sur une période donnée (jour ou semaine) nous donne une approximation satisfaisante du profil de l'évènement.

Le **graphique 3** illustre l'approche méthodologique adoptée en présentant les volumes de publications associés à quatre catastrophes naturelles majeures de l'année 2010, normées par le nombre de publications de chacune le jour-même de la catastrophe. Nous en déduisons une courbe moyenne que nous approximons à l'aide d'une fonction exponentielle.



Graphique 3 : *Normalisation du nombre de publications suite à une catastrophe naturelle.*



Graphique 4 : *Normalisation du nombre de publications suite à une catastrophe naturelle.*

Les sections suivantes décrivent dans le détail le calcul des différents profils correspondant aux types d'évènements identifiés.

4.2 Catégories d'évènements

Les 6 types d'évènements identifiés et analysés dans le cadre de cette étude sont résumés dans la **Figure 1.**

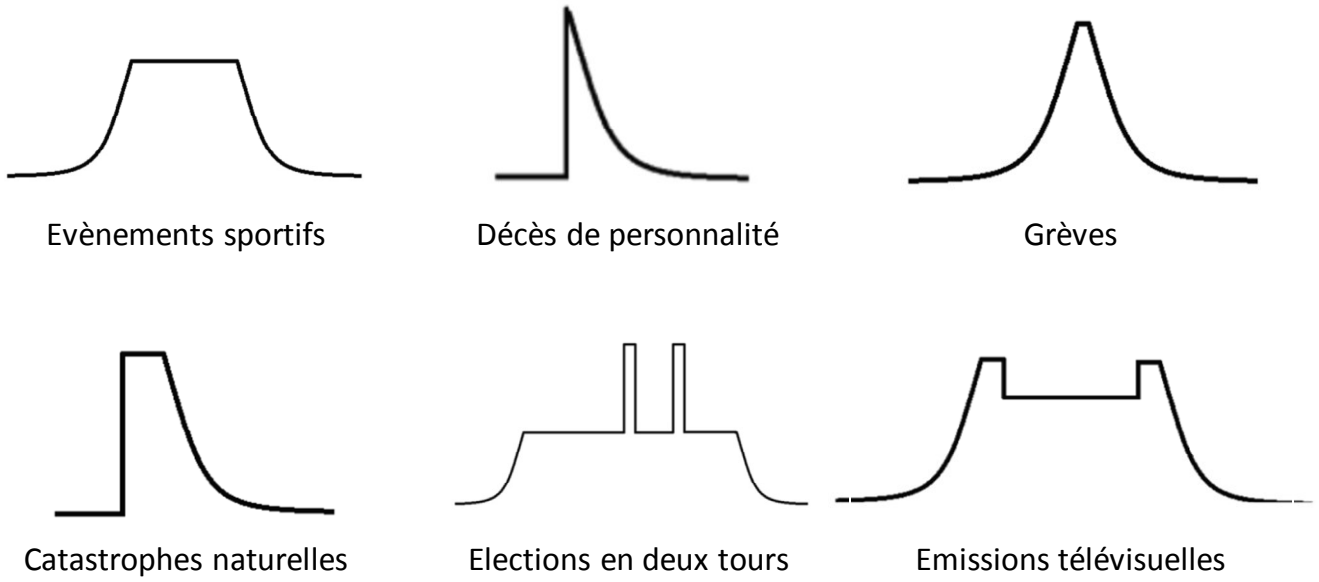


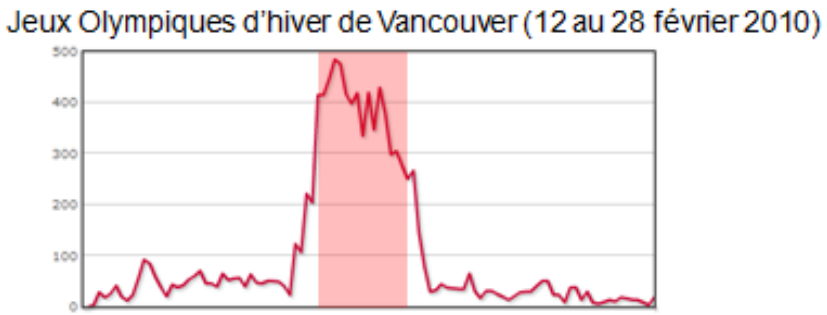
Figure 1 : *Patterns en fonction des types d'évènements d'actualité.*

4.3 Catégorie "Evènements sportifs"

Ce type d'évènement présente une caractéristique spécifique intéressante : en effet, le volume d'actualité est globalement constant sur une période que nous appelons le plateau. La durée du plateau est parfaitement connue puisqu'elle correspond aux dates officielles de l'évènement proprement dit, par exemple :

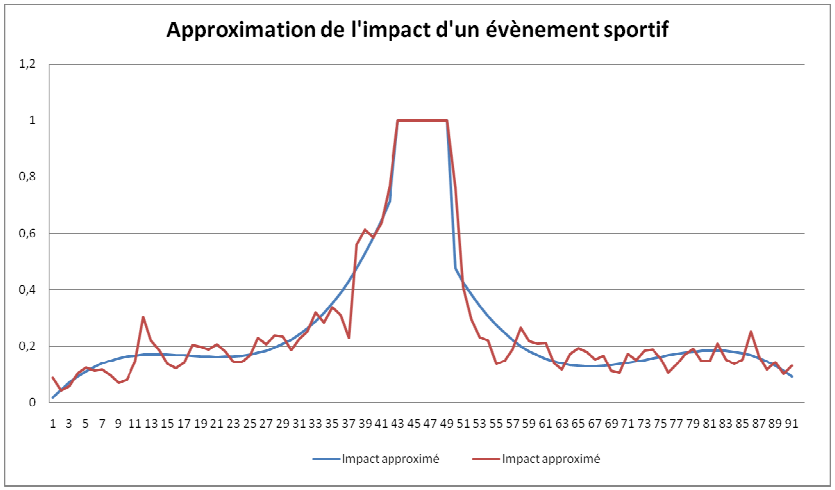
- **Jeux Olympiques d'hiver de Vancouver :** 12 au 28 février 2010 (17 jours)
- **Tournoi de Roland Garros :** 23 mai au 6 juin 2010 (15 jours)
- **Tour de France :** 3 au 25 juillet 2010 (23 jours)
- **Tour d'Italie :** 8 au 30 mai 2010 (23 jours)
- **Coupe du monde de football :** 11 juin au 11 juillet 2010 (31 jours)

Pour ce type d'évènement, la normalisation des données s'effectue donc sur la base de la moyenne du volume quotidien d'actualités publiées lors du plateau. Nous considérons ensuite les deux périodes précédant et suivant le plateau afin d'en extraire respectivement les courbes de croissance et de décroissance.



Graphique 5 : *Nombres de publications durant les jeux olympiques de Vancouver.*

Graphique 7 : *Nombres de publications durant la coupe du monde de football 2010.*



Graphique 6 : *Comparaison de la fonction d'approximation exponentielle du pattern d'un évènement sportif avec les données moyennes réelles.*

Soit E la période de l'évènement sportif, l'impact pour un jour X est donc :

$$\begin{cases} \text{pour } X < J : e^{-0{,}787 - 0{,}076\,X + 0{,}003\,X^2 - 4\,E{-}05\,X^3} \\ \text{pour } X \in E : 1 \\ \text{pour } X > J : e^{-0{,}529 - 0{,}0.56\,X + 2{,}464E{-}03\,X^2 - 3{,}3\,E{-}05\,X^3} \end{cases}$$

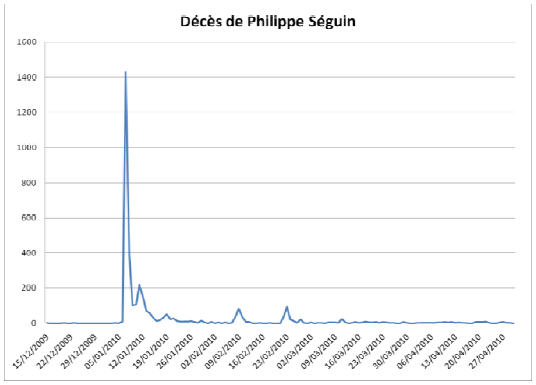
Par approximation, nous utilisons en pratique uniquement les impacts ayant au moins 5% d'impact quotidien, ce qui équivaut à environ prendre uniquement en compte la période J – 40 et J + 40.

Nous avons donc montré dans cette section que nous sommes en mesure de fournir une fonction mathématique représentée sur le ***Graphique 6*** qui modélise la trainée avant et la trainée après la réalisation d'un évènement sportif d'envergure mondial. Nous verrons en section 5 comment cette fonction peut être utilisée dans nos algorithmes de prévisions. Nous détaillerons d'abord les modélisations des autres types d'évènements.

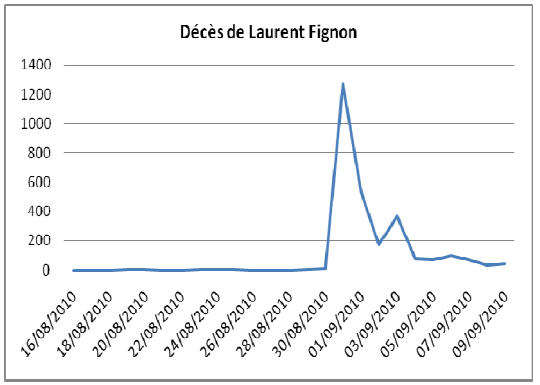
4.4 Catégorie « Décès de personnalité »

La disparition de personnalités génère une activité de publications particulièrement soudaine et importante (dépêches, articles de presse biographique, billets de blogs de fans, republication d'anciennes vidéos, etc.). A titre d'exemple, lors du décès du chanteur Mano Solo survenu en janvier 2010, nous avons réalisé la requête booléenne suivante permettant d'identifier l'ensemble des publications traitant de sa disparition : ("mano solo" OR manosolo) AND (deces OR decede OR mort). La disparition de l'artiste a généré 300 publications dès le premier jour de l'évènement, puis un nombre décroissant d'actualité pendant la semaine suivante. Le mois qui suit l'évènement est encore faiblement impacté par l'évènement, jusqu'à ce que le nombre de publications revienne à un niveau quasi nul. Nous avons réalisé le même type d'étude sur les décès de Jean Ferrat, Denis Hopper, Philippe Séguin et d'une dizaine d'autres personnalités. ***Les Graphiques 8, 9*** et ***10*** présentent le nombre de publications pour trois de ces personnalités :

Graphique 8 : *Nombre de publications sur le décès de Mano Solo en Janvier 2010*

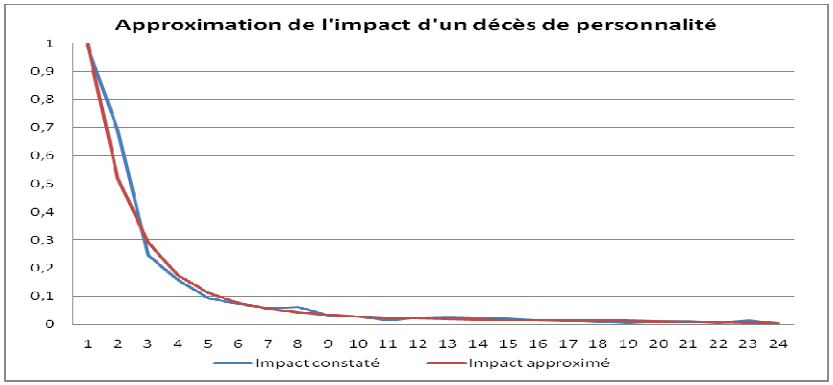


Graphique 9 : *Nombre de publications sur le décès de Philippe Séguin en Janvier 2010*



Graphique 10 : *Nombre de publications sur le décès de Laurent Fignon en Août 2010*

Nous proposons le pattern présenté en **Graphique 11** qui modélise ce type d'évènement ainsi que la fonction de décroissance associée.



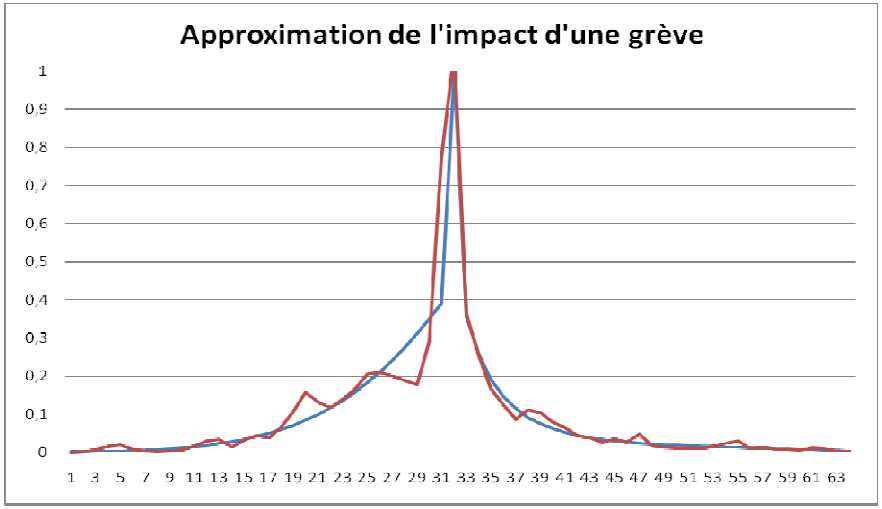
Graphique 11 : *Comparaison de la fonction d'approximation exponentielle du pattern de décès de personnalité avec les données moyennes réelles.*

Soit J le jour de grève, l'impact pour un jour X est donc :

$$\begin{cases} \text{pour X} < J : 0 \\ \text{pour X} = \text{J} : 1 \\ \text{pour X} > J : e^{-0{,}0067 - 0{,}7\,\text{X} + 0.041\,\text{X}^2 - 9{,}1\text{E}-04\,\text{X}^3} \end{cases}$$

4.5 Catégorie « Grèves »

Pour définir le profil de ce type d'évènement, nous nous sommes basés sur l'analyse des différentes grèves et mouvements sociaux (SNCF, RATP, fonctionnaires, chauffeurs routiers, etc…) ayant ponctué l'actualité 2010 et le graphique 12 présente la fonction d'approximation associée.



Graphique 12 : *Comparaison de la fonction d'approximation exponentielle du pattern d'une grève avec les données moyennes réelles.*

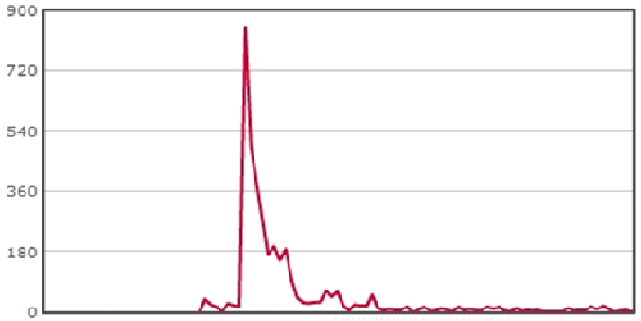
Soit J le jour de grève, l'impact pour un jour X est donc :

$$\begin{cases} \text{pour X} < J : e^{-0{,}831 + 0{,}103\,\text{X} + 0{,}003\,\text{X}^2 - 1{,}59\,\text{E}-05\,\text{X}^3} \\ \text{pour X} = \text{J} : 1 \\ \text{pour X} > J : e^{-\,0{,}682 - 0{,}364\,\text{X} + 0{,}014\,\text{X}^2 - 2{,}47\text{E}-04\,\text{X}^3} \end{cases}$$

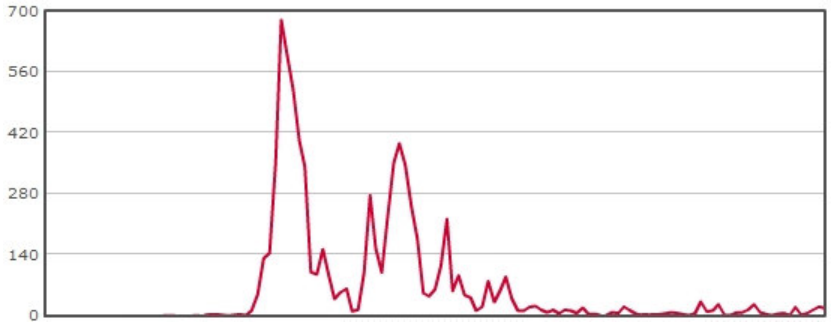
Par approximation, nous utilisons en pratique uniquement les impacts de J – 30 et J + 30.

4.6 Catégorie « Catastrophes naturelles »

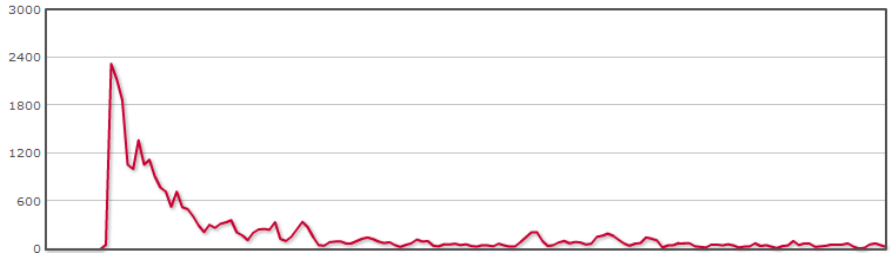
Pour définir le profil de ce type d'évènement, nous nous sommes basés sur l'analyse de quatre catastrophes naturelles majeures de l'année 2010.



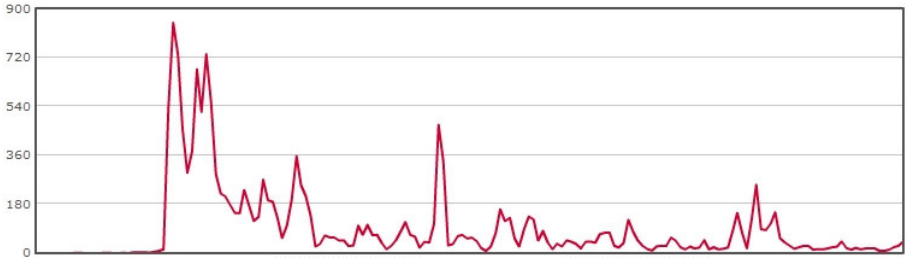
Graphique 13 *: Nombre de publications suite au séisme en Chine*



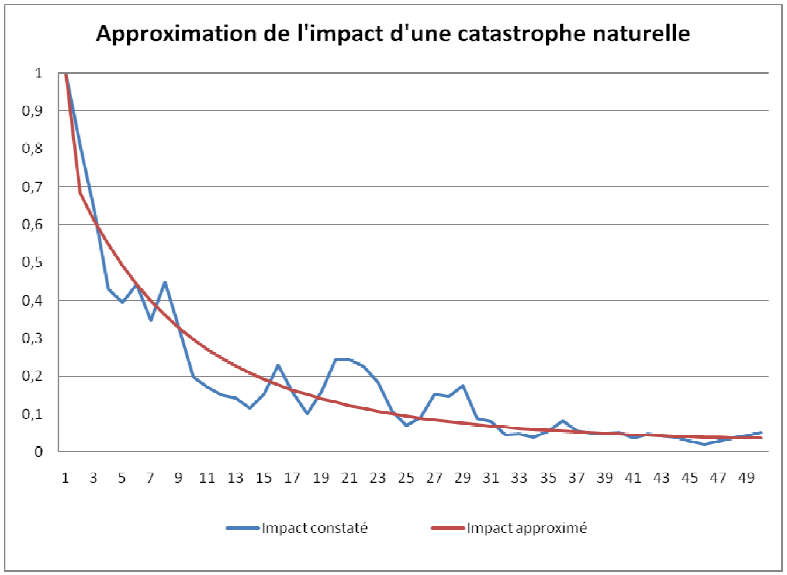
Graphique 14 *: Nombre de publications suite à l'Eruption du volcan islandais Eyjafjoll*



Graphique 15 *: Nombre de publications suite au séisme en Haïti*



Graphique 16 *: Nombre de publications suite à la tempête européenne Xynthia*



Graphique 17 : *Comparaison de la fonction d'approximation exponentielle du pattern d'une catastrophe naturelle avec les données moyennes réelles.*

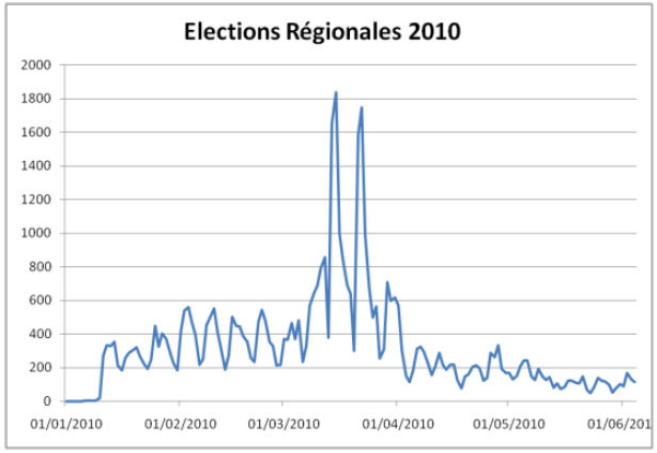
Soit J le jour de la catastrophe naturelle, l'impact pour un jour X est donc :

$$\left\{\begin{array}{c} \text{pour X} < J : 0 \\ \\ \text{pour X} = \text{J} : 1 \\ \\ \text{pour X} > J : e^{-0{,}264 - 0{,}117\,\text{X} + 0{,}0014\,\text{X}^2 - 5{,}36\,\text{E}-6\,\text{X}^3} \end{array}\right\}$$

Par approximation, nous utilisons en pratique uniquement les impacts de J à J + 50.

4.7 Catégorie « Elections en deux tours »

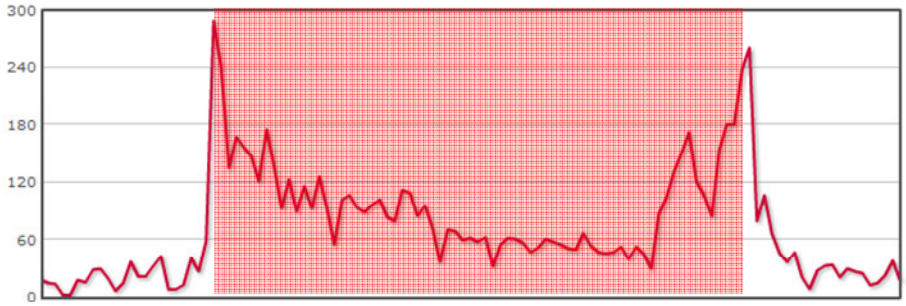
Faute de données suffisante (une seule élection de ce type en France au cours de l'année 2010), nous n'avons pas pu modéliser le comportement de ce type d'évènement, qui présente néanmoins un profil caractéristique.



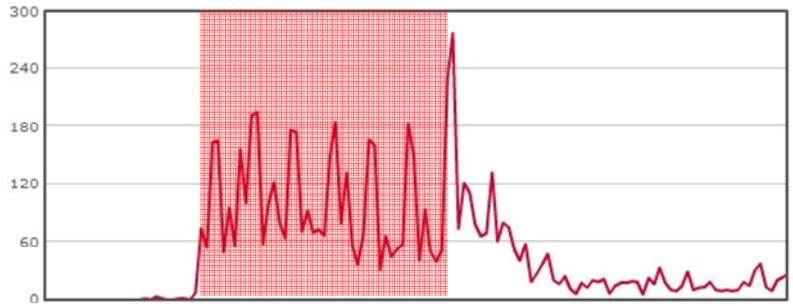
Graphique 18 : *Impact des élections régionales 2010*

4.8 Catégorie « Emissions télévisuelles »

De même, faute de temps, nous n'avons pu modéliser le comportement de ce type d'évènement, qui présente néanmoins un profil caractéristique, comme le montre les 2 graphes suivants ***Graphique 19*** et ***Graphique 20.***



Graphique 19 *: Impact de l'émission La Ferme Célébrités en Afrique diffusée sur TF1 entre le 29 janvier et le 9 avril 2010*



Graphique 20 *: Impact en terme de publications de l'émission Koh-Lanta, le choc des héros diffusée sur TF1 entre le 26 mars et le 21 mai 2010*

4.9 Extrapolations

Bien entendu, tous les évènements ne peuvent pas être classés dans les catégories proposées ci-dessus et nous envisageons à terme de proposer de nouvelles familles d'évènements accompagnées de leurs paramètres d'impacts, notamment en ce qui concerne la sortie de nouveaux produits ou services. Cependant, ces quelques familles permettent déjà de mesurer les impacts d'une part significative des évènements ayant une influence sur publications d'articles en ligne. Il est aussi intéressant de constater que de nombreux évènements d'actualité ont un impact uniquement le jour de leur publication. Ces évènements sont classés dans une catégorie *pic* et leur occurrence (dans le passé, comme dans le futur) est signalée à l'aide d'une indicatrice binaire. C'est le cas par exemple des jours fériés, des évènements religieux ou des vacances scolaires.

Nous allons voir dans la section suivante que ces patterns sont directement corrélés aux patterns d'audiences de sites internet. Les données, qui ne seront pas détaillées ici pour des raisons de confidentialité, proviennent du moteur de prévisions utilisé au sein du groupe TF1 pour prévoir les audiences des sites internet. Nous constatons que les évènements d'actualité génèrent en premier lieu du trafic à destination des sites dédiés à l'actualité puis, en conséquence directe, sur tous les autres sites internet. C'est la raison pour laquelle ces patterns ont été introduits dans les algorithmes d'apprentissage d'audiences, avec un impact de premier ordre pour les sites d'actualité et de second ordre pour les autres sites.

5 Impacts de l'analyse sur les algorithmes de prévisions

5.1 Méthodologie

Les algorithmes utilisés pour prévoir les audiences de chaque tag de publicités des sites internet du groupe TF1 ont été présentés dans un précédent papier *[AJE 2010]*. Le nombre de travaux publiés traitant de prévisions d'audiences (internet ou TV) est limité. Yahoo Research a publié des travaux qui ne détaillent que très peu les algorithmes sous-jacents en se focalisant davantage sur la probabilité de taux de clics *[WANG09]*. Les publications sur les prévisions d'audiences télévisuelles traitent plutôt d'analyse de panels *[CAT94,WEB02]*. Au cœur des serveurs de distribution de publicités, les prévisions sont bien souvent des extrapolations linéaires du passé. Nous avons présenté des algorithmes basés sur des régressions linéaires par moindres carrés multiples que nous ne détaillerons pas ici.

Les algorithmes de prévisions développés par le e-lab utilisent les patterns d'évènements décrits précédemment. Ces algorithmes sont utilisés en production pour le calcul des audiences des sites du groupe TF1, ainsi que pour un autre partenaire du e-lab pour des prévisions de fréquentations et de chiffre d'affaires. Ces derniers travaux plus récents donneront lieu à une autre publication dans les mois qui viennent.

La prévision fait appel aux patterns au cours de deux étapes de calcul :

- Tout d'abord, ils permettent d'effacer l'impact des évènements passés de l'historique des audiences. On obtient ainsi une audience « nettoyée » où le bruit associé à cette catégorie d'évènement a été supprimé et permet de ne pas interférer dans le calcul des coefficients des autres indicatrices (jour de la semaine, mois de l'année, vacances, saison, etc.). En historique (colonne « Audience quotidienne » du ***Tableau 2***), on dispose du nombre de pages vues par jour passé

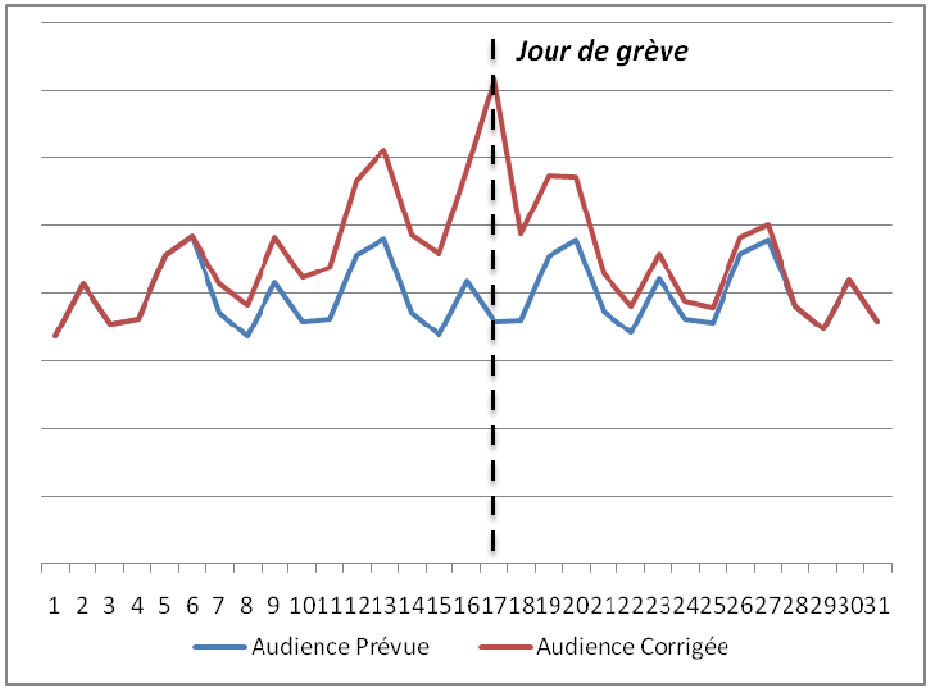
ainsi que d'un calendrier des évènements avec leur date d'apparition. Chacun de ces évènements appartient à une catégorie citée plus haut et dispose donc d'un pattern associé. Pour les évènements de type *pic* qui ont lieu un jour unique sans trainée avant ou après, on fait appel à une indicatrice binaire.

Jour	Audience quotidienne	Est-ce un lundi ?	Est-ce un jour de décembre ?	Est-ce les vacances de noël ?	etc. (environ cinquante indicatrices)	Y-t-il eu un décès de personnalité ?
Jour passé 1	12681	1	1	1	…	0
Jour passé 2	15311	0	1	1		**1**
Jour passé 3	18942	0	1	1		**0.5204**
Jour passé 4	13546	0	1	1		**0.2908**
Jour passé 5	13576	0	1	1	…	**0.1747**
Jour passé 6	13235	0	1	1	…	**0.1122**
Etc.	…	…	…	…	…	…
Jour futur 1	Calculée	0	0	0	…	0
Jour futur 2	Calculée	0	0	0	…	0
Jour futur 3	Calculée	0	0	0	…	0

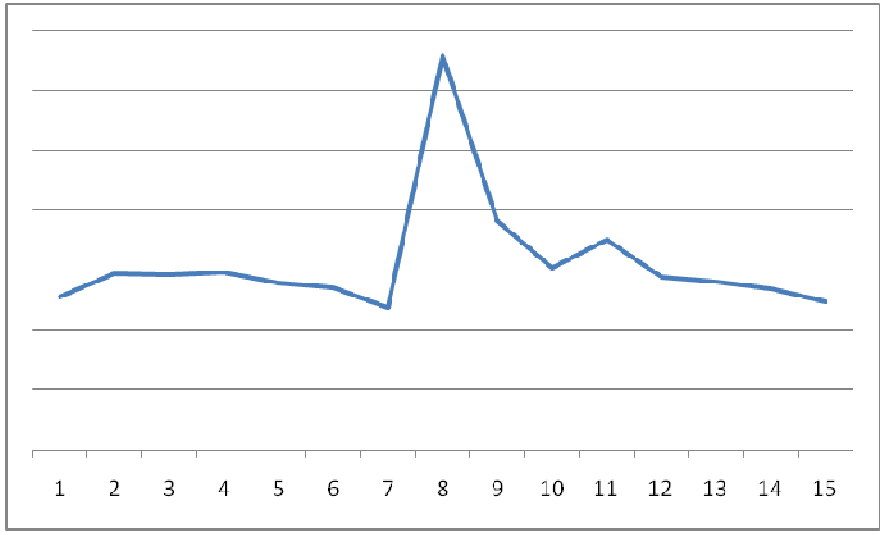
Tableau 2 : *Apprentissage par moindres carrés à l'aide d'indicatrices. On sait qu'il y a eu un décès de personnalité le jour passé 2.*

- Ensuite, nous sommes en mesure d'affecter à chaque famille d'évènement un impact moyen : l'impact est le coefficient de l'indicatrice associée dans l'apprentissage par moindres carrés. Pour tous les évènements futurs qu'il est possible d'anticiper, nous intégrons le pattern couplé à cette puissance apprise pour ce site dans le calcul des prévisions. A titre d'exemple, prenons l'hypothèse qu'une grève a un impact de + 11% sur l'audience de la catégorie d'un site (chiffre calculé à partir des données historiques de cette catégorie en utilisant son coefficient calculé de la colonne « grève »). Si une nouvelle grève est programmée pour une date future, nous sommes en mesure d'intégrer cette information aux prévisions en appliquant le pattern corrigeant l'audience prévue pour ce jour (+11%) ainsi que les audiences des jours précédents et suivants (+11% multiplié par le coefficient réducteur de la journée). Le ***graphique 20*** illustre une prévision d'audience et l'application d'un pattern de grève pour obtenir une prévision d'audience corrigée.

Pour les évènements imprévisibles qui se déclenchent spontanément, il est nécessaire de corriger la prévision dès le premier jour de l'occurrence de l'évènement. Chaque jour, il est donc nécessaire d'intégrer les nouveaux évènements de la journée en cours et de relancer le calcul des prévisions.



Graphique 21 : *Visualisation de l'impact d'un évènement ayant lieu le jour 10 sur une prévision future (avec une trainée avant et après, proportionnelle à la prévision).*



Graphique 22 : *Impact du premier tour des élections 2007 sur l'audience quotidienne de la page d'accueil du site TF1.fr*

5.2 Résultats

L'apprentissage sur ces données historiques « nettoyées » est de meilleure qualité et permet d'obtenir des résultats plus précis, comme le montre le ***Tableau 3*** présentant la différence entre le taux d'erreur pour une prévision sans évènements et une prévision avec, pour 4 benchmarks réel. Ces benchmarks sont des prévisions réalisées sur cinq sites du groupe TF1. Nous ne détaillerons pas les volumétries, ni les taux d'erreurs ici mais uniquement la différence entre les deux taux. La prévision consiste à calculer le nombre de pages vues futures sur une rubrique d'un site internet. On vient ensuite comparer cette prévision au réalisé connu après coup. On mesure 2 prévisions : la prévision à -30 jours et la prévision à 7 jours. L'apprentissage sur données historiques permet d'extraire pour chaque catégorie d'évènements la puissance moyenne, minimale et maximale de cette famille d'évènements pour un site et une période donnés.

Nom du site	Différences entre le taux d'erreur sans et avec prise en compte des évènements à J - 30	Différences entre le taux d'erreur sans et avec prise en compte des évènements à J - 7
www.tf1.fr	- 1,17 %	- 1,46 %
www.plurielles.fr	- 3,12 %	- 2,84 %
www.wat.tv	- 2,34 %	- 2,64 %
www.lci.fr	- 2,72 %	- 2,91 %

Tableau 3 : *Différences entre le taux d'erreurs des prévisions avec et sans la prise en compte des évènements d'actualité pour le mois de Décembre 2009 à partir des données historiques de 1er Janvier à 31 Novembre 2009.*

On constate que l'utilisation des évènements dans les fonctions d'apprentissage permet d'améliorer la qualité des prévisions. Le taux d'erreur est mesuré comme la différence en valeur absolue entre l'audience réelle et l'audience prévue, divisée par l'audience réelle. La colonne 2 du ***Tableau 3*** présente la différence entre le taux d'erreur moyen du mois de Décembre 2009 avec et sans utilisation des évènements dans le moteur des prévisions 30 jours avant le jour J. Par exemple, la prévision du 1er décembre 2009 est calculée le 1er Novembre 2009 avec des données historiques du 1er janvier 2009 au 31 octobre 2009, celle du 2 décembre est calculée de même avec des données historique du 1er Janvier 2009 au 1er Novembre 2009, etc... La colonne 3 présente les mêmes résultats pour la prévision à J-7 : par exemple la prévision du 1er décembre est calculée à partir des données historiques du 1er Janvier 2009 au 23 Novembre 2009.

6 Conclusion

Ce travail est un travail préliminaire qui annonce des recherches futures pour nos équipes dans le domaine de la prévision : l'utilisation de patterns dans les moteurs de prévisions couplée à une gestion calendaire des évènements permet d'aider à anticiper les impacts sur le calcul des prévisions. En cas de réalisation soudaine d'un évènement, il est alors possible d'inclure leur impact et de réajuster les prévisions calculées auparavant. Nous avons vu que cette prise en compte des évènements permet d'améliorer les résultats de prévisions et d'autres résultats sur la prévision dans un autre domaine seront publiés très prochainement.

7 Bibliographie

[PAGE98] Lawrence Page, Sergey Brin. *The Anatomy of a Large-Scale Hypertextual Web Search Engine - Computer Science Department, Stanford University*, 1998.

[LI06] Yu Li, Xiaofeng Meng, Qing Li and Liping Wang. *Hybrid Method for Automated News Content Extraction from the Web - Web Information Systems, WISE 2006*.

[AJE 2010] Jeanjean A., Martin B. (2010). *Prévisions d'audiences et optimisation de plannings de publicités internet*. ROADEF 2010 11 (6). pp. 643-653.

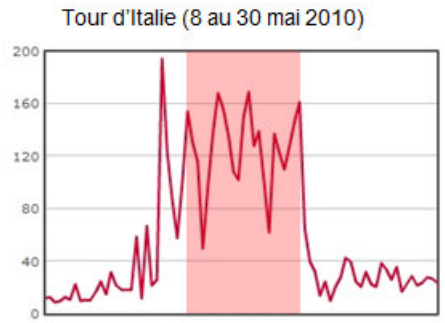
[WANG09] X. Wang, A. Broder, M. Fontoura, V. Josifovski. *A Search-based Method for Forecasting Ad Impression in Contextual Advertising*. 18th International WWW Conference, April 2009.

[CAT94] P. Cattin, R. Festa, A. Le Diberder, *A Model for Forecasting the Audience of TV Programs. Worldwide Electronic and Broadcasting Audience Research Symposium*, – Vol. 1. Pages 513-524.

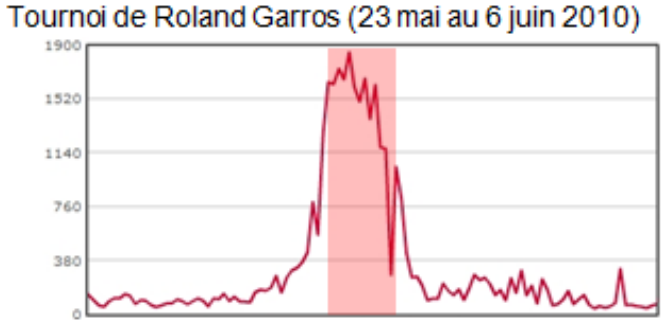
[WEB 02] Weber, René. *Methods to Forecast Television Viewing Patterns for Target Audiences*. Communication Research in Europe and Abroad – Challenges of the First Decade. Berlin: DeGruyter, 2002.

8 Annexes

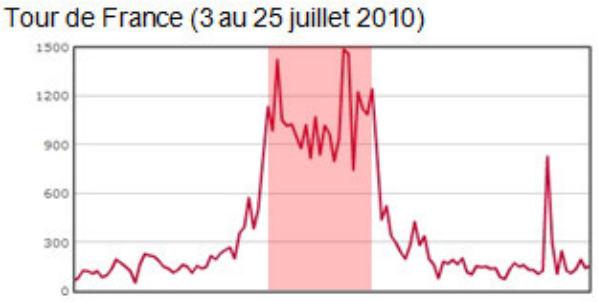
Voici en annexes les courbes d'impacts des trois autres évènements sportifs étudiés ainsi que d'une autre émission de TF1:



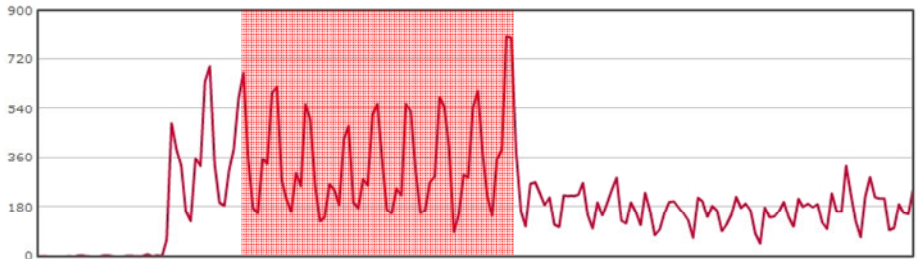
Graphique 23 : *Nombre de publications suite au Tour d'Italie 2010*



Graphique 24 : *Nombre de publications suite au tournoi de Roland Garros 2010*



Graphique 25 : *Nombre de publications suite au Tour de France 2010*



Graphique 26 *: Impact de l'émission L'Île de la tentation Saison 8 diffusée sur TF1 entre le 29 avril et le 18 juin 2010*