Filtrage d'information en contexte de veille

Nassim Moussouni

UFR IDIST – Lille 3

Spécialité Master 2 PRISME

PRoduits de l'Information spécialisée et Médiation Electronique

Mémoire de stage effectué chez Cofidis France

du 1er avril au 28 septembre 2012

Filtrage d'information en contexte de veille

Par Nassim Moussouni

Tuteur de stage : Mr. Stéphane Chaudiron

Responsable de stage : Mr. Jean-Yves Calais

Filtrage d'information en contexte de veille

SOMMAIRE

REMERCIEMENTS

Je tiens par ces quelques lignes à remercier tous ceux qui, de près ou de loin, m'ont aidé à la réussite de mon stage.

- Mr. Stéphane Chaudiron, mon tuteur de stage, qui m'a beaucoup conseillé pendant toute l'année universitaire et m'a aidé à orienter mon mémoire de stage vers ce sujet passionnant. Je le remercie également pour son dévouement pour le Master.
- Jean-Yves Calais, mon responsable de stage, qui m'a fait confiance, et m'a beaucoup conseillé et motivé. J'a appris beaucoup à ses côtés et je le remercie pour sa disponibilité et son écoute.
- Diane et Nathalie, mes deux collègues de bureau, pour leur accueil chaleureux. Elles m'ont beaucoup soutenu et permis une intégration rapide. Grâce à elles, j'ai une meilleure vision des études marketing et des créations publicitaires.
- Tout le service marketing de Cofidis France pour leur disponibilité.
- Et à tous les étudiants et enseignants du Master PRISME de la promo 2011/2012.

Introduction

Avec l'apparition d'Internet et sa généralisation à toutes les couches de la société, la production d'information n'est plus l'œuvre des acteurs classiques. En effet, les supports se sont diversifiés et les auteurs de l'information se sont multipliés. Des journalistes ou des experts dans leurs domaines analysent l'actualité et assurent une présence éditoriale sur Internet à travers leurs propres sites Internet. Cette tendance s'est amplifiée avec l'apparition du web 2.0 et les nouvelles technologies d'édition, de partage et de diffusion de l'information qui permettent à n'importe quel internaute de générer de l'information.

La multiplicité des producteurs de l'information s'est accompagnée par la diversification des supports sous lesquels ces informations sont diffusées. En effet, à coté des sites classiques, sont apparus d'autres formats d'expression tels les blogs, les forums, les sites de partage multimédia et les réseaux sociaux.
Tout cela a contribué à accroître le volume des informations disponibles sur la toile. Ainsi, la boulimie d'informations a atteint de nos jours des sommets. En 2005, le PDG de Google Eric Schmidt a estimé à 5 millions de téraoctets, le volume des informations sur Internet [2] et avec l'apparition des réseaux sociaux, ce chiffre est beaucoup plus élevé aujourd'hui.
Il devient ainsi de plus en plus indispensable pour tous les internautes en général et pour les professionnels de l'information en particulier de consacrer de plus en plus de temps à l'extraction de l'information pertinente. Cela amène donc à installer des systèmes de filtrage d'information efficaces et personnalisés.

Si le web actuel présente une richesse immense en termes de contenu, Il aurait été plus facile de trouver l'information voulue si les données sur le web étaient structurées, mais pour la plupart d'entre elles, ce n'est pas le cas. Elles sont soit non structurées ou au mieux semi-structurées. L'abondance de ces informations

mal structurées rend difficile l'automatisation de la recherche d'information en général, et le filtrage d'information en particulier.

C'est à partir de ce constat que plusieurs recherches ont été menées pour améliorer le filtrage d'information. Différents modes de filtrage sont utilisés de nos jours pour tirer le meilleur. Le plus traditionnel est le filtrage sur contenu qui comme son nom l'indique permet de sélectionner l'information contenue dans les différentes structures du document. Mais avec le web 2.0, est apparu le mode de filtrage collaboratif ou social qui se base les recommandations des internautes devenus depuis cette révolution du web de vrais producteurs d'information.

Cofidis France est une société de crédit à la consommation. L'une des tâches de son *Département Veille et Etudes* est d'assurer une veille constante, exhaustive et efficace. Constante car elle doit être quotidienne et durable dans le temps. Exhaustive car elle doit couvrir toutes les thématiques susceptibles d'impacter directement ou indirectement son environnement (concurrents, produits, réglementations, etc.). Et enfin efficace car elle doit permettre la remontée d'informations pertinentes avec des outils qui répondent aux besoins spécifiques de l'entreprise.

Le stage mené durant ces six mois tente de répondre à ces besoins.
En effet, ma mission consiste à optimiser le dispositif déjà existant et apporter via des outils gratuits ce que l'outil payant *KB Crawl* ne permettait pas. Cette amélioration inclura l'utilisation du filtrage d'information sur le contenu. Ce mode de filtrage se basant sur la sélection ou non d'un document en fonction de son contenu a été adopté en raison du format semi-structuré XML des sources sélectionnées. Il est non seulement plus adapté pour une meilleure efficacité et un ciblage précis du filtre, mais aussi pour la proportion prise par les flux RSS qui sont sous format XLM. Adoptés en premier lieu par les blogs, les flux RSS se sont popularisés et admis dans de nombreux sites d'information en ligne. En parallèle,

plusieurs outils dédiés à leur manipulation sont apparus. Ils permettent d'en créer, de s'y abonner, de les regrouper, de les indexer et de les transmettre.

C'est donc de manière naturelle que l'optimisation du processus de veille quotidienne et automatisée du *Département Veille et Etudes* de Cofidis se base sur ces flux RSS.

Pour y arriver, plusieurs outils on été associés. Ainsi, ce rapport expliquera comment j'ai procédé pour manipuler des flux RSS avec Yahoo Pipes, et comment j'ai intégré la notion de filtrage dans la création de flux RSS pour des sites qui n'en proposent pas avec Feed43, pour enfin les intégrer dans l'agrégateur de flux RSS Google Reader.

Cela demande des compétences dans les langages de structuration web tels le HTML et le XML et quelques notions informatiques telles les expressions régulières et le langage PERL.

II- Présentation de la société COFIDIS France et de son Département Veille et Etudes

I-1- Historique et activités de COFIDIS France

Historique :

COFIDIS est une société de crédit à la consommation appartenant au Groupe Crédit Mutuel-CIC. Créée en 1982, elle est le fruit de deux expertises :

- La vente à distance, métier du Groupe 3 Suisses International (3SI) ;
- Le crédit à la consommation, métier de Cetelem.

Cette alliance a donné naissance à un concept original : la vente à distance de crédits à la consommation.

Présence à l'international :

Dès 1985, COFIDIS a développé son activité à l'international, avec l'ouverture de COFIDIS Belgique. Dans les années 1990, elle étend ses activités en Europe Latine, puis en Europe Centrale dans les années 2000.

Aujourd'hui, COFIDIS est implantée dans huit pays européens : La France, la Belgique, l'Espagne, l'Italie, le Portugal, la République tchèque, la Hongrie et la Slovaquie.

Un actionnaire majoritaire et un autre historique :

Fin 2008, en pleine crise économique et financière, 3SI et la Banque Fédérative du Crédit Mutuel (BFCM) concluent un contrat de cession. En mars 2009, la BFCM devient alors actionnaire majoritaire du Groupe COFIDIS Participations auquel appartiennent COFIDIS, CREATIS et MONABANQ. 3SI reste son actionnaire historique.

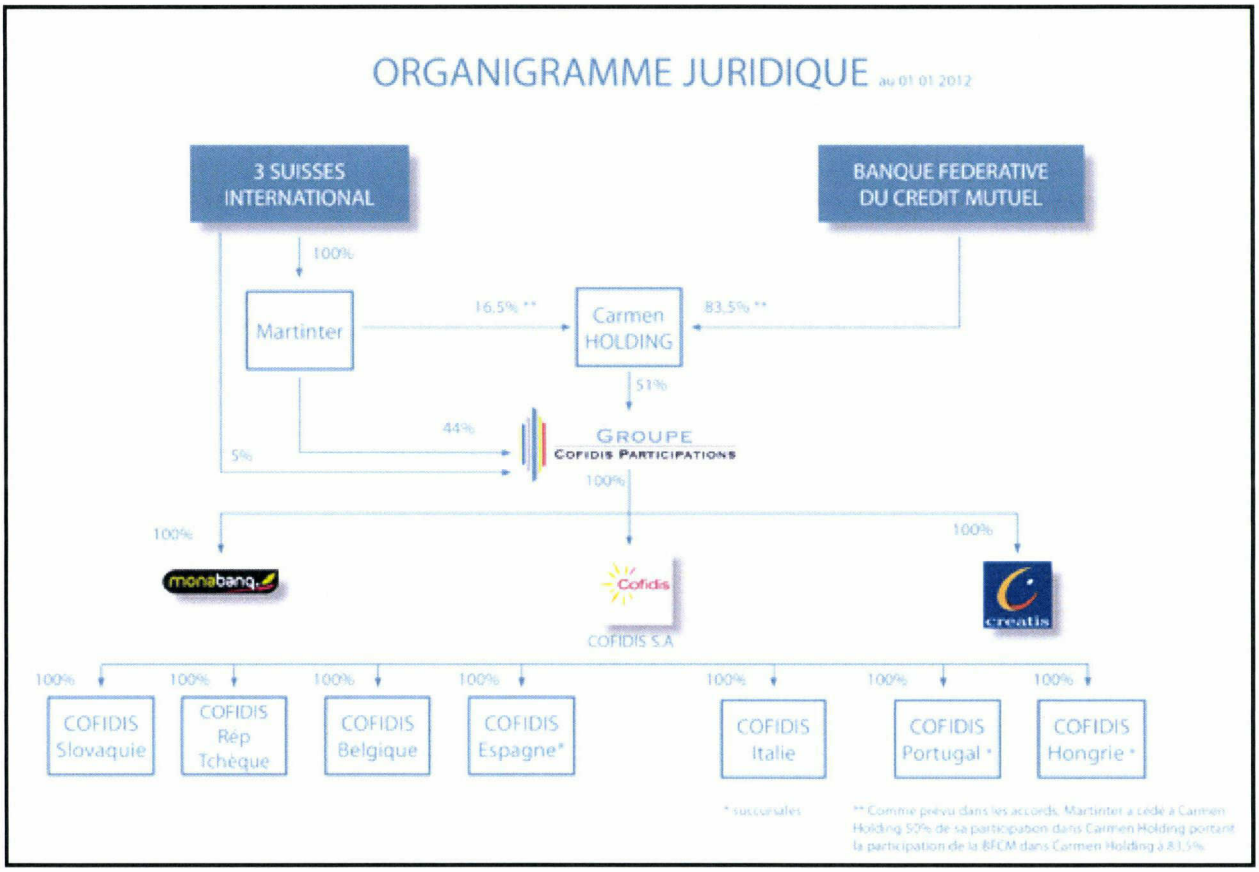


Figure1 : organigramme juridique du Groupe COFIDIS Participations (Source : cofidis.com)

Investissement dans le sponsoring cycliste :

Pour développer la notoriété de la marque, COFIDIS s'est lancé dans le sponsoring cycliste et ce depuis 1996. Cela lui permettant de promouvoir les valeurs de la marque, telles la proximité, la solidarité et l'esprit d'équipe. Ces dernières considérées comme essentielles pour le spécialiste de la relation à distance.

Chiffres clés et métiers de COFIDIS :

Basée à la haute Borne à Villeneuve d'Ascq, COFIDIS emploie plus de 1600 employés. Elle compte près de 3,5 millions de clients et 850 enseignes partenaires qui proposent les solutions de financement de COFIDIS au travers de ses trois produits : 1 euro.com, le « 3 fois carte bancaire » et la carte 4 étoiles.

Grâce à sa large gamme de produits, COFIDIS a satisfait plus de 7 millions de clients, et ce grâce à ses canaux de distribution multiples : la vente à distance, Internet et les magasins.

L'offre de crédit de COFIDIS englobe ainsi :

- ***le crédit renouvelable*** : réserve de crédit permanente dans laquelle le titulaire du compte peut puiser dans la limite de son disponible, ce type de crédit est décliné en 3 produits selon le montant accordé :
  - ***Accessio*** : accordé pour des montants de financement compris entre 500 et 3 000 € ;
  - ***Modulcio*** : accordé pour des montants de financement compris entre 3 500 et 6 000 € ;
  - ***Alto*** : accordé pour des montants de financement compris entre 6 500 et 10 000 €.

  Le crédit renouvelable est aussi associé à des carte d'enseignes où le titulaire peut faire des achats chez les partenaires de COFIDIS. On distingue ainsi :
  - ***La carte 4 étoiles*** : elle permet de faire bénéficier son titulaire de privilèges exclusifs et de traitements de faveur tout au long de l'année dans chaque enseigne partenaire. Le client peut payer comptant ou à crédit.
  - ***1euro.com*** : c'est un crédit renouvelable permettant à ses souscripteurs de financer leurs achats en ligne au comptant ou en plusieurs fois dans plus de 800 sites marchands. Le coût décliné tourne en général autours d'un euro.

- ***Le prêt amortissable*** **ou personnel :** le prêt sur mesure (PSM) de COFIDIS est un prêt personnel qui ne peut être utilisé qu'une seule fois. Il reste fixe et est déterminé lors de la souscription.
- **Le rachat de crédit :** il permet de regrouper plusieurs prêts d'un client en un seul décliné en un prêt personnel pouvant aller de 3 000 à 80 000 €, déterminant ainsi une mensualité adaptée à la situation financière du client.

- **L'assurance :** COFIDIS propose la garantie assurance par l'intermédiaire des ACM (Assurance Crédit Mutuel) pour une prise en charge des mensualités du client en cas de chômage, de maladie, d'invalidité et de décès.
- **La téléphonie mobile :** véritable diversification stratégique, COFIDIS a lancé ***Cofidis Mobile*** au mois de Novembre 2011. Elle propose une gamme de forfaits téléphoniques. COFIDIS s'appuie pour cela sur le savoir-faire d'une autre filiale du Crédit Mutuel, NRJ Mobile.

I-2- Marché du crédit en France

Le marché du crédit en général et du crédit à la consommation en particulier connait depuis quelques années des moments difficiles. En effet, la crise économique et financière de 2008 a fortement impacté le secteur. Ainsi, cette crise rend les ménages plus prudents sur leurs dépenses et passent de moins en moins par des crédits à la consommation. La loi Lagarde entrée en vigueur au mois de mai 2011, a accentué cette situation délicate. En effet, le durcissement de la réglementation en matière d'octroi du crédit a eu des effets délétères sur le crédit renouvelable et sur la production globale du crédit à la consommation. L'environnement concurrentiel a connu lui aussi des mutations avec le lancement de l'offre de crédits à la consommation par la Banque Postale. Ces mutations, cumulées aux prochaines règles prudentielles de Bâle3 et CRD4, et le spectre d'une éventuelle suppression du crédit renouvelable par le nouveau gouvernement risquent de brider encore plus les professionnels du crédit.

I-3- Le Département Veille et Etudes

Si l'activité de veille parait essentielle dans n'importe quelle situation et dans n'importe quelle structure, elle l'est encore davantage dans une grande entreprise telle que COFIDIS. Ainsi, pour répondre aux besoins internes, il est essentiel de mettre en place une structure de veille complète et adaptée à l'organisation interne de l'entreprise.

Vue la situation difficile délicate que traverse le secteur du crédit à la consommation, il est primordial de mettre une surveillance optimale de l'environnement financier, concurrentiel, réglementaire, et normatif de COFIDIS.

I-3-1- Organisation interne

Le Département Veille et Etudes (DVE) est un service rattaché au pôle marketing lui-même rattaché à la direction commerciale. Il fonctionne selon un système transversal car pouvant couvrir toutes les directions de l'entreprise. Il est composé d'un chef spécialisé dans la veille assisté par deux autres collaborateurs.

I-3-2- Activités du département et moyens mis en œuvre

a- Activités

La veille et les études sont les deux missions dont est chargé le DVE.

Le DVE est chargé d'assurer différents types de veilles, allant de la veille financière à la veille concurrentielle, en passant par la veille commerciale, partenariat, technologique ou marketing.

Ses missions consistent à surveiller l'environnement du marché du crédit à la consommation ou tout autre environnement impactant directement ou indirectement son secteur d'activité. Il fournit ainsi des alertes, des livrables récurrents et des topographies marché pour le comité directeur, ainsi que des fiches comparatives des produits.

Il réalise également des études Ad' hoc, ainsi que des cartographies sur la concurrence. Il assure enfin la gestion d'études marketing : étude de marché, usages et altitudes, pré et post tests de communication, tests de nouveaux concepts produits, en qualitatif, quantitatif ou sémiologiques.

b- Moyens mis en œuvre

Pour bien mener ses missions, le DVE a à sa disposition, différents outils de veille.

On peut distinguer deux sortes de moyens : les payants et les gratuits.

b-1- Outils de veille payants : ils permettent non seulement un gain de temps mais aussi une automatisation de la veille.

- **Outil de veille concurrentielle :** propriété d'un cabinet de veille, cet outil en mode SaaS, permet de recevoir des alertes mails sur la boite électronique du DVE. Il assure une veille tarifaire et concurrentielle dans la limite du secteur du crédit à la consommation.
- **KB Crawl :** en mode SaaS, cet outil de veille est un aspirateur de sites et est conçu pour détecter tous les changements qui surviennent sur les sites Internet. Accessible via un portail dédié et personnalisé, il peut être accessible pour plusieurs utilisateurs à la fois. Il peut être paramétré à tout moment via ces différents modules :

➢ *Sourcing et paramétrage* : à tout moment, l'utilisateur peut intégrer de nouvelles sources d'informations pour étendre et approfondir son périmètre de veille. il existe une possibilité de cibler la ou les zones à surveiller via le module *Scraper ;*

➢ *Crawl :* ce module permet à l'outil de rechercher et d'indexer toutes les informations de la source balayée et de les valider dans sa base de données. C'est à partir de cet historique que l'outil permet la détection des changements ;

➢ *Planificateur :* KB Crawl permet via ce module l'automatisation de la surveillance à fréquence déterminable, de toutes les sources intégrées ;

➢ *Déclencheur d'alertes* : c'est à travers ce module que l'utilisateur définit les paramètres permettant le déclenchement des alertes mails. Ces paramètres peuvent être la détection de requêtes de mots clés, les changements survenus dans des documents ou l'apparition ou la disparition de pages web.

➢ *Organisation :* il est possible d'organiser les sources dans des dossiers et sous-dossiers ;

➢ *Publier :* ce module permet de déterminer le ou les destinataires des alertes mails

- **Outil de veille publicitaire :** toujours en mode SaaS, cet outil permet d'effectuer une veille marketing en prospection sur les investissements publicitaires des différents acteurs du crédit à la consommation, sur tous les supports : Médias TV, radio, presse, affichage, bannières internet, courriers ;

Cela permet d'être alerté des investissements et des dernières créations publicitaires et des stratégies de communication. A partir de ces données, peuvent être constitués des livrables accompagnés d'analyses.

- **Outil de veille panéliste** : il est dédié au scan de mailings, e-mailings reçus par des panélistes et mis à disposition sur l'outil en ligne. Il permet de répondre à des besoins d'études ad' hoc et de faire le livrable mensuel et de connaître la stratégie de fidélisation des acteurs du crédit à la consommation.
- **Abonnements presse :** ils peuvent être reçus électroniquement ou sur papier. Ce sont des périodiques qui traitent des secteurs impactant directement ou indirectement les activités de COFIDIS.

b-2- Outils de veille gratuits

- **Google Reader** : cet agrégateur de flux permet de recevoir les alertes Google sous forme de flux RSS.
- **Blog veille** : accessible en intranet seulement, il permet de diffuser les dernières actualités à l'ensemble des collaborateurs.

I-3-3- Ressources d'information

La provenance des informations peut être externe ou interne à COFIDIS. En effet, certaines informations circulent entre les différentes directions. Le DVE tire ainsi certaines dépêches de la part d'autres services.

Les sources externes sont essentiellement des revues de presse, des articles de presse en ligne, des études provenant de cabinets ou d'instituts et des alertes mails des différents prestataires externes.

II- Etat de l'art du filtrage d'information

II-1- Principe du filtrage d'information

Le filtrage d'information (FI) consiste à concevoir des mécanismes destinés à faire parvenir à l'utilisateur l'information qui l'intéresse directement.

Un système de filtrage d'informations (SFI) peut être assimilé à un assistant personnel et doit être capable d'identifier les documents qui correspondent ou pas aux besoins en information des utilisateurs **[4]**. Il peut être défini comme un processus qui permet d'extraire à partir d'un flot d'informations (News, e-mail, actualités journalières, etc.) celles qui sont susceptibles d'intéresser un utilisateur ou un groupe d'utilisateurs ayant des besoins en information relativement stables. » **[3]**

Les systèmes de filtrage d'information sont conçus pour les données non structurées ou semi-structurées, par opposition à l'application aux bases de données qui elles sont des données structurées. La notion de structuration concerne à la fois le format du document et son contenu.

Ils sont portés sur de l'information textuelle constituée de données non structurées incluant des données multimédia telles que audio, vidéo, images qui d'ailleurs sont difficiles à gérer par les moteurs de recherche.

Le filtrage concerne un flux d'information en provenance d'une ou plusieurs sources extérieures (*ex. news)* ou adressé directement par d'autres sources (*ex. email).*

Le filtrage doit prendre en compte le profil de l'usager qui spécifie au système ses caractéristiques. Il inclue également la suppression de données à partir d'un flux entrant, plutôt que de rechercher des données dans le flux.

La distinction entre le filtrage et les processus connexes que sont la récupération, le routage, la catégorisation et l'indexation, n'est pas toujours évidente.

Nous verrons plus loi, un comparatif entre la recherche d'information et le filtrage d'information.

II-2- Types d'informations existantes

Données structurées : bases de données : sont des informations disposées de façon à être traitées automatiquement sans une intervention humaine.

Données semi-structurées : e-mails (en-tête bien défini, corps du texte non structuré).

Données non structurées : sont des données textuelles, images, vidéos, audio.

II-3- Evolution du filtrage d'information

Avec le développement du *WordWideWeb* ainsi que d'autres réseaux d'information, la recherche dans le domaine du filtrage d'information s'est accélérée.

La DSI, Diffusion Sélective de l'Information est l'une des premières formes de filtrage de l'information numérique **[1]**. Elle consiste à envoyer continuellement l'ensemble des données nouvelles faisant référence à une requête personnalisée préalablement enregistrée. Il s'agit d'une veille documentaire axée sur un thème précis destinée à un utilisateur ou un groupe d'utilisateurs aux intérêts relativement communs appelés *Profils* [3]. Elle reste néanmoins difficile à implémenter sur certains systèmes tels les messageries étant donné la masse d'information sur Internet. Ces dernières ont fait l'objet de recherches aboutissant à un filtrage sous forme de mail-boxes introduites par Denning **[4]**. En effet, cette deuxième approche laisse à l'utilisateur la possibilité de fixer des règles de filtrage appliquées à ses messageries électroniques, ces dernières étant plus ou moins structurées sous forme de champs tels les champs de renseignement sur l'expéditeur, l'objet, la date, etc. **[5]**. Par ce procédé, Denning, qui a inventé le terme *« filtrage d'information » (information filtering)* a voulu élargir le concept de génération d'information en incluant sa réception. Ce procédé faisant appel à la coopération des usagers, a été repris et appliqué pour les articles de presse et ceux provenant d'Internet **[4]**. Malone en 1987 définit trois modèles pour le filtrage d'information : le cognitif, l'économique et le social appelé récemment collaboratif. S'ensuit après plusieurs projets qui ont pout but d'améliorer l'extraction et le filtrage d'information.

Parmi les plus remarquables, on peut citer le projet *Tapestry* lancé par *United States Advanced Research Projects Agency des Etats-Unis (DARPA)* qui a sponsorisé *Message Understanding*

Conference (MUC) afin de développer des techniques d'extraction d'information pour la sélection de messages couplées aux techniques statistiques de présélection pouvant être soumis à des traitements automatiques du langage naturel (TAL) encore plus sophistiqué.
Mais c'est le projet international *TREC (Text REtrieval Conference)* sponsorisé conjointement en 1992 par DARPA et *National Institut of Standards and Technology (NIST)* qui s'intéresse à l'évaluation des systèmes de recherche et de filtrage d'information sur une collection standard de textes **[11]**.

II-4- Les modes de filtrage d'information

D'une manière générale, il existe trois grandes familles de FI :

- *Le filtrage collaboratif :* initialement appelé *filtrage social* **[6]** ce mode se base sur les appréciations des utilisateurs pour la sélection des documents. Il se base sur l'hypothèse que les personnes à la recherche d'information devraient pouvoir se servir de ce que d'autres ont déjà trouvé et évalué **[7].** Ses limites sont essentiellement la rareté des annotations
- *Le filtrage sur contenu :* également appelé *langage cognitif*, est un autre mode de filtrage dont le principe est que la décision de sélection ou non d'un document se base exclusivement sur le contenu de ce même document. Il peut être considéré comme de la recherche d'information vue que la fonction de relation entre une requête et un corpus de documents joue le rôle de filtre entre le profil qui est la requête appliquée sur les documents entrants [8].

  Etant donné que les données sont de type textuel, il est possible de structurer l'information sous un format convenable de type fichier XML, de sorte que le filtre s'applique sur une profondeur modulable selon les besoins. En effet, et nous le verrons plus loin qu'il est possible d'extraire l'information à partir d'une partie ou de la totalité des documents entrants. A cela s'ajoutera une étape de normalisation du texte en le débarrassant des caractères spéciaux ou des balises HTML ainsi que l'introduction de morceaux de textes tels la source du document ou les dates de publication.
- *Le filtrage hybride :* ce procédé combine les approches des deux autres familles **[9]**. En procédant ainsi, le but est de tirer les avantages et de réduire les inconvénients qui leurs

sont liés. En effet, ces systèmes gèrent des profils axés sur le contenu, et la juxtaposition de ces profils forme des communautés permettant ainsi le filtrage collaboratif.

Si Burke en a identifié sept différents types de filtrage hybride, toutes se basent sur trois approches principales **[10]** :

- La première considère le profil comme des groupes d'utilisateurs considérés comme similaires pour un thème donné.
- La deuxième approche considère que les profils thématiques des utilisateurs sont utilisés pour les comparer dans la phase de calcul des corrélations entre utilisateurs du filtrage collaboratif.
- Le « boosting », qui consiste à utiliser les résultats de différentes techniques et à les combiner.

II-5- Filtrage d'information et recherche d'information

Le filtrage d'information est une branche de la recherche d'information. Leurs similitudes sont telles, que les distinguer devient difficile. En effet, les deux concepts sont en général les moyens les plus utilisés pour accéder à l'information sur le web. La méthode *Pull* et la méthode *Push*.

Selon Belkin, la recherche d'information a pour fonction d'amener à l'utilisateur les documents qui vont lui permettre de satisfaire son besoin en information » **[1]**. Ce dernier fait la démarche d'aller sur le moteur de recherche, formule ses besoins en information via des requêtes et rapatrie les informations qu'il juge pertinentes. C'est une procédure individuelle car l'œuvre d'un seul utilisateur et répondant à un besoin à court terme. Les informations recherchées sont établies dans des bases de données statiques incluant de nouvelles et surtout beaucoup d'anciennes informations du fait de la rareté de leurs mises à jour. C'est la méthode *Pull*.

Du côté opposé, le filtrage d'information « achemine des documents qui se présentent vers des groupes de personnes, en se basant sur leurs profils à long terme élaborés à partir de données d'apprentissage » **[1].** Il ne nécessite donc pas de formulation du besoin par le ou les utilisateurs. D'où l'intérêt tiré de cette approche, qui permet un gain de temps, d'efforts et évite ainsi un travail itératif lassant et ennuyeux. Appliqué généralement sur des flux de

données, le filtrage d'information permet ainsi de n'acheminer que de nouvelles informations. Souvent utilisé par un groupe d'utilisateurs aux intérêts communs, le filtrage d'information est une approche pérenne dans le temps. Elle s'inscrit sur du long terme. C'est la méthode *Push*.

Caractéristiques	Recherche d'information	Filtrage d'information
Approche générale	Collection de l'information en la cherchant	Déplacement de l'information reçue
Utilisateur	Un seul utilisateur	Un groupe d'utilisateurs
Profils	Besoins ponctuels sur du court terme	Besoins récurrents sur du long terme
Origine de l'information	Base de données statique	Flux de données dynamique
L'information	Mises à jour rares	Réception d'informations nouvelles
Approche sur le contenu	Requêtes de mots clés	Filtrage sur le contenu via des mots clés

Tableau 1 : Tableau comparatif : recherche d'information **vs** filtrage d'information **[1]**

II-6- Filtrage d'information et veille

a- Les fonctions de la veille

Définitions :

Plusieurs définitions ont été proposées pour définir au mieux ce qu'est la veille. Je retiens celle d'un organisme officiel (AFNOR) qui définit la veille selon la norme expérimentale française **[14]** dans « *prestation de veille et prestation de mise en place d'un système de veille* », comme étant : « l'activité continue et en grande partie itérative visant à une surveillance active de l'environnement technologique, commercial...etc., pour en anticiper les évolutions », et l'anticipation comme la « détection d'une situation avant qu'elle se soit réellement manifesté » **[14].** La veille est donc « un Système d'Information » ouvert sur l'extérieur ayant pour objet l'écoute de l'environnement de l'entreprise pour capter et anticiper les grandes tendances à venir, et ainsi de conforter le processus de décision interne. **[15]**

Typologies et objectifs de la veille

➨ **Veille stratégique :** la plus connue. Elle consiste à mettre en place un processus de veille faisant appel à l'ensemble des types de veille pratiqués. Elle inclut des domaines de plus en plus nombreux comme la veille image (ou veille E-réputation). La veille stratégique consiste à

collecter puis analyser les informations dans le but de permettre aux personnes concernées de prendre les meilleures décisions possibles. Elle ne sera optimale que si elle implique de surveiller tous types d'informations (web, terrain, mais aussi d'intégrer l'information interne à l'entreprise). **[16]**

➨ **Veille technologique :** c'est l'activité qui met en œuvre des techniques d'acquisition, de stockage et d'analyse d'informations, concernant un produit ou un procédé, sur l'état de l'art et l'évolution de son environnement scientifique, technique, industriel ou commercial, afin de collecter, organiser, puis analyser et diffuser les informations pertinentes qui vont permettre d'anticiper les évolutions, et qui vont faciliter l'innovation. **[17]**

➨ **Veille concurrentielle:** consiste à surveiller les actions et l'environnement des concurrents afin d'anticiper au mieux ses stratégies. Elle impliquera de surveiller et d'analyser notamment l'organisation des concurrents, leur politique commerciale, marketing, financière, leurs produits et leur stratégie de distribution. La **veille concurrentielle** doit aussi s'attacher à surveiller l'environnement du concurrent : partenaires, fournisseurs, influences, lobbyings et ses marchés afin d'imaginer des scenarii de positionnement et d'évolution. **[18]**

➨ **Veille réglementaire et juridique :** La veille réglementaire est un dispositif de surveillance des dispositions réglementaires pouvant affecter l'activité commerciale et marketing de l'entreprise. Elle vise à identifier le plus tôt possible les projets de réglementation de manière à pouvoir les anticiper ou parfois même les influencer (lobbying). Particulièrement importante dans les domaines où les règlements ou normes sont très présents, elle permet d'adapter les produits, services et pratiques commerciales dans les meilleurs délais, ou de permettre de profiter de nouvelles opportunités commerciales (ex : obligation de détecteur de fumée). **[19]**

➨ **Veille Image :** C'est la recherche, le traitement et la diffusion (en vue de leur exploitation) de renseignements relatifs à l'image, la notoriété de l'entreprise ou d'une marque, et ce sur tous types de supports (Internet, informations terrain etc.). L'e-réputation, veille notoriété et veille d'opinion font partie de la veille image.

b- Le filtrage d'information dans le contexte de la veille

Le processus de veille est schématisé le plus souvent par un cycle qui comporte les quatre grandes étapes de tout système de veille.

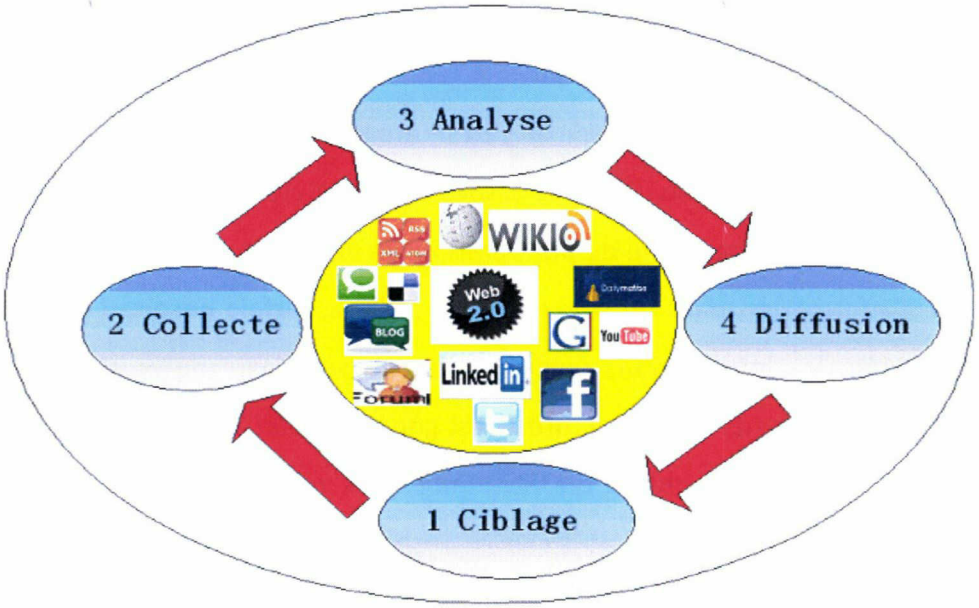


Figure 2 : Cycle classique d'un système de veille

Entre la partie de ciblage et de paramétrage de l'information et celle consistant à la collecter, il est primordial de mettre en place un système de filtrage efficace, sous peine de voir un flux d'information entrant volumineux. Ce qui peut engendrer un enfouissement de données importantes au milieu de données non pertinentes.

Le plus souvent, dans un système de veille, les personnes chargées de le mettre en place sont confrontées à l'infobésité du web. La multiplicité des supports d'information est une autre problématique à laquelle il fallait trouver des solutions. Ainsi, le filtrage d'information prend toute son ampleur dans le métier du veilleur. Ce dernier, doit non seulement posséder des compétences techniques pour paramétrer des agents intelligents permettant le filtrage d'information, mais éventuellement les améliorer ou les coupler à d'autres outils.

Cependant, dans un contexte de veille, le filtrage d'information est souvent basé sur le modèle de recherche d'information. Ainsi, les informations et les profils sont représentés par des listes de mots pondérés. **[20]**

De ce fait, on peut distinguer les flux d'information résultant de l'association des deux modèles. Cela est rendu possible grâce à des outils permettant d'effectuer la requête composée de mots clés et d'en ressortir un flux d'information correspondant à cette requête. L'exemple qui illustre le mieux cette approche sont les moteurs de recherche proposant un flux RSS pour chaque requête. On passe ainsi de documents au langage web type HTML à des documents XML caractéristiques des flux RSS.

L'autre approche consiste à remonter un flux d'information résultant de l'application de filtres (mots clés) sur le contenu. Elle n'est pas associée au modèle de recherche d'information même si elle emploie la même approche de pondération via des mots clés.

Le filtrage d'information dans un processus de veille a parfois besoin d'être appliqué sur des items précis. Or, les informations électroniques ne sont pas toutes structurées. De ce fait, l'utilisation de documents XML structurés s'est généralisée et est devenue incontournable pour les veilleurs. Cela facilite le ciblage du filtre qui peut être appliqué sur différents items tels les liens (qui contiennent parfois les mots clés correspondant au profil), les titres ou les descriptifs. Cela diminue le risque que le flux entrant contienne des informations non pertinentes ne correspondant pas au profil de référence.

III- Intégration du processus de filtrage d'information pour l'optimisation du dispositif de veille du Département Veille et Etudes de COFIDIS France

III-1- Problématique, analyse et évaluation de l'existant

Comme précisé ci-dessus, la veille concurrentielle dans le contexte du crédit à la consommation est très importante. En effet, dans ce secteur, les acteurs du crédit procèdent souvent à des changements de taux, au lancement de promotions et à de nouvelles créations publicitaires. De ce fait, il est indispensable de disposer d'outils pouvant traquer ces changements qui souvent sont minimes mais permettent d'avoir une vision de leur stratégies commerciales à court ou à long terme et d'anticiper les menaces et les opportunités.

A coté de cela, le cadre réglementaire et normatif du secteur du crédit à la consommation ne cesse de changer et est soumis à des encadrements de plus en plus réguliers. D'où la nécessité de disposer de sources d'informations qui permettent un suivi quotidien.

Face à ces deux constats, le Département Veille et Etudes dispose d'outils dédiés.

Si ceux assurant la surveillance des changements de taux et de création de publicitaires donnent entière satisfaction, les autres dispositifs tournés beaucoup plus sur une veille marché provenant de sources d'information Internet, affichent des lacunes au niveau de l'exhaustivité, de la pertinence et du ciblage d'information, et de l'organisation de ces mêmes informations.

Exhaustivité **:** Google Reader qui est un agrégateur de flux RSS tire les remontées des requêtes de Google Alerte. Cependant, ces requêtes ne sont pas mises à jour et ne couvrent pas tous les noms des produits, services, marques et filiales des acteurs du secteur d'activité. Cela conduit au risque de passer à coté d'informations qui peuvent être stratégiques.

Se satisfaire des alertes Google, c'est aussi rester dépendant du traitement de l'information jugée pertinente par l'algorithme de Google. Il n'ya pas en effet, une maîtrise parfaire des sources d'information, qui présentent certes des flux RSS selon les thèmes choisis, mais aucun filtrage par mots-clés n'est possible. C'est pour cette raison que le dispositif actuel se contente des alertes fournies par Google.

Pertinence : KB Crawl, qui assure la surveillance des changements de sites, peut être paramétré de sorte qu'il assure la détection de changement de contenus, l'apparition de mots clés préalablement définis et la détection de la disparition ou d'apparition de nouvelles pages. Pour mieux cibler les zones à surveiller, le module *Scraper* peut être exploité et permettre cela. Par contre, ce dernier ne peut s'appliquer que sur une page, et cela pose des difficultés quant il s'agit de surveiller un site entier ou une partie de ce site. En effet, si l'utilisateur donne une profondeur de surveillance, le *Scraper* ne s'appliquera pas au reste des pages du site. A moins d'appliquer *Scraper* à toutes les pages (ce qui n'est pas envisageable pour la plupart des sites car disposant de plusieurs pages), il n'existe pas de solutions qui permettent un ciblage à l'échelle du site. Cela est du en grande partie à la non structuration de l'information traitée. Sans le *Scraper,* les informations remontées proviendront non seulement du corps de la page (titre de la page et son contenu), mais le plus souvent d'autres rubriques du site, telles les anciens articles, les menus, l'en-tête et le pied de page. Cela induit beaucoup d'informations non pertinentes et une perte de temps due au traitement d'énormes quantités d'informations.

Organisation : dans Google Reader, chaque thème surveillé est composé de trois flux RSS. L'un remonte les informations d'actualité, le deuxième découle des blogs et enfin le troisième flux provient des alertes vidéo. Cela encombre l'outil notamment quand le nombre de thèmes tend à augmenter. Google Alerte permet de sélectionner le type de support sur lequel se trouve l'information : actualité, blogs, vidéos, discussions (forums) et livres. Il permet certes de rassembler ces flux en un seul, mais ne permet pas d'en sélectionner quelques uns. C'est soit un seul flux pour chaque support ou la totalité.

III-2- Les attentes du Département Veille et Etudes

Les attentes du Département Veille et Etudes s'appuient sur les problématiques su-citées. Elles sont explicitées lors de l'entretien et recadrées au fur et à mesure du déroulement des tâches. L'optimisation du dispositif de veille est l'attente principale. Elle doit pouvoir apporter des solutions concrètes permettant de faciliter une veille quotidienne et automatisée.

III-3- Définition de la mission et des tâches à effectuer

De ces attentes ont découlé différentes tâches.

Après un état des lieux, il fallait mettre en place des outils facilitant la lecture des sources tout en permettant leur convergence vers un seul outil pour les remontées, les lectures, le traitement et le stockage. Il fallait aussi pouvoir constituer un pack de sources conséquent et être capable de filtrer l'information entrante.

Même s'il a fait l'objet de paramétrages, KB Crawl devait être optimisé et ainsi maximiser son exploitation.

En parallèles à ces missions, je devais mener un travail sur un projet émanant de demandes spécifiques d'une direction. Ce projet incluant une surveillance du web pour mener une veille technologique et en ressortir les dernières innovations.

J'étais chargé également de la rédaction de livrables veille, d'alertes et d'études Ad' Hoc.

III-4- Réalisation de la mission

Après analyse de l'existant, il était clair qu'une optimisation était nécessaire. Comme tout dispositif de veille, il était primordial que je m'imprègne dans le secteur du crédit à la consommation. Cela passait par un benchmark des besoins.

Ainsi pour l'exhaustivité de la veille sur toutes les entités (nom des produits, services, marques et filiales), j'ai recensé ces dernières à partir d'une base de données en ligne.

J'ai élaboré et ce en concertation permanente avec mes collègues du DVE, la liste de toutes les sources susceptibles d'être intégrées dans le nouveau dispositif.

Pour ne pas chambouler les pratiques déjà existantes, nous avons décidé du maintien de l'agrégateur de flux RSS *Google Reader.* De plus, il fournit l'essentiel des services nécessaires à la pratique de la veille.

III-4-1- Identification des outils et des moyens ainsi que leurs mises en œuvre

III-4-1-1- Le langage XML et les flux RSS

Le nouveau dispositif que j'ai mis en œuvre devait satisfaire tous les points déjà exprimés. C'est le résultat de l'association de plusieurs outils gratuits. Chacun devait résoudre une problématique particulière.

La solution que j'ai apportée se base entièrement sur le filtrage d'information sur le contenu. Les outils ainsi sélectionnés avaient comme matière grise des documents semi-structurés, en l'occurrence les fichiers au langage HTML (langage des sites web) et surtout XML qui est le langage de structuration des flux RSS.

XML (*Extensible Markup Language*) ou « langage de balisage extensible », est un langage de balisage qui s'est imposé progressivement comme le support privilégié pour l'échange des données et leur stockage **[21].** En effet, face aux documents en ligne souvent mal structurés, il devenait complexe de cibler la recherche d'information en général et son indexation ainsi que son filtrage d'une manière spécifique.

Un document XML doit comporter une structuration bien spécifique :

- **Prologue** : Il contient deux déclarations certes facultatives mais conseillées :

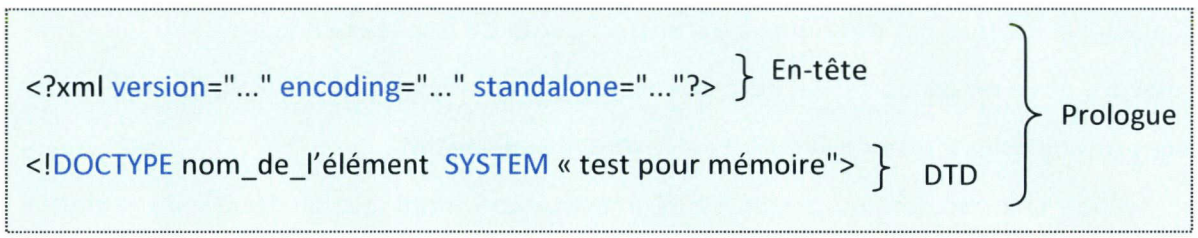


Figure 3 : structure générale d'un document XML

- **En-tête** : située au tout début du document, elle peut contenir trois attributs, *version*, *encoding* et *standalone* **[22].** « *version* » précise la version XML utilisée (1.0 et 1.1). « *encoding* », précise le codage des caractères utilisé (US-ASCII, ISO-8859-1, UTF-8, et UTF-16 sont les plus utilisés et les plus rencontrés). Enfin, « *standalone* » indique si le fichier est relié ou pas à des déclarations externes. Sa valeur par défaut étant « no », il est possible de lui attribuer les valeurs « yes » ou « no».

- **La DTD (Document Type Definition)** : son rôle est de définir la structure du document. Il spécifie quels éléments peuvent apparaître dans le contenu d'un élément donné ainsi leur ordre.**[22]**

- **Corps du document :**

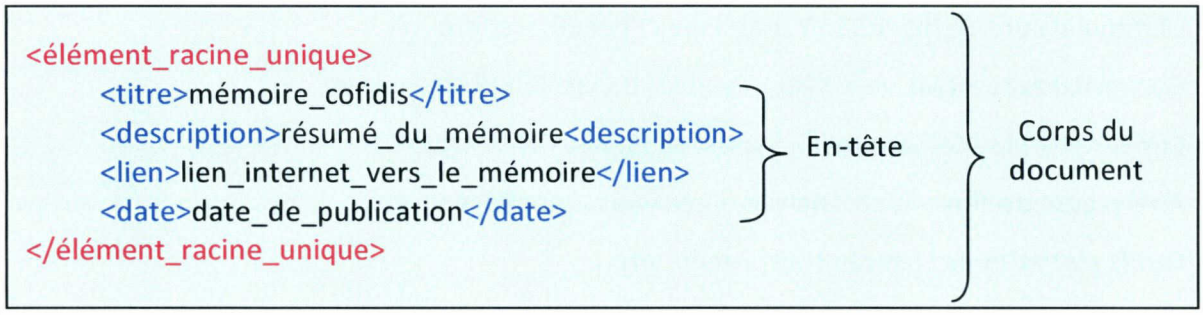


Figure 4 : structure du corps d'un document XML

Le corps du document XML est une arborescence d'éléments (composés chacun d'une balise ouvrante et d'une autre fermante) imbriqués, avec un élément racine unique. Les attributs peuvent être associés aux éléments pour apporter plus d'informations.
Lors de l'indexation et du filtrage d'information sur les documents XML, il est possible via certains outils de déterminer sur quels éléments de structuration les appliquer et pouvoir manipuler les flux RSS. Nous verrons plus loin comment procéder.

Les flux RSS : créé en 1999 par Netscape **[23],** le flux RSS, appelé également ***fil RSS*** est un format d'échange de données. Adopté en premier lieu par les blogs, Il s'est généralisé et a conquis les sites d'actualité. Cette popularisation s'explique par la simplicité de structuration des informations contenues et la possibilité d'accéder au contenu sans passer par le site d'origine. Ajouté à cela le mode *push de l'information* que permet le flux RSS. En effet, l'utilisateur n'est plus obligé d'aller tirer lui même l'information. C'est cette dernière qui est poussée vers l'utilisateur. Cela a un intérêt double, le gain de temps et la centralisation des informations recueillies.

Naturellement, plusieurs outils ont vu naissance depuis l'apparition des flux RSS. Nous pouvons les classer en plusieurs catégories **[24]**:

- **Lecteurs sous différents systèmes d'exploitation** : RSS Owl; Feed Reader, etc.
- **Lecteurs en ligne (agrégateurs de flux RSS)** : Google Reader; Bloglines, Netvibes, etc.
- **Pages personnalisables** : iGoogle, Gritwire, etc.
- **Manipulateurs de flux RSS** : Yahoo Pipes ; FeedRinse, etc.
- **Convertisseurs HTML vers XML** : Feed43, PagetoRSS, FeedFire, etc.
- **Convertisseurs RSS vers mail** : SendMeRSS, RSS FWD, etc.
- **Analyseurs de flux** : Feed Analysis, FeedHaus, ReadBurner, etc.
- **Outils statistiques** : Feedburner, Feediz, etc.

Cette liste n'est bien évidemment pas exhaustive. Le but étant de montrer l'ampleur prise par la technologie de syndication de contenu.

III-4-1-2- Les outils

Les outils présentés ont été identifiés et choisis en fonction des besoins exprimés. Leur validation s'est faite après des tests concluants qui ont apporté les solutions aux problématiques citées plus haut.

a- Yahoo Pipe

Présentation générale :

Yahoo Pipes est une application web créée par Yahoo et qui permet de manipuler les flux RSS. En version béta depuis le 7 février 2007, c'est un éditeur très puissant et complet qui permet une multitude d'utilisations grâce à ses différents modules.

De prime abord, Yahoo Pipes n'est pas très intuitif et est complexe à utiliser. Mais pour un utilisateur avisé, cet outil est un vrai dispositif de veille. Il offre en plus une interface graphique permettant de manipuler les flux RSS agrégés avec des Pipes (tuyaux) qui sont des canaux de communication entre les différents modules **[25]**. Ainsi il est possible de faire de la

programmation mais avec un minimum de connaissances en celle-ci et aux langages HTML, XML et les expressions régulières, Yahoo Pipes utilisant le langage Perl pour cela **[26]**.

Présentation des applications :

Page d'accueil :

Pour pouvoir utiliser Yahoo Pipes, il faut posséder un compte Yahoo, Gmail ou Facebook pour pouvoir stocker les flux créés. Il est accessible via cette adresse http://pipes.yahoo.com/pipes/.

La page d'accueil permet via la rubrique « *My Pipes* » de visualiser les différents Pipes créés. La rubrique « Pipes » affiche la totalité des flux, alors que la rubrique « favorites » affiche les flux mis en favoris.

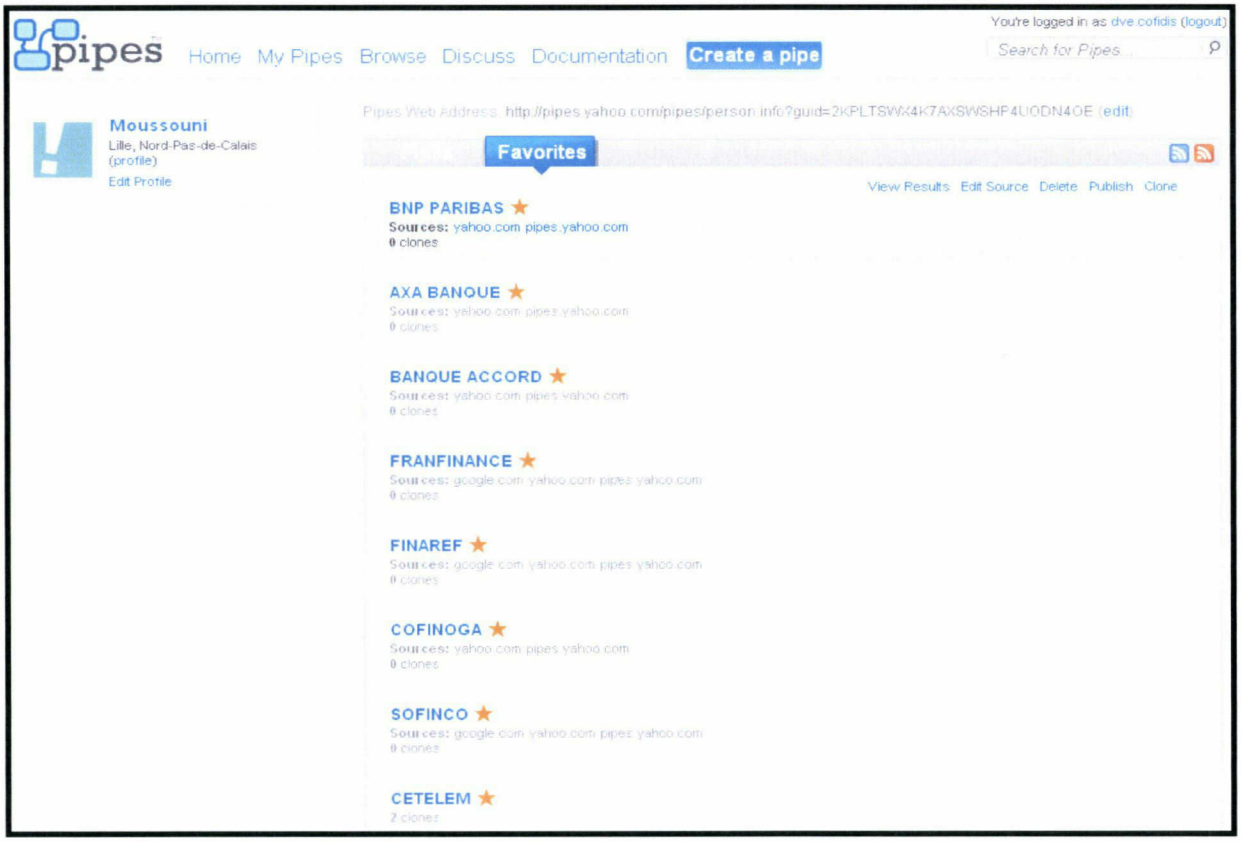


Figure 5 : page d'accueil de Yahoo Pipes

Il est possible de rejoindre la communauté « Yahoo Pipes » dans la rubrique « *Discuss* », qui permet d'échanger des astuces et de trouver des réponses aux différentes interrogations.

Dans la rubrique « *Browse* », sont répertoriés tous les Pipes rendus publics par leurs auteurs. Un moteur de recherche est dédié pour cibler les flux souhaités. Cela permet de trouver des Pipes intéressants pour pouvoir les utiliser ou les optimiser.

Enfin la rubrique d'aide « *Documentation* », permet de mieux utiliser l'outil. Elle classe tous les tutoriels officiels (en anglais) pour tous les modules présents dans Yahoo Pipes.

Modules de Yahoo Pipes

Yahoo Pipes compte sur la partie gauche de chaque Pipe, des rubriques, qui elles mêmes sont constituées de modules. Ils sont tous répertoriés dans ce tableau.

RUBRIQUES	MODULES		
Sources	Find First Site Feed	YQL	Fetch Data
	Yahoo Local	Item Builder	RSS Item Builder
	XPath Fetch Page	Flicker	Feed Auto-Discover
	Fech Feed	Fetch CSV	
User inputs	Private Text Input	Number Input	
	Date Input	Text Input	
	URL Input	Location Input	
Operators	Count	Truncate	Location Extractor
	Rename	Split	Reverse
	Create RSS	Filter	Unique
	Regex	Sort	Web Service
	Tail	Union	
	Sub-Element	Loop	
Url	URL Builder		
String	Yahoo Shortcuts	String Tokenizer	String Regex
	String Replace	Sub String	Private String
	String Builder	Term Extractor	
Date	Date Builder	Date Formatter	
Location	Location Builder		
Number	Simple Math		
Favorites	Tous les Pipes mis en favroris		
My Pipes	Tous les Pipes créés et enregistrés		
Deprecated	Fetch Page		

Tableau 2 : Liste des modules disponibles dans Yahoo Pipes

Caractéristiques des principaux modules de Yahoo Pipes

- ***Fetch Feed*** : permet d'intégrer un ou plusieurs flux RSS ;
- ***Feed Auto Discovery :*** avec ce module, il est possible de détecter et récupérer tous les flux RSS d'un site web. Nous verrons comment extraire tous les articles de ces flux RSS ;
- ***Module Filter*** : permet de filtrer (permettre ou bloquer) les remontées des flux RSS par mots clés se trouvant dans le titre, la description ou le lien ;
- ***Module Unique*** : permet de limiter les liens renvoyant vers le même titre ou le même lien (supprime les doublons) ;
- ***Module Sort*** : permet de classer les remontées selon un certains nombre de critères. Utile pour classer selon la date de publication ;
- ***Module Split*** : permet de dupliquer le flux RSS ;
- ***Module Regex*** : il permet de changer la configuration des différents items composant le flux RSS. Nous pouvons par exemple ajouter, supprimer ou modifier le titre, la description, la date ou tout item détecté par Yahoo Pipes.

  Il permet en effet de supprimer certaines parties d'un tweet et qui ne permettent pas d'avoir un tweet propre pour qu'il soit utilisé pour le ***Module Filter***.
- ***Module Union*** : il permet de relier les différents modules.
- ***Module Loop*** : il permet d'intégrer d'autres modules issus de la famille « String ».

  Ce qui permet par exemple de remplacer un item par un autre, ou pour hiérarchiser la structure du flux RSS.
- ***String Builder :*** il permet de structurer les items et d'assigner les résultats à tel ou tel item.
- ***String Regex*** : avec ou sans des expressions régulières, il est possible de remplacer un item par un autre et de le personnaliser.
- ***Fetch Page*** : les flux RSS sont tronqués, c'est-à-dire qu'ils contiennent un résumé de l'article. Pour pouvoir afficher tout l'article, il est possible d'appliquer une extension du descriptif grâce à ce module. Des notions du langage HTML sont nécessaires.

b- Google Alerte

Google Alerte est un service de notification qui permet de recevoir les dernières actualités en rapport avec la requête saisie en amont. Les informations ainsi remontées peuvent être reçues sous forme d'alertes mail. Pour ma part, j'ai préféré les recevoir sous forme de flux RSS que je récupère et intègre à Yahoo Pipes.

L'intérêt de Google Alerte est le fait qu'il se base sur l'immense pouvoir d'indexation de Google qui de plus, remonte des informations pertinentes selon son propre algorithme.

Ainsi, pour créer une alerte, il faut spécifier à l'outil les termes recherchés, le type de contenu (actualité, blogs, vidéos, forums, etc.) et la fréquence d'envoi. L'intérêt de Google Alerte est qu'il est possible de limiter le bruit au maximum en optimisant les requêtes avec des opérateurs booléens.

c- Feed43

Présentation générale :

Feed43 est outil qui permet de générer des flux RSS de n'importe quel site qui n'en propose pas. Il est certes puissant mais un peu complexe à utiliser. En effet, il demande quelques connaissances en langage HTML. Cette complexité dépend de la page à analyser et des éléments à extraire.

Ainsi, Feed43 peut être très utile lorsque la surveillance d'un site ne peut se faire via des flux RSS non proposés par le site. C'est le cas notamment de sites de concurrents et de certains sites d'actualité.

Son intérêt se trouve aussi lorsque l'on veut surveiller les résultats d'une recherche. En effet, lors d'une requête, chaque résultat présente une URL spécifique. Grâce à cette dernière, nous pouvons suivre les nouveaux résultats en créant un flux RSS grâce à Feed43. C'est du filtrage sur URL, qui dans KB Crawl se traduit par le module « macro » qui d'ailleurs n'est pas très performant.

d- Google Reader

Google Reader est un agrégateur de flux RSS. C'est outil gratuit qui permet de lire un ou plusieurs fils RSS de manière rapide et de les classer dans différents dossiers. Il offre la possibilité d'archiver les articles jugés pertinents dans une liste de favoris appelée « liste de suivi ».

Même s'il existe d'autres agrégateurs de flux RSS avec des tableaux de bord plus dynamiques et des fonctionnalités plus avancées, Google Reader satisfait largement aux activités du Département Veille et Etudes qui l'utilise depuis des années. C'est pour cette raison et pour ne pas chambouler les habitudes des collaborateurs qui l'utilisent, qu'il a été décidé de le garder pour la suite des activités de veille.

Fonctionnalités de Google Reader :

L'outil regorge de plusieurs fonctionnalités pour optimiser les remontées, la lecture, l'archivage, et la diffusion de l'information.

De plus, il existe des extensions pour les navigateurs Internet spécialement conçues pour Google Reader permettant une présentation plus souple et plus ergonomique.

- ***Ajout de nouveaux flux RSS*** ;
- ***Liste de suivi*** : répertorie les remontées mises en favoris ;
- ***Tendances*** : cette rubrique dispose de toute une série de statistiques d'utilisation : nombre d'abonnements, nombre d'éléments publiés par jour ou le pourcentage d'éléments non lus ;
- ***Parcourir*** : permet de s'abonner à des thèmes prédéfinis et à des recommandations ;
- ***Diffusion*** : il est possible de partager des articles en les diffusant via e-mail aux destinataires ;
- ***Tags*** : pour chaque remontée, il est possible de lui assigner des tags pour leur identification ultérieure ;
- ***Recommandations*** : permet de recommander une information via Google+, le réseau social de Google.

III-4-2- Identification des sources à intégrer

Pour constituer un pack de sources conséquent et pertinent, j'ai opté pour les méthodes de sourcing classiques. Elles doivent toutefois respecter le périmètre de surveillance.

Il était possible de se contenter des alertes Google qui permettent de recevoir l'information pertinente. Mais cette pertinence dépend de l'algorithme de Google. Cela pourrait amener à ne pas recevoir des informations sensées être pertinentes au niveau de l'entreprise.

Pour éviter de tels risques, j'ai répertorié les principales sources traitant des thèmes dont opère l'entreprise. Elles sont regroupées comme suit :

- Quotidiens nationaux ;
- Quotidiens régionaux ;
- Médias en ligne (médias parallèles, médias collaboratifs) ;
- Médias spécialisés dans le secteur financier ;
- Sites institutionnels et gouvernementaux ;
- Sites d'organismes bancaires, de crédit et d'études ;
- Sites des concurrents ;
- Bases de données en ligne.

Pour la plupart de ces sources, il existe des formes d'abonnements sous format RSS. Cela correspond à la stratégie de recueil d'information que j'ai choisie et est adaptée aux outils sélectionnés.

Pour ceux qui n'en proposent pas, j'ai créé des flux RSS via l'outil Feed43.

A rappeler que le sourcing s'est fait tout au long de la mise en œuvre du nouveau dispositif.

III-4-3- Contraintes et solutions apportées

Les premières contraintes qui apparaissent sont l'absence de flux RSS pour une source donnée. Il se peut aussi qu'une source permettant de s'abonner à un thème via des flux RSS, ne le permette pas pour une autre rubrique susceptible d'être intéressante. Enfin, certaines informations contenant un mot clé ne sont pas remontées par le flux habituel, mais par des recherches via le moteur de recherche du site.

Devant ces contraintes, il fallait trouver une solution permettant comme même de remonter ces informations.

Comme le dispositif à mettre en place se base sur les flux RSS, il fallait donc être en mesure d'en en créer. C'est ce que permet l'outil Feed43. Ce dernier est un outil puissant qui convertit les pages web en HTML (pages sans flux RSS) à des pages au format XML (langage des flux RSS). Il fait partie de la famille des outils de conversion HTML vers XML.

Le principe repose sur l'extraction d'items via des balises régulièrement répétables au langage HTML. Cela permet un filtrage précis et personnalisé de l'information voulue. L'outil se charge ensuite d'afficher les résultats de l'extraction. L'illustration qui suit, expliquera également comment remonter automatiquement les résultats d'une recherche.

1- Renseigner l'URL de la page web

Nous voulons par exemple créer un flux RSS pour la page résultant de la recherche effectuée sur le champ de recherche du journal Les Echos.

Après avoir effectué la requête "crédit à la consommation", nous obtenons des résultats sur une page dont l'URL est :

http://recherche.lesechos.fr/?exec=1&texte=%22cr%C3%A9dit+%C3%A0+la+consommation%22&ok=

Après avoir cliqué sur « creat a new feed », on atterrit sur la page suivante.

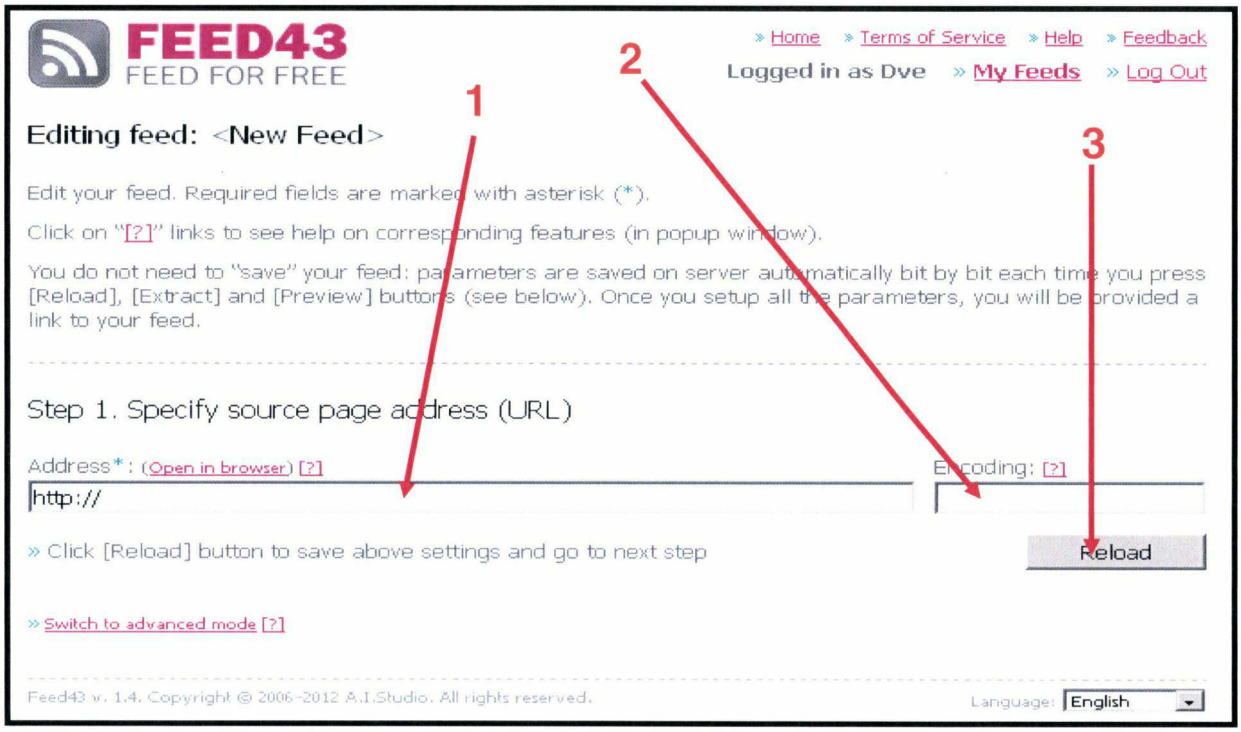


Figure 6 : renseignement de l'url d'un site web pour en créer un flux RSS

1- La première étape est de copier cette url et de la coller dans le champ « Address »

2- le champ « Encoding » permet de spécifier le langage du code source de la page (le plus souvent c'est UTF-8). Il est le plus souvent automatiquement généré par feed43, donc inutile de le remplir.

3- En cliquant sur le bouton « Reload », Feed43 génère 3 champs.

2- Définir les règles d'extraction

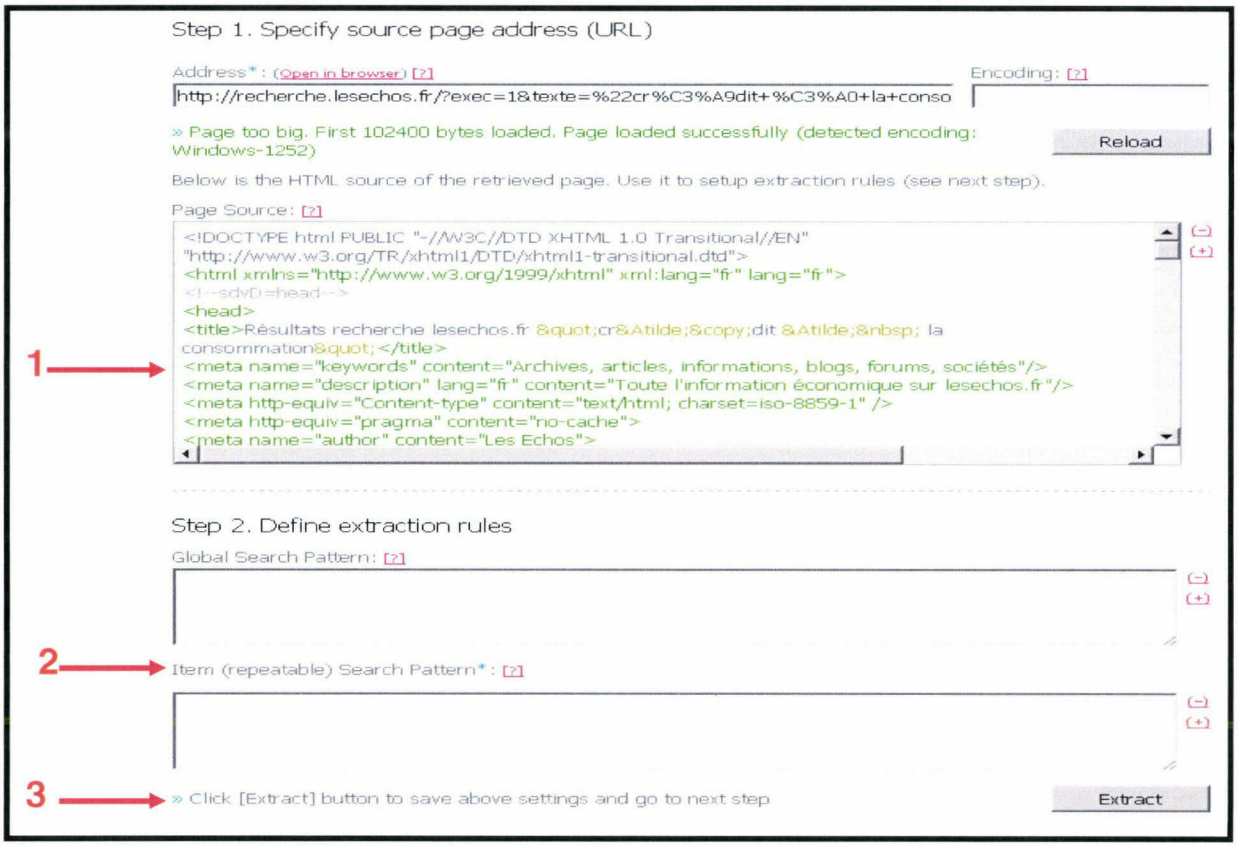


Figure 7 : définition des règles d'extraction

1- Le premier champ est pré-rempli par l'outil et correspond au code source de la page intégrée ;

2- Le deuxième champ correspond au caractère permettant l'extraction des données qui nous intéressent. Il s'agira toujours du caractère : **{%}.**

3- Le troisième champ est celui qui fait intervenir les connaissances en langage HTML.

Pour chaque résultat et pour chaque page web, il y a toujours des balises qui se répètent. Le but est de les identifier. Pour cela il faut toujours comparer entre les codes

sources des différents résultats. Prenons les codes sources correspondants aux premier et deuxième résultats de la requête "crédit à la consommation"

Code source du premier résultat	Code source du deuxième résultat
 Paris accélère à la hausse <img src="http://www.lesechos.fr/acces/88/investir_00459088.gif" class="acces_gif" border="0"> constitué par l'orientation du <SPAN class="keywordsHighlight">crédit</SPAN> à la <SPAN class="keywordsHighlight">consommation</SPAN> (attendue en baisse à 11 milliards de dollars, contre 17,12 milliards précédemment) à 21 heures Article du 07/08/2012	 Le Cac 40 attendu sur une note hésitante <img src="http://www.lesechos.fr/acces/68/investir_00459068.gif" class="acces_gif" border="0"> constitué par l'orientation du <SPAN class="keywordsHighlight">crédit</SPAN> à la <SPAN class="keywordsHighlight">consommation</SPAN> (attendue en baisse à 11 milliards de dollars, contre 17,12 milliards précédemment) à 21 heures Article du 07/08/2012

Tableau 3 : codes sources de deux résultats issus d'une recherche dans un moteur de recherche

On souhaite extraire le titre, la description, le lien vers l'article et enfin la date de publication.

Nous remarquons qu'il existe des balises répétitives pour chaque résultat :

<h3 style="font-family:Georgia,serif;">**saut de ligne**
<a href="**lien vers l'article**">**le titre**<img src="**une image**</h3>**saut de page**
 description **autres renseignements non utiles**
date de publication

Figure 8 : les balises répétitives identifiées

Les informations à extraire seront remplacées par la caractère **{%}.**

Dans notre cas, les informations à extraire sont : le lien vers l'article, le titre, la description et la date de publication. Les informations à ignorer seront remplacées par le caractère **{*} :** les sauts de ligne et les autres renseignements inutiles.

Nous aurons ce langage d'extraction qui sera intégré au champ 3 :

```
<h3 style="font-family:Georgia,serif;">{*}<a href="{%}">{%}<img src="{*}</h3>{*}
<div class="RESUME">{%}</div>{*}<font color="#CC0000">{%}</font>
```

Figure 9 : le langage d'extraction intégré dans Feed43

3- Extraction des items

En cliquant sur « Extract », nous obtenons les résultats suivants :

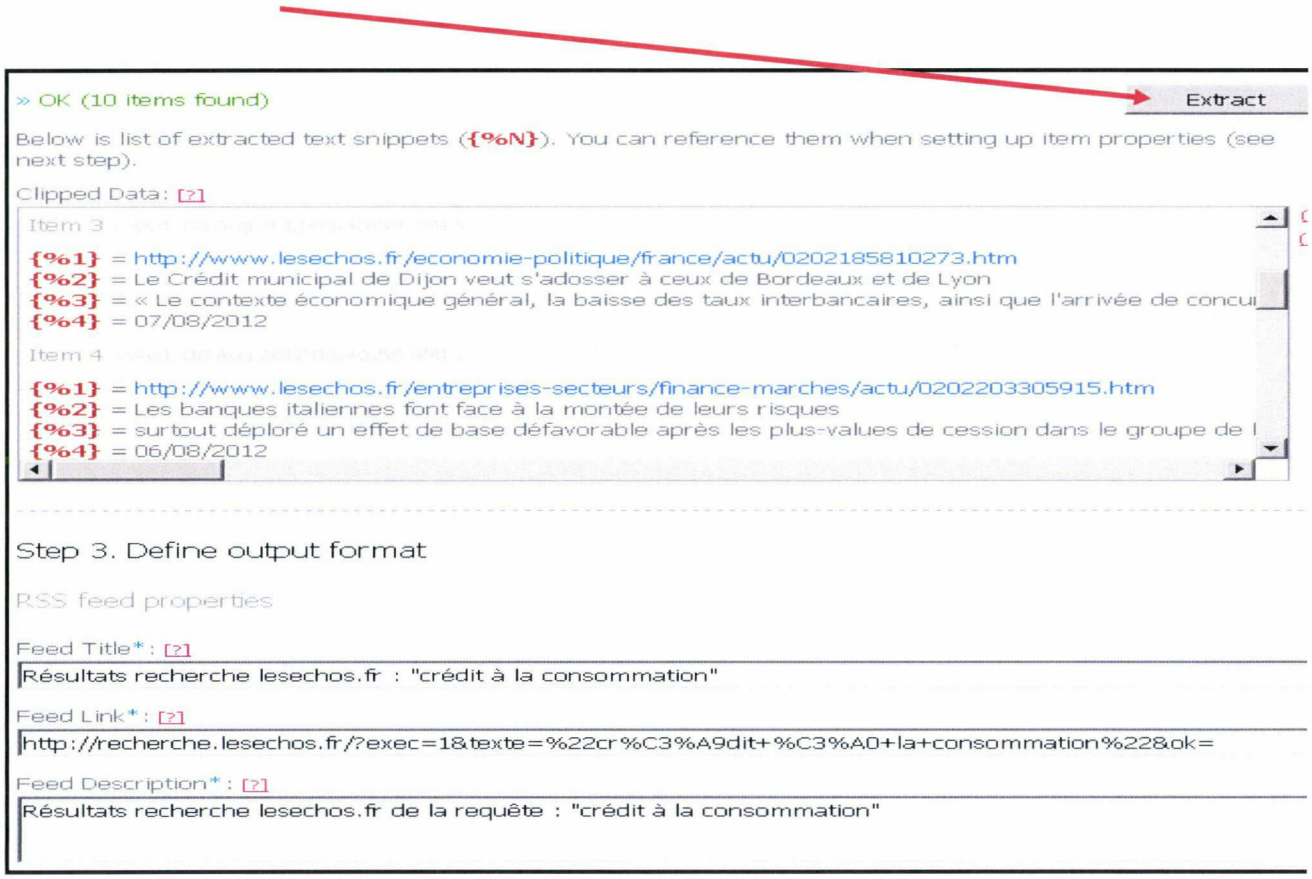


Figure 10 : extraction des items selon le langage incorporé

Le champ « Clipped Data » nous donne un aperçu des résultats obtenus.

Chaque résultat est précédé d'un caractère {%numero de l'item}. Ce sont ces caractères qu'on va utiliser pour faire correspondre chaque item à « titre », « description » ou « lien ».

Les trois autres champs « Feed Title », Feed Link » et « Feed Description » sont générés automatiquement pour décrire le flux RSS généré. Ils peuvent être modifiés selon les besoins.

4- Définition des règles de sortie

Cette étape permet d'indiquer de quelle façon les items seront structurés dans le fichier xml du flux RSS.

Pour cela, il faut remplir les trois derniers champs. On utilisera pour cela les caractères précédant chaque résultat.

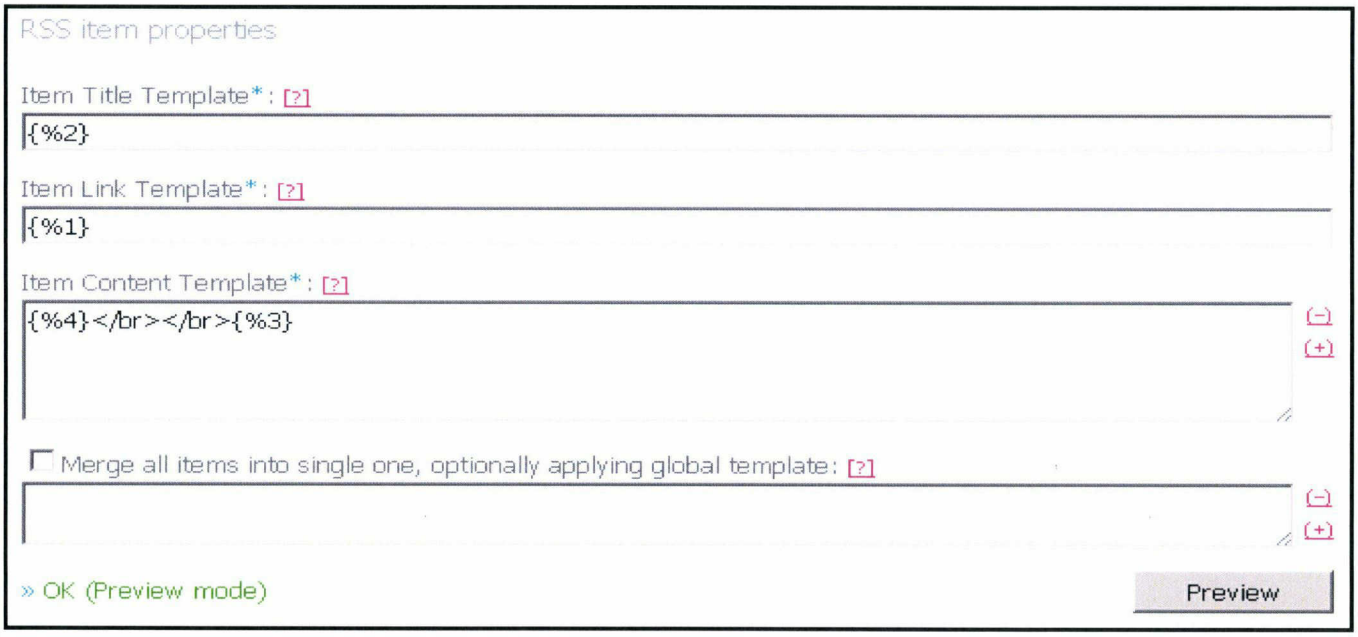


Figure 11 : structuration des items dans le document XML

Dans le champ « Item Title Template », on indiquera quel caractère correspond au titre dans les résultats de recherche. Dans notre cas il s'agit du caractère **{%2}** ;

Dans le champ « Item Link Template », on indiquera quel caractère correspond au lien vers l'article. Dans notre cas il s'agit du caractère **{%1}** ;

Dans le champ « Item Content Template », on indiquera la date de publication **{%4}** suivie de deux sauts de page **</br></br>** suivis de la description de l'article **{%3}**.

Le dernier champ, qui est optionnel, permet d'englober tous les items en un seul. Il n'est pas intéressant dans le contexte de notre veille.

En cliquant sur « Preview », on obtient enfin l'aperçu de notre flux RSS généré.

Figure 12 : flux RSS créé à partir d'une page qui n'en contient pas

Tout en bas, se trouve l'adresse du flux RSS permettant de l'intégrer dans un agrégateur.

Figure 13 : service d'édition du flux RSS proposés par Feed43

Il est constitué d'une suite de chiffres, mais il est possible de le modifier pour qu'il soit plus parlant et mieux indéniable, en cliquant sur « Change file name ».

***NB** : le nom ne doit contenir que des lettres, des chiffres et les caractères (-) et (_).*

Une fois le flux créé, il est possible de l'intégrer dans un outil d'agrégation.

L'autre problématique est celle liée à la non exhaustivité des alertes Google. En effet, les requêtes renseignées ne couvrent pas tout le périmètre dont opère l'entreprise. Pour y remédier, j'ai répertorié tous les produits et services, les noms de marques et filiales des acteurs du secteur d'activité. Cela grâce à la base de données actualisée fournie par l'un des prestataires externes. J'ai également utilisé les opérateurs booléens qui permettent la combinaison de plusieurs recherches d'une manière simultanée et précise. Ils affinent les résultats et permettent un filtrage sur le contenu. Ce dernier peut s'opérer sur le corps de l'article, sur le titre ou sur l'URL.

Google propose toute une palette d'opérateurs booléens. Ils sont résumés dans ce tableau **[27]** :

Opérateur booléen	Rôle et type de recherche
AND (+)	Recherche combinée : combine plusieurs recherches simultanément
OR (|)	Recherche aléatoire : recherche d'une ou plusieurs recherches
NOT (-)	Recherche restrictive : élimine les mots ou expressions non désirables
" "	Permet une recherche d'expressions exactes
?	Remplace un seul caractère
Troncature *	Remplace une chaine de caractères.
allintitle (intitle)	Recherche dans le titre de la page
allinurl (inurl)	Recherche dans l'adresse url de la page
Filetype	Permet de spécifier sur quel type de fichier rechercher (PDF, Word, etc.)
Site:	Recherche sur un site donné

Tableau 4 : liste des opérateurs booléens les plus utilisés

Cette étape a ainsi permis de mettre à jour les remontées mais également réduit le bruit (informations non pertinentes).

Si le dispositif actuel se contentait des alertes Google, c'est parce qu'aucun filtrage par mots-clés n'est possible sur les flux RSS. D'où le choix de s'équiper de KB Crawl qui permet la détection de mots clés. En effet, cet outil remonte tous les articles contenant le ou les mots clés renseignés lors de la phase de paramétrage. S'il remplit bien sa tâche, nous avons détecté quelques problématiques liées à son crawl.

Une veille efficace présuppose des informations nouvelles et un nombre de remontées non pertinentes réduit au maximum. KB Crawl ne permet pas toujours tout cela. Même s'il dispose du module *Scraper*, qui permet de délimiter la zone à surveiller, cela pose problème dès lors qu'on applique une profondeur de surveillance, car *Scraper* ne s'applique qu'à une seule page et pas à tout le site. L'appliquer à toutes les pages du site réglerait le problème, mais avec le nombre de pages que contient chaque site, cela devient vite impossible à réaliser. Ce qui fait que même s'il détecte un ancien article, il le remontera encore une fois. Ce dernier pouvant se trouver dans des rubriques d'anciens articles que les journaux en lignes proposent de lire en supplément car similaires à l'article principal, ou des menus, l'en-tête et le pied de page. La non structuration régulière des informations ne fait qu'empirer les remontées aléatoires de KB Crawl.

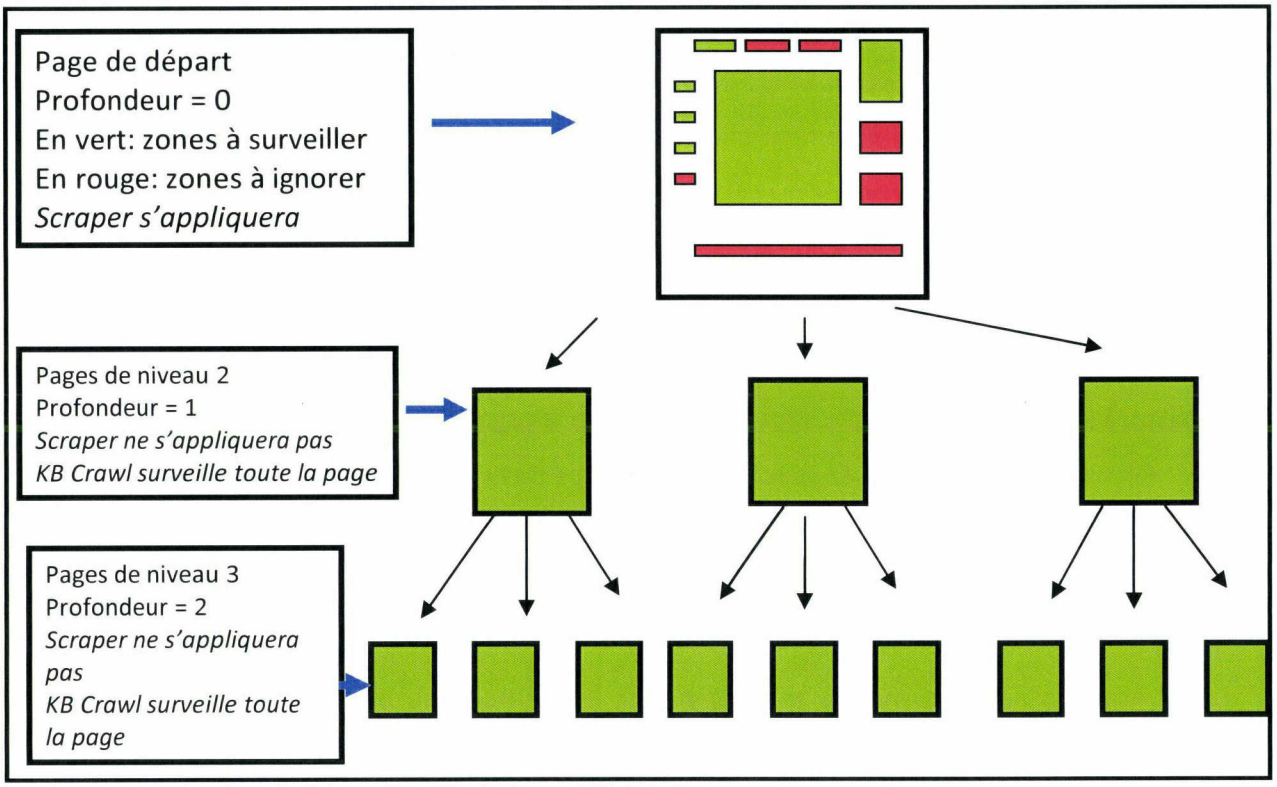


Figure 14 : Modèle de fonctionnement du module *Scraper* de KB Crawl

Ces remontées aléatoires font que KB Crawl génère beaucoup d'alertes, le plus souvent non pertinentes car s'agissant pour la plupart d'entre elles d'informations anciennes. Le traitement de ces alertes prend beaucoup de temps aux personnes chargées de cette mission.

Pour y remédier, dans un premier temps, j'ai optimisé le paramétrage de certaines sources sur KB Crawl. Cela a fait diminuer les remontées non pertinentes sans pour autant les faire disparaitre. Au même temps, j'ai commencé à mettre en place le dispositif de remontées d'informations en mode Push via les flux RSS qui se base sur le filtrage d'information sur les contenus XML. Ces derniers présentent l'avantage de contenir des informations structurées et pouvant être mises en surveillance d'une façon efficace puisque les remontées ne concerneront que les nouveaux articles. L'autre avantage est l'absence de rubriques annexes.

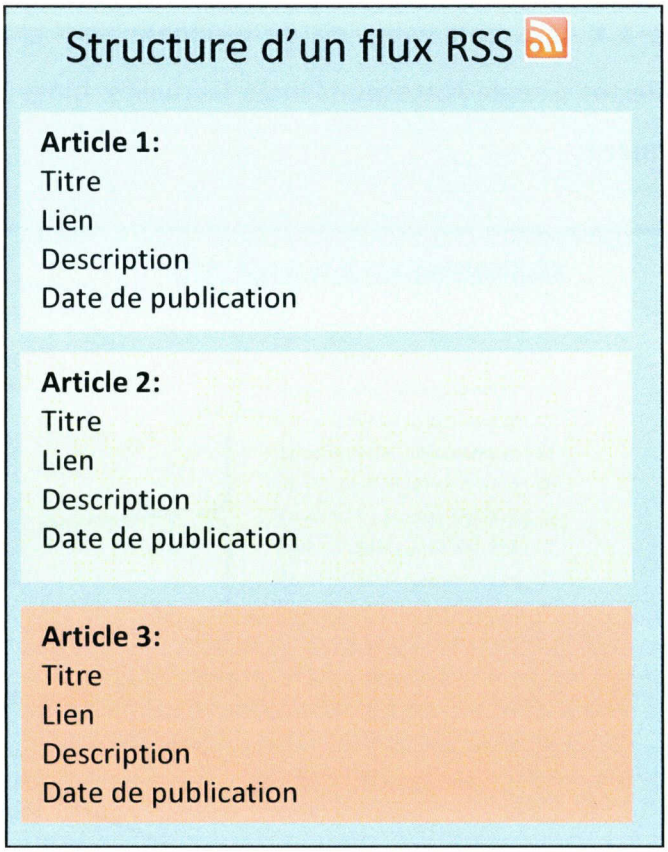


Figure 15 : structure de base d'un document XML

Ainsi, il est possible d'appliquer un filtrage uniquement sur les informations contenues sur l'un des items présents dans le flux. D'une manière générale, les items sont le titre de l'article, son descriptif, le lien URL vers cet article et la date de publication.

Vouloir l'exhaustivité c'est aussi assumer qu'il peut exister plusieurs dizaines de flux à intégrer. Ainsi, face au nombre de sources répertoriées, il était devenu difficile de se repérer sur l'agrégateur Google Reader qui contenait énormément de flux provenant de ceux générés par

Google Alerte et de ceux provenant des sites sélectionnés. Il fallait donc trouver une solution pour réduire le nombre de flux sans pour autant en supprimer.

Yahoo Pipes est un outil permettant de manipuler les flux RSS. Il permet de fusionner plusieurs flux RSS et d'en sortir au final qu'un seul. C'est donc avec son module *Fetch Feed* que j'ai réussi à réduire le nombre de flux présents sur Google Reader.

Selon les besoins, nous pouvons utiliser soit *Fetch Feed* soit le module *Feed Auto Discovery.*

Module Fetch Feed et fusion des flux RSS :

Le but recherché à travers ce module est de fusionner pour chaque thème tous les flux RSS provenant des alertes Google correspondantes (actualité, blogs et vidéos) ainsi que ceux des sources sélectionnées.

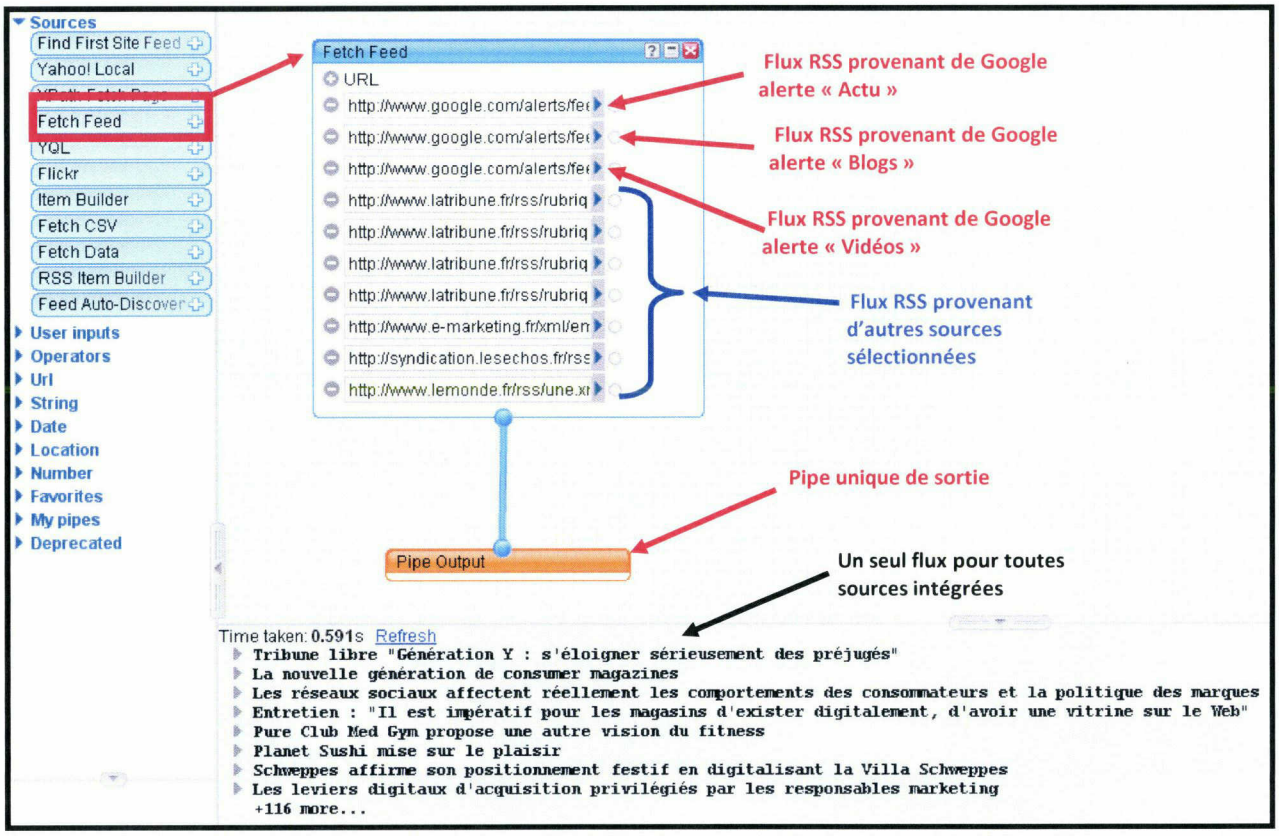


Figure 16 : fusion de plusieurs flux RSS grâce au module *Fetch Feed* de Yahoo Pipes

Après avoir intégré les flux RSS, nous aurons à la sortie un seul flux contenant toutes les remontées de toutes les sources intégrées.

Ce module permet d'avoir la main complète sur les flux à intégrer. Mais il arrivait que tous les flux d'un site intéressent le domaine de veille. Pour ne pas les chercher un par un, j'ai opté pour une technique permettant non seulement de répertorier tous les flux RSS d'un site, mais aussi d'afficher tous les articles de ces mêmes flux. C'est ce que permet le module *Feed Auto Discovery*.

Après avoir renseigné les adresses URL des sites web **(1)**, on arrive à visualiser les flux RSS **(2)**. Pour afficher tous les articles correspondant à tous ces flux **(4)**, j'ai intégré le module *« Fetch Feed »* au module *« Loop »***(3),** tout en cochant le bouton *« emit all results »*.

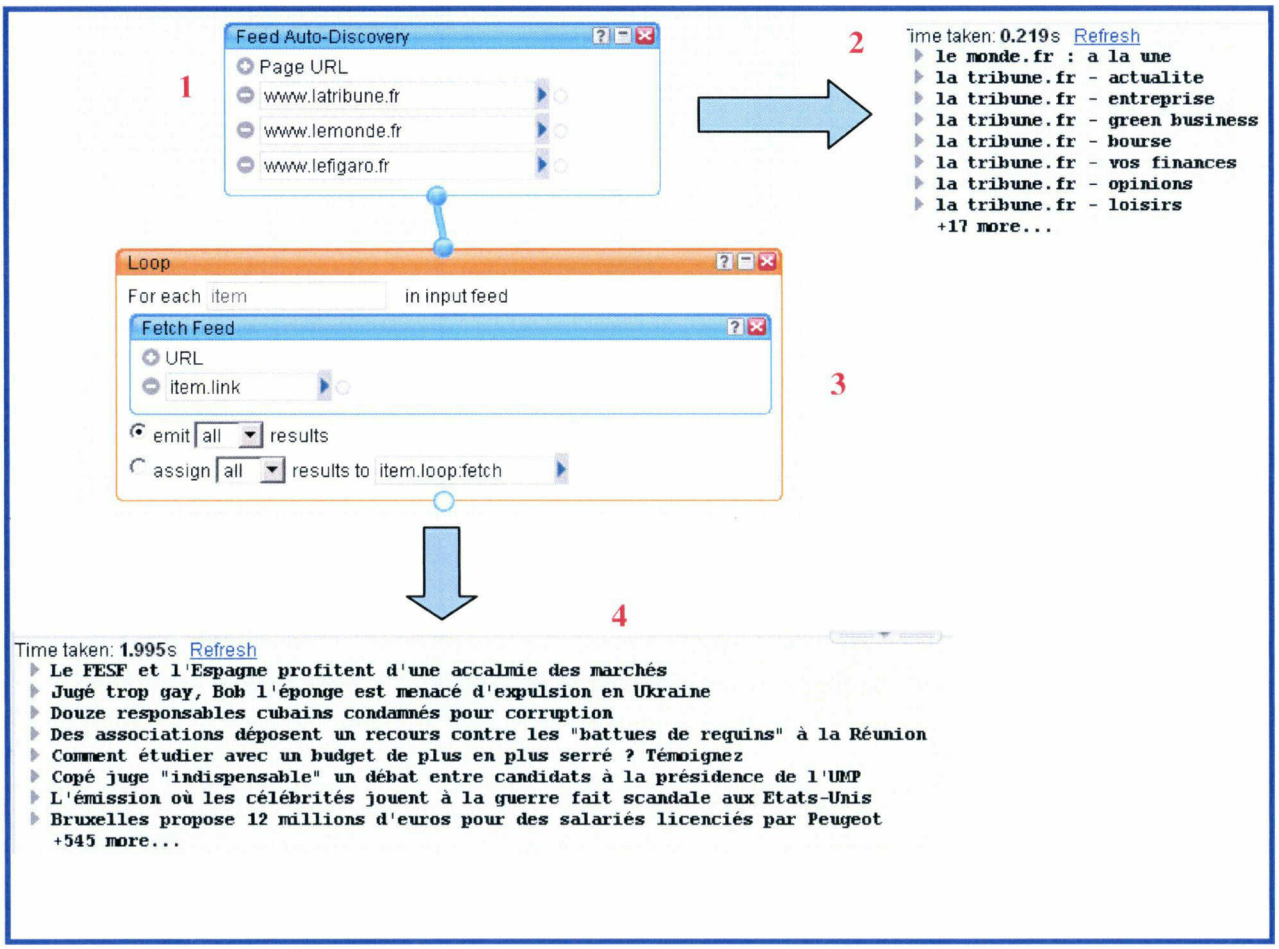


Figure 17 : mode de détection automatique des flux RSS et extraction des articles contenus

Extension des articles :

Avant d'opérer un filtrage sur les flux RSS, une autre problématique ressortait. Pour un internaute ordinaire qui a besoin de ces flux RSS, ces derniers suffisent à ses besoins. Le titre et

le descriptif peuvent être très explicites sur la suite de l'article. Mais pour un veilleur en recherche d'exhaustivité, le titre et la description peuvent ne pas contenir un mot clé appartenant à ses thèmes mis sous surveillance. En effet, dans les flux RSS, la totalité des articles n'est pas affichée car les responsables des sites et pour des raisons de génération de trafic, préfèrent ramener l'internaute vers leurs sites web pour consulter la totalité de l'article. Ce sont des flux RSS tronqués. Pour résoudre cela, j'ai opté pour Yahoo Pipes, qui permet l'extension des descriptions des flux RSS.

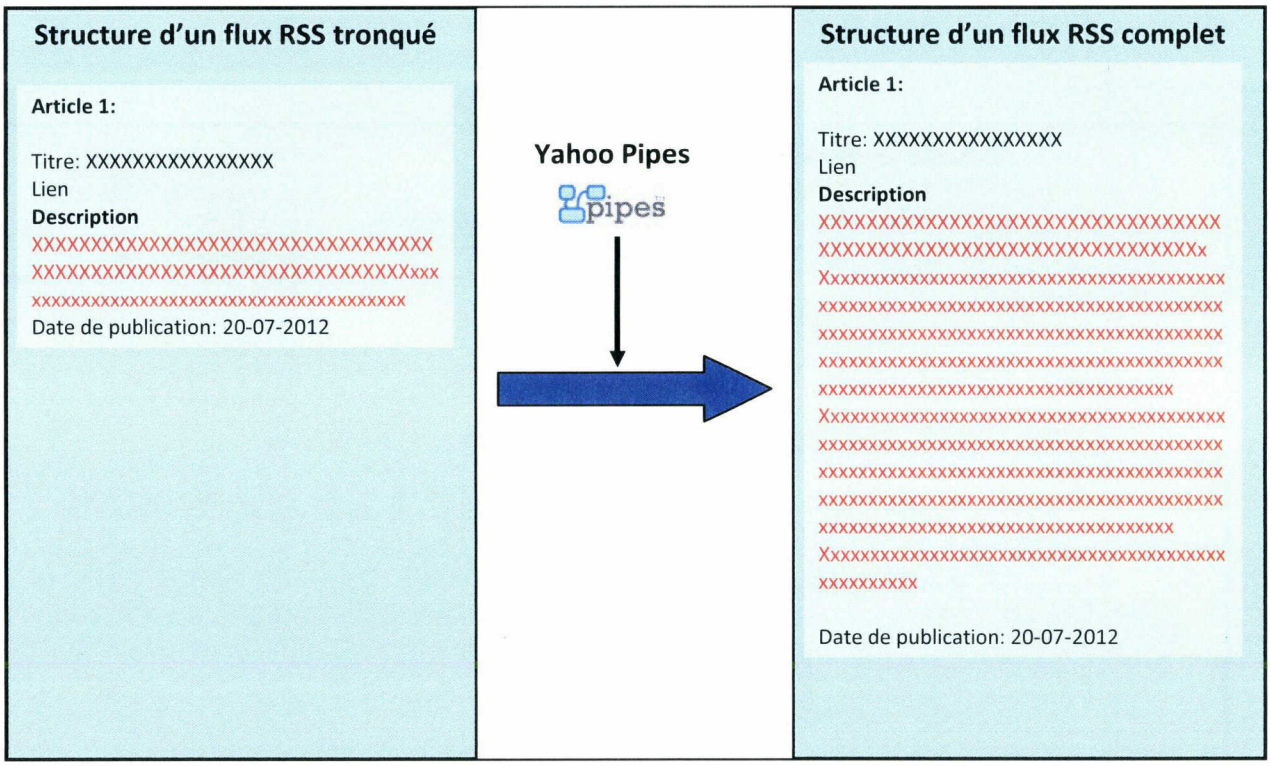


Figure 18 : extension de l'item « *description* » (en rouge) d'un flux RSS grâce à l'outil Yahoo Pipes

Procédure :

Pour pouvoir faire une extension d'un item, il est nécessaire de maitriser le langage HTML.

Cette extension ne s'appliquera que sur les sources sélectionnées car celles de Google Alerte ont déjà subit le filtrage par mots clés lors de la composition des requêtes.

Pour illustrer comment j'ai procédé, j'ai choisi de faire une extension sur un flux du site d'actualité de *Orange Economie*. Chaque source ayant sa propre structuration HTML, il est nécessaire de traiter chaque source indépendamment des autres.

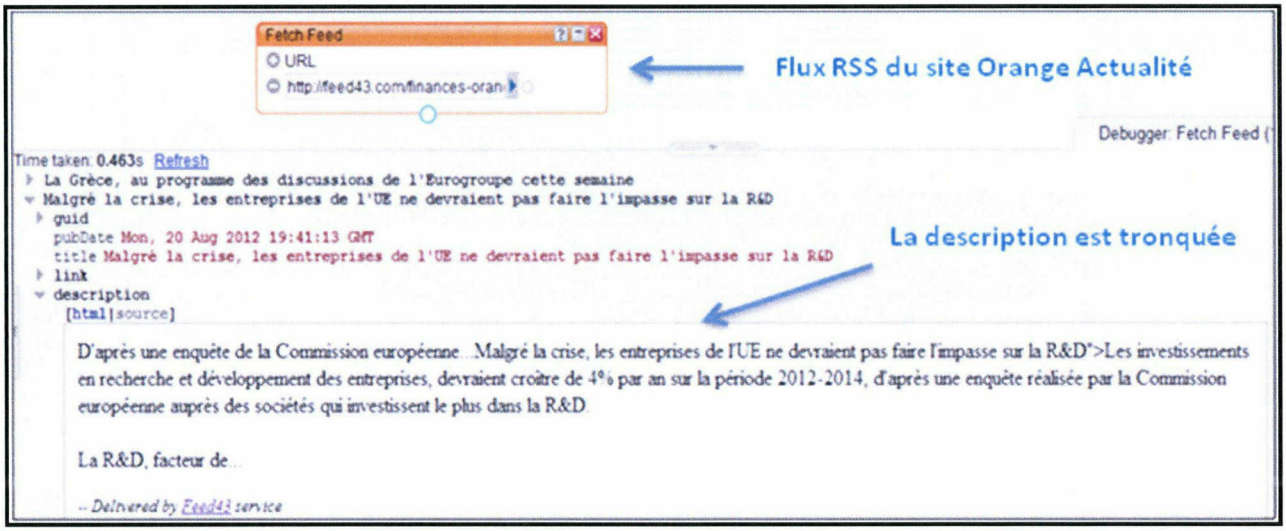


Figure 19 : l'item « description » est tronqué. Il ne contient pas la totalité de l'article.

La page web contenant les articles de Orange Economie présente pour chacun des articles une structure HTML régulière.

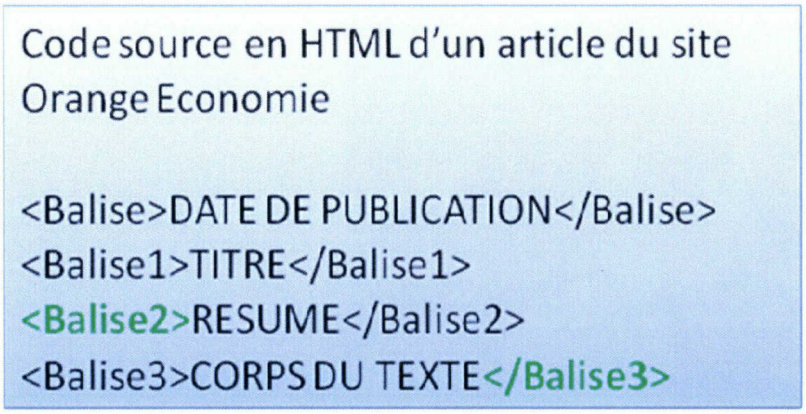


Figure 20 : code source d'un article en langage HTML

En détectant les balises entourant les parties essentielles comme le titre, le corps du texte et la date de publication, nous pouvons extraire l'intégralité de l'article et l'assigner à l'item *« description »*. Pour cela on utilisera les modules *« Loop »* et *« Fetch Page »*.

Dans l'exemple ci-dessus, on va demander à Yahoo Pipes que pour chaque lien détecté, d'extraire les informations contenues entre la balise ouvrante <Balise2> et la balise fermante </Balise3>, et d'assigner ces informations dans l'item « description ».

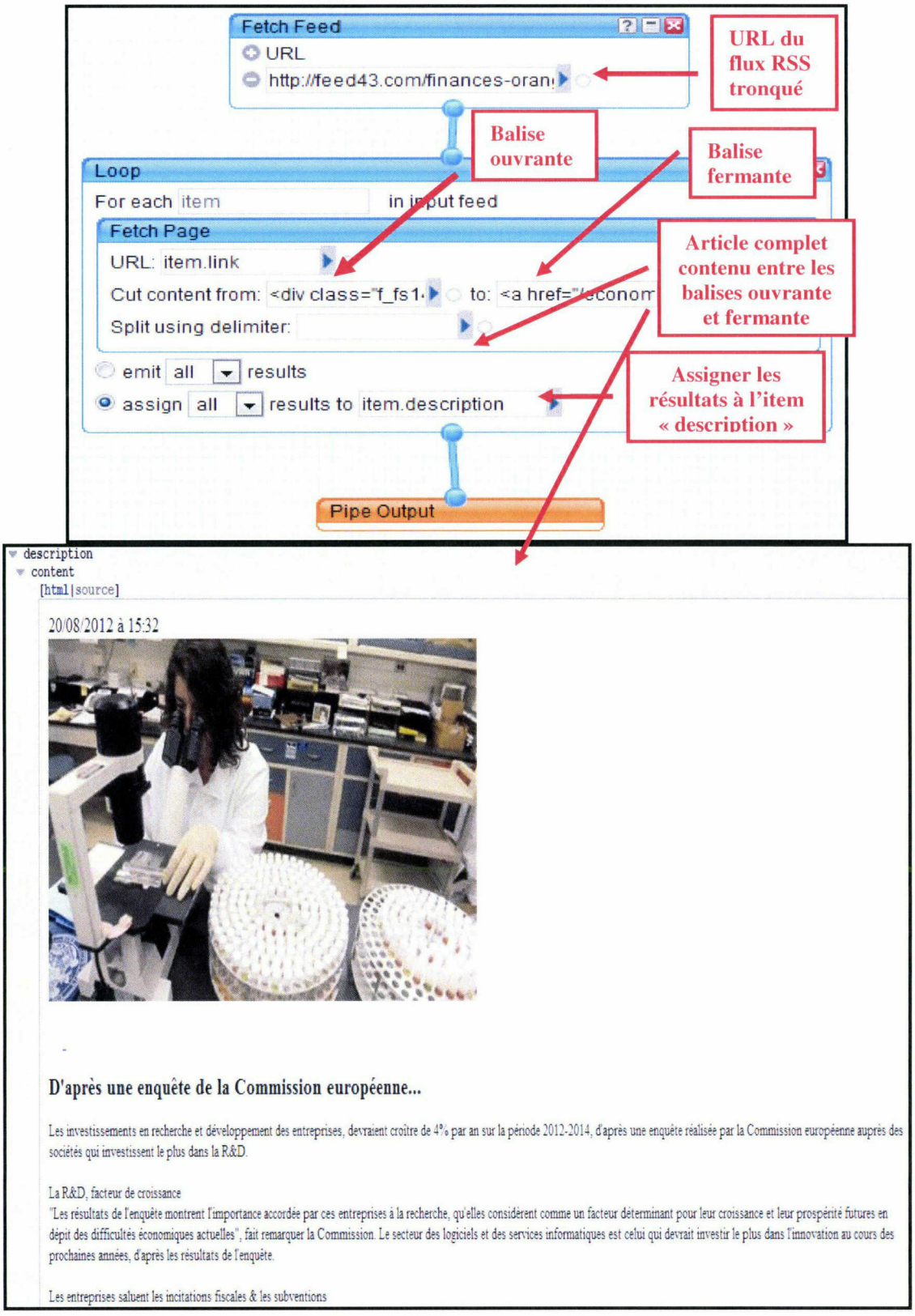


Figure 21 : l'item « description » contient l'intégralité de l'article

Filtrage sur le contenu

Une fois l'intégralité des articles extraite, il fallait appliquer un filtrage sur ce contenu via des mots clés. Yahoo Pipes apporte encore une fois la solution en permettant un filtrage sur différents items : le titre, la description, la date de publication ou l'url, grâce au module « *Filter* ».

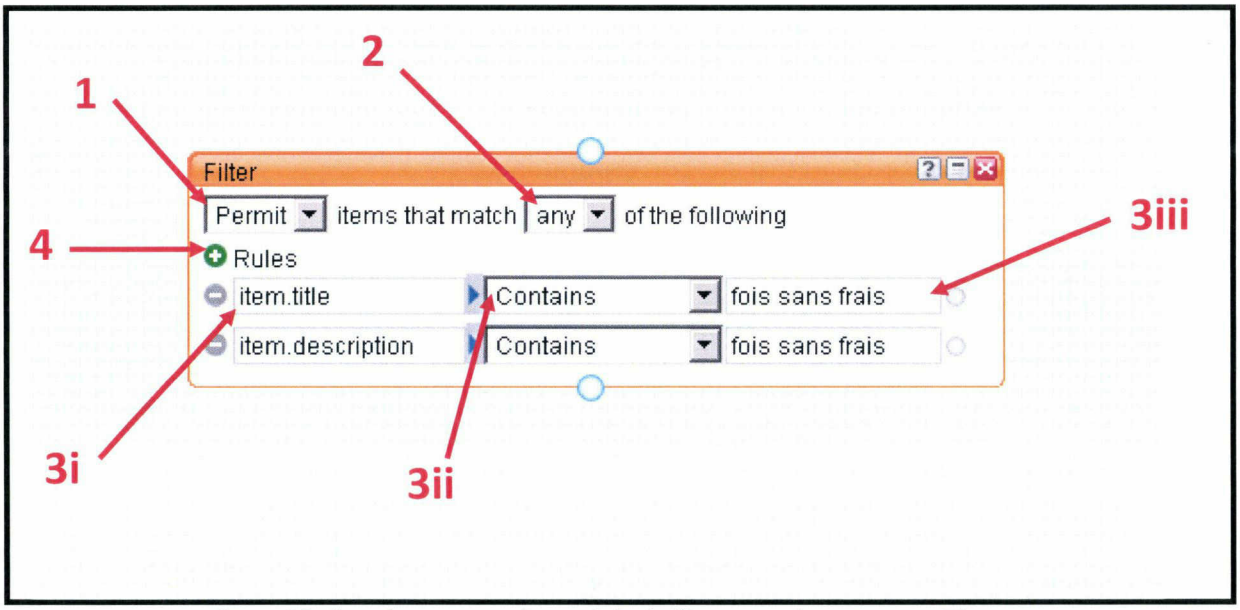


Figure 22 : fonctionnement du module de filtrage sur le contenu « *Filter* »

Ce module peut prendre en compte des mots clés comme il peut aussi bloquer certains.

Dans notre cas, nous allons permettre la remontée de tout article contenant l'expression « *fois sans frais* » dans le titre et la description. Pour cela nous procéderons comme suit :

1- dérouler la première liste et sélectionner « *Permit* » (si nous sélectionnons « *Block* », nous bloquerons le ou les mots clés indiqués) ;

2- dérouler la liste suivante et sélectionner « *any* ». Cela voudra dire que le filtre s'appliquera sur le titre OU la description (si nous avions sélectionnée « *all* », le mot clé doit se trouver ET dans le titre ET dans la description) ;

3- Sous « *Rules* », il y a trois rubriques, (3i) la première permet d'indiquer dans quelle partie de l'article il faut appliquer le filtre. Sélectionner « title ». (3ii) La deuxième permettra de soit permettre soit bloquer le ou les mots-clés. (3iii) La troisième rubrique est celle où on saisira la ou les mots-clés.

4- Pour rajouter un autre filtrage, cliquer sur le signe Rules. Nous aurons les mêmes rubriques, où cette fois nous sélectionnerons « *descritpion* » au lieu de « *title* ».

Caractéristiques du module « *Filter* » :

Le module *Filter* est insensible à la casse mais l'est pour la suite des caractères. Il ne permet malheureusement pas de composer des requêtes textuellement comme sur des moteurs de recherche à moins d'utiliser les expressions régulières. Par contre, il offre des champs imitant ces derniers grâce à des rubriques correspondant aux opérateurs booléens.

Le tableau ci-dessous montre ces correspondances :

Opérateurs booléens de Google	Leurs correspondants dans le module *Filter sans REGEX*	Leurs correspondants dans le module *Filter avec REGEX*
AND (+)	*All , Permit*	
OR (|)	*any*	|
NOT (-)	*Block , Does not contain ,*	^
" "	Le module *filter* est fidèle à la suite des caractères	Le module *filter* est fidèle à la suite des caractères mais sensible à la casse
allintitle (intitle)	*Item.title*	*Item.title*
allinurl (inurl)	*Item.link*	*Item.link*
Aucune correspondance	**Aucune correspondance**	*Matches regex* (utilisation des expressions régulières)

Tableau 5 : liste des opérateurs booléens et leurs correspondances dans le Yahoo Pipes

Il arrive parfois de vouloir remonter des informations qui présentent des caractéristiques répétitives mais pas toutes identifiables, comme par exemple des articles qui débutent ou finissent par tel ou tel mot, ou seulement les titres contenant des chiffres. Pour y parvenir, Yahoo Pipes via son module *Filter* permet cela en sélectionnant *« Matches Regex »* au lieu de *« contain »*. Cela autorise à utiliser un filtrage selon les expressions régulières (REGEX).

Ainsi, les expressions régulières suivantes permettent :

\d → tous les articles contenant au moins un chiffre
^Comment → tous les titres ou descriptions commençant par le mot «comment»
dette$ → tous les titre ou descriptions finissant par le mot « dette »
comme\s → remonte les titres ou descriptions contenant le mot « comme » suivi d'un espace. Cela évite de remonter des mots ayant la même racine tels commentaire, commerce, etc.

Figure 23 : quelques expressions régulières et leurs fonctions

En rouge, sont les opérateurs utilisés pour exprimer les expressions régulières en langage Perl. Il faut par contre faire attention car ce module est sensible à la casse. Pour éviter cela, deux solutions s'imposent : soit renseigner toutes les variantes du mot. Par exemple si on recherche le mot cofidis quelques soient ses variantes, on écrira cette requête : cofidis|Cofidis|COFIDIS, soit utiliser un opérateur qui ignore la casse tel « i » ou employer les crochets entourant la lettre en majuscule et en minuscule : cofidis s'écrira ainsi [Cc][Oo][Ff][Ii][Dd][Ii][Ss].
La liste des principaux opérateurs en expressions régulières est en annexe 1.

Chronologie des retombées :

Yahoo Pipes affiche les résultats selon l'intégration des flux RSS. Ainsi, il ne tient pas compte de la date de publication des articles. Pour forcer la publication chronologique, on va utiliser le module « *Sort* ».. Pour cela, il faut sélectionner « *item.pubDate* » et choisir « *descending* » pour classer les articles selon la date la plus récente à la plus ancienne.

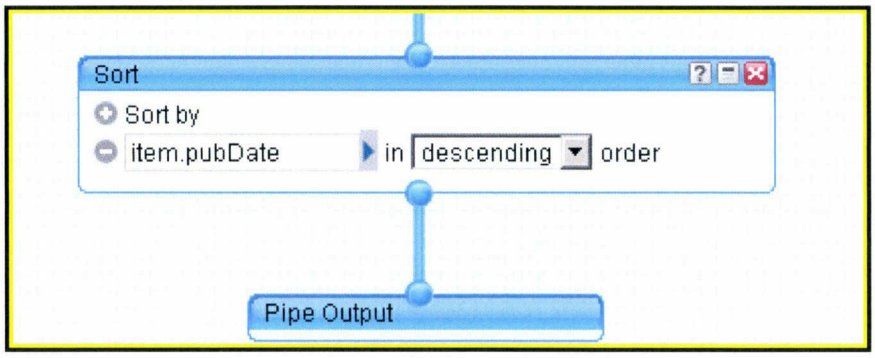


Figure 24 : le module « *Sort* » permet de classer les articles selon leurs dates de publication

Supprimer les doublons :

Pour diverses raisons (un même article sur différentes rubriques du site, ou un même titre pour plusieurs sources), il arrive parfois qu'un même article soit remonté plus d'une fois. Cela crée des doublons indésirables. Pour les supprimer, j'ai utilisé le module dédié « *Unique* » qui permet de supprimer les doublons ayant la même chaîne de caractères dans le titre, la description ou le lien URL. Plusieurs options s'offrent selon les besoins. Ainsi, si on désire supprimer tous les doublons provenant de plusieurs sources, on va sélectionner « *Title* » et dans certains cas « description » même s'il est très rare que deux sources différentes ont la même description. Si par contre, on veut supprimer un doublon provenant de la même source, on l'appliquera sur « *Title* », « *description* » et « *Link* ».

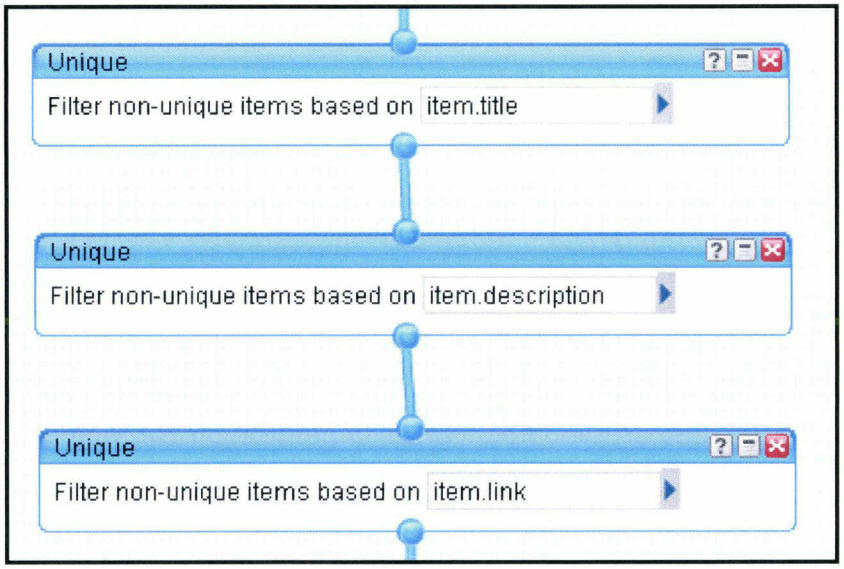
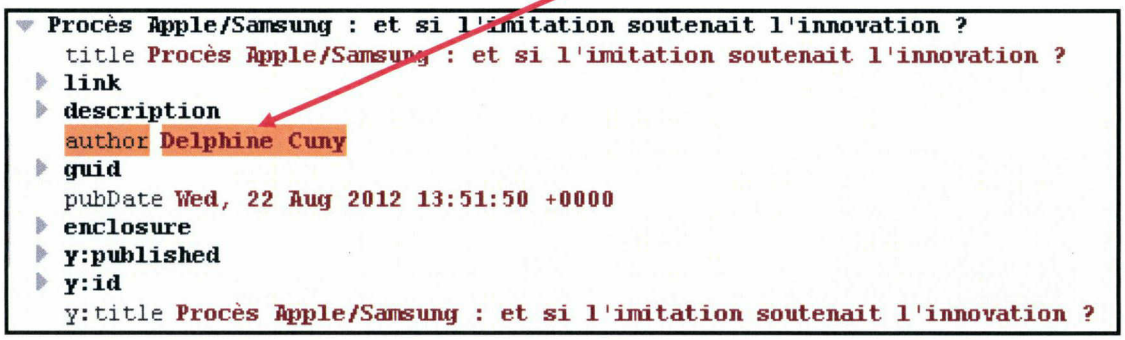


Figure 25 : le module « Unique » permet de supprimer les doublons

Problèmes d'affichage et d'horodatage sur Google Reader :

Google Reader affiche toujours sur sa partie droite la date et l'heure de la remontée de l'article. Cet horodatage des articles n'est pas exact.

Il faut savoir que Google Reader affiche en dessous de chaque titre, la source d'information d'où provient l'article. Il se réfère pour cela à l'item « *Author* » dans Yahoo Pipes.



D'un autre côté, il lui arrive d'afficher l'auteur de l'article au lieu du nom de la source, ou carrément ne pas reconnaître la source en affichant dans ce cas le message « *Author Unknow* ». Pour arriver à visualiser la vraie date et la vraie heure de publication, ainsi que la source d'information même si elle n'est pas reconnue, j'ai effectué certaines manipulations sur Yahoo Pipes. nous devons extraire la partie de domaine de chaque site. Elle se trouve dans l'item *Link*. Pour cela, nous procéderons comme suit :

1- nous ne devons jamais toucher au lien (*Link*) sinon le lien sur Google Reader ne fonctionnera pas. Donc, on recopie ce lien et le plaçons dans « *Author* ». Cela est rendu possible grâce aux modules **(1i)** « *Loop* » situé sous la rubrique « *Operators* »et **(1ii)** « String *Regex* » situé sous la rubrique « *String* ».

2- Pour chaque source remontée, nous demanderons à Yahoo Pipes que **(2i)** pour chaque item « *Author* », **(2ii)** de remplacer cet item par **(2iii)** le lien «*link*» et de **(2iv)** l'assigner à « *Author* ».

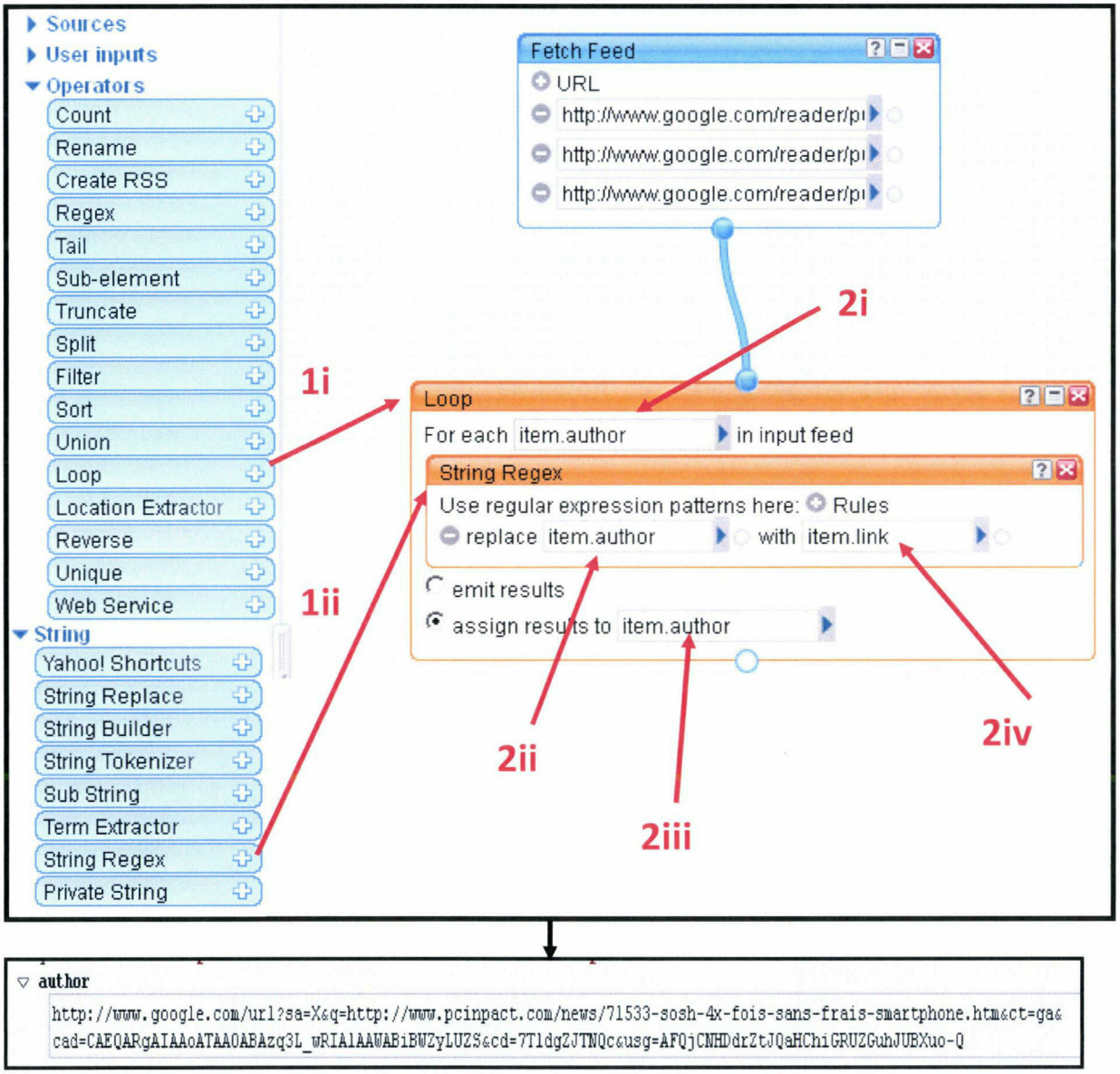


Figure 26 : mode d'épuration des API Google Alerte

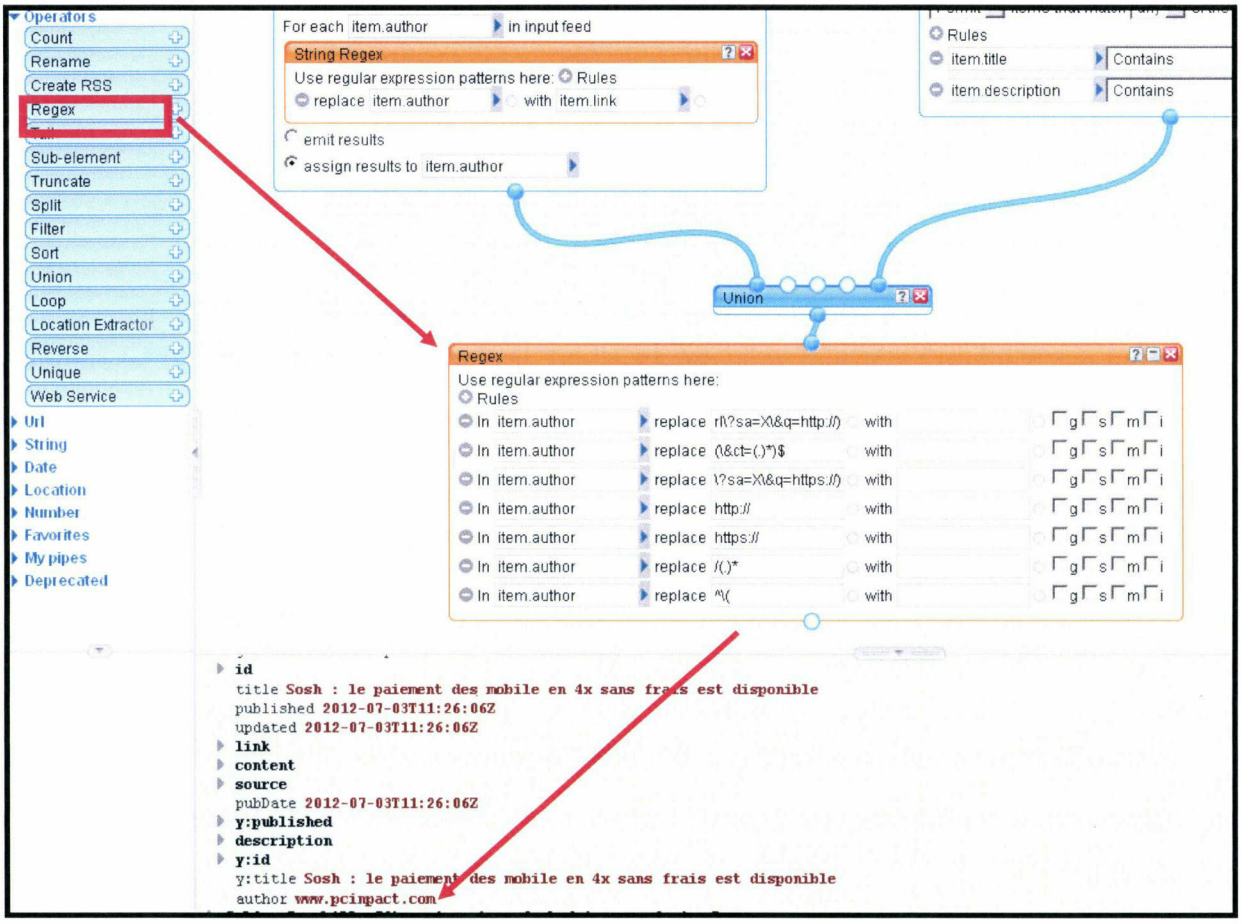
Nous allons maintenant récupérer la partie du lien qui nous intéresse.

Nous utiliserons le module *« Regex »* situé sous la rubrique *« Operators »*

Yahoo Pipes récupère dans l'item *Link,* l'API de Google Alerte. Elle est composée d'une partie fixe (en rouge), du nom de domaine de la source (en vert) et d'une dernière partie appartenant à la source et propre à chaque article (en bleu)

Nous allons lui demander via ce module de supprimer les parties du lien qui ne nous intéressent pas (celles en rouge et en bleu) et de ne récupérer que la partie en vert.

http://www.google.com/url?sa=X&q=http://www.pcinpact.com/news/71533-sosh-4x-fois-sans-frais-smartphone.htm&ct=ga&cad=CAEQARgAIAAoATAAOABAzq3L_wRIAlAAWABiBWZyLUZS&cd=7TldgZJTNQc&usg=AFQjCNHDdrZtJQaHChiGRUZGuhJUBXuo-Q



Intégration des expressions régulières pour tirer l'essentiel de l'url

En agissant sur l'item « *Author* », nous allons via les *expressions régulières* récupérer cette partie.

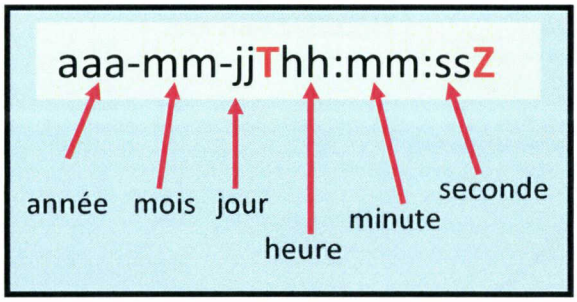
Dans le module Regex, je lui demande de remplacer la partie rouge http://www.google.com/url?sa=X&q= par rien du tout. J'utilise pour cela cette expression régulière : (http://www.google.com/url\?sa=X\&q=http://). On remarquera que j'ai intégré cette partie entre parenthèses et ajouté deux barres d'échappement permettent de représenter les caractères figurant déjà dans les expressions régulières (? et &).

Pour supprimer la partie en bleu /news/71533-sosh-4x-fois-sans-frais-smartphone.htm&ct=ga&cad=CAEQARgAIAAoATAAOABAzq3L_wRIAlAAWABiBWZyLUZS&cd=7TIdgZJTNQc&usg=AFQjCNHDdrZtJQaHChiGRUZGuhJUBXuo-Q, j'ai demandé à Yahoo Pipes de la remplacer aussi par rien du tout, via cette expression régulière : /(.)*

Qui signifie : remplacer tous les caractères se trouvant après le slash (/) par rien du tout.

Ensuite, j'agis sur la date de publication récupérée par Yahoo Pipes.

Elle est sous ce format



Le but est d'épurer ce format en le débarrassant des lettres en rouge T et Z et les remplacer par des espaces pour pouvoir mieux distinguer la date et l'heure et de placer le jour avant le mois et l'année. J'ai utilisé cette expression régulière pour cela :

Remplacer [\d*]- [\d*]- [\d*]T[\d*]: [\d*]: [\d*]Z par $3-$2-$1 $4:$5:$6

$1 $2 $3 $4 $5 $6

\d signifie tout caractère numérique. Et du fait que j'ai placé l'étoile * à l'intérieur des crochets, l'ensemble [\d*] est interprété par le sigle $. Les chiffres suivant $ correspondent à l'ordre de priorité.

Il ne reste plus qu'à assembler l'ensemble « source + date et heure » et les assigner à l'item « *Author* » pour que Google Reader affiche cet ensemble. nous allons placer les dates et heures juste après la source.

Nous utiliserons pour ce faire, le module « *Loop* » situé sous la rubrique « *Operators* » et le module « *String Builder* » situé sous la rubrique « *Strings* ».

Nous demanderons au couple de modules de structurer la source comme suit :
Source suivie de *espece : espace*, suivis date et heure de publication.

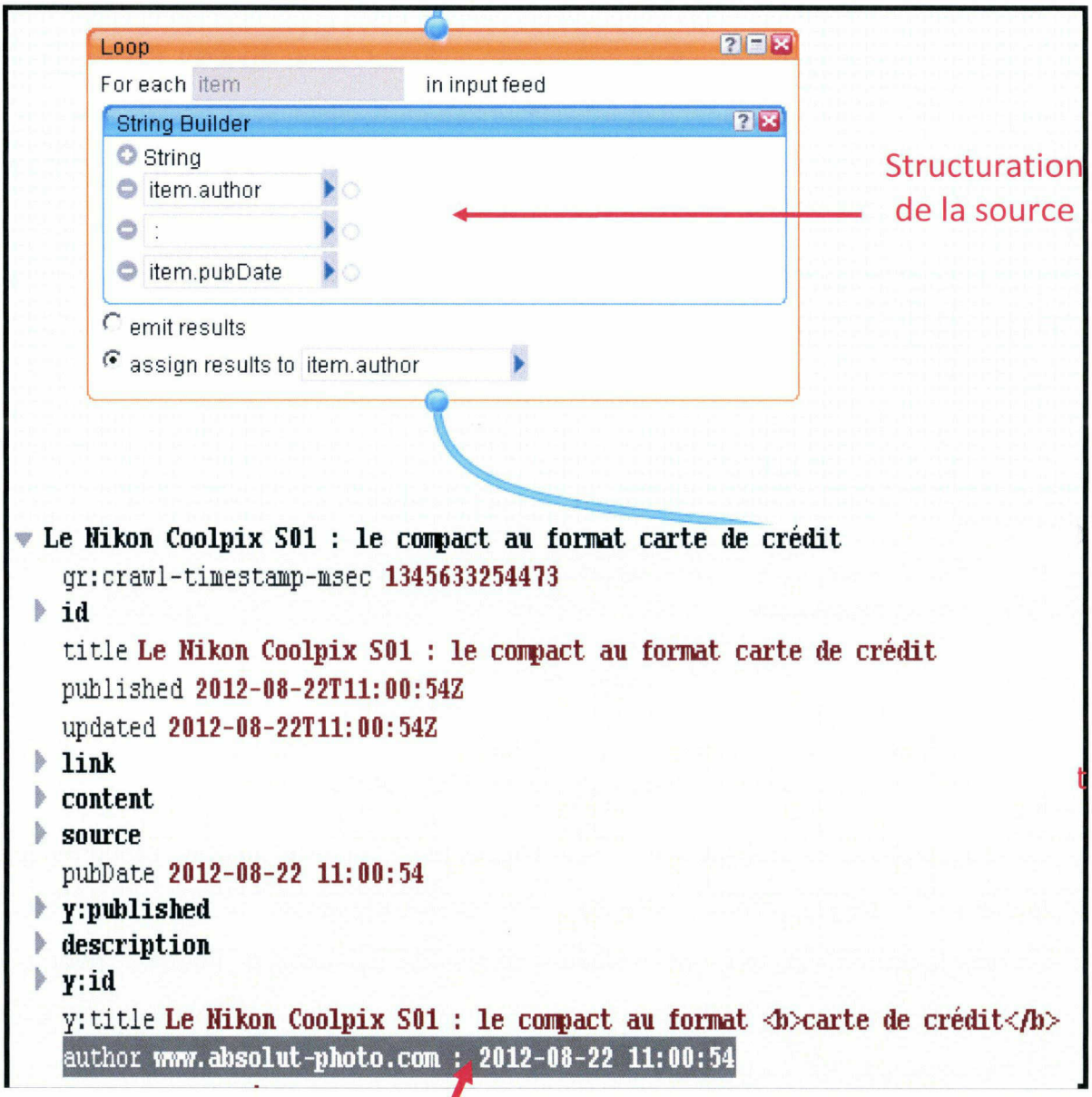


Figure 27 : structuration de la source et des date et heure de publication dans l'item « *Author* »

Ainsi, nous obtenons une meilleure présentation de la source et de la date de publication.

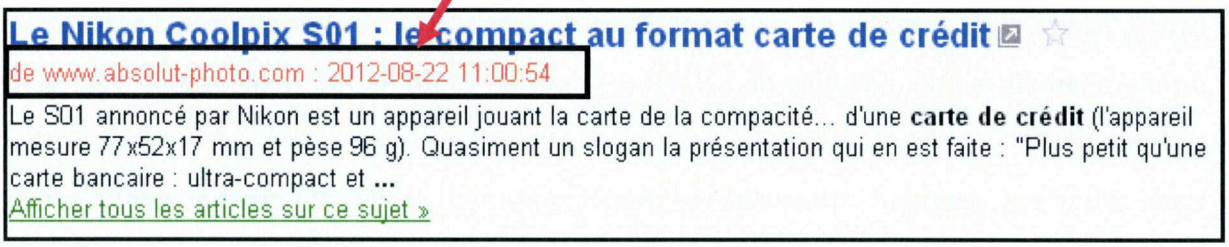


Figure 28 : format final de la source et de la date et heur de publication

Cas particulier : filtrer les tweets :

Twitter est un site de microblogging qui a connu un grand succès mondial. Il a conquis plusieurs millions d'internautes en général et les veilleurs en particuliers. Sa principale force est que les messages envoyés sont courts donc traitent directement le cœur de l'information. Il permet de s'informer de l'actualité d'une façon immédiate et rapide.

Si cela était rendu possible grâce aux flux RSS pour chaque recherche, Twitter a depuis peu décidé de les supprimer. Mais heureusement il existe une parade en jouant sur les API suivantes :

- Trouver les tweets contenant un mot :
  http://search.twitter.com/search.atom?q=**cofidis**
- Trouver les tweets d'un utilisateur :
  http://search.twitter.com/search.atom?q=from%3A**cofidis**
- Trouver les tweets destinés à une personne :
  http://search.twitter.com/search.atom?q=to%3A**cofidis**
- Trouver les tweets mentionnant une personne :
  http://search.twitter.com/search.atom?q=%40**cofidis**
- trouver des tweets contenant un hashtag :
  http://search.twitter.com/search.atom?q=%23**cofidis**

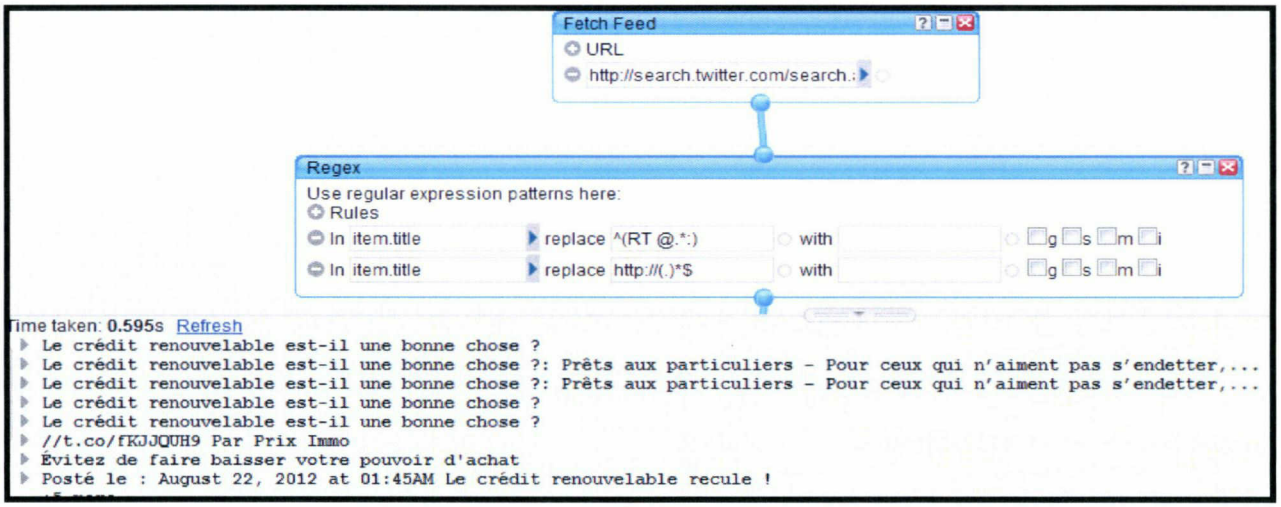
Après avoir résolu ce problème et obtenu les flux RSS à intégrer sur dans Yahoo Pipes, il fallait résoudre un autre problème, celui lié à l'infobésité provenant de Twitter. En effet, il y a énormément de tweets qui sont retweetés, ce qui crée beaucoup de doublons pour une même information. Il aurait été facile d'appliquer le module de dédoublonnage *« Unique »,* mais cela ne marchera pas car aucun tweet ne contient la même chaîne de caractères même pour des tweets identiques et cela pour plusieurs raisons :

la première raison est l'ajout automatique par Twitter de morceaux de chaînes de caractères lors d'un retweet. Ces morceau sont RT suivi de @moncompte au début du tweet. L'autre raison est l'url pointant vers la source d'information et qui est de la forme http://lienverslasource à la fin du tweet. Ces url sont raccourcies par différents sites, ce qui ne donnera jamais la même chaîne de caractères pour cette url même si elle pointe vers la même source.

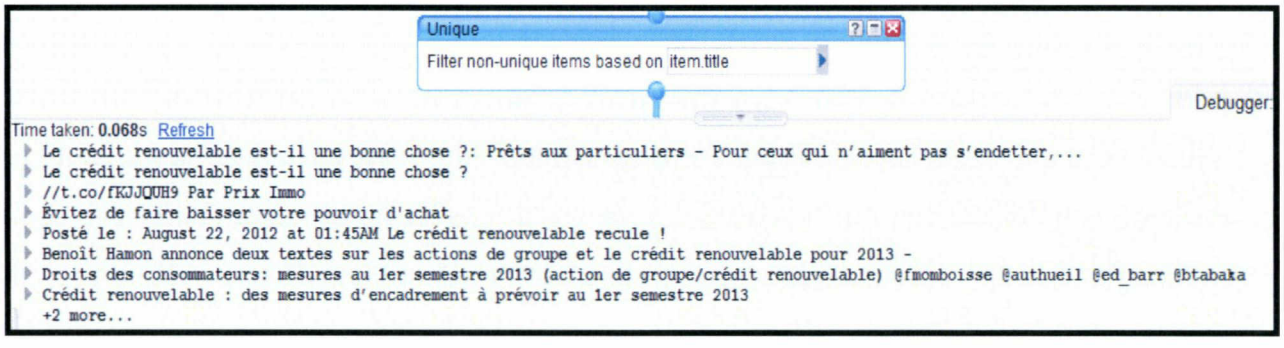
Pour éviter ces doublons, nous devons donc supprimer ou plutôt chercher-remplacer les RT @moncompte et http://lienverslasource par du vide, comme ça le tweet sera « propre ».
Nous utiliseront pour cela les expressions régulières avec le *Module Regex :*

- Pour RT @moncompte → ^(RT @.*:)
- Pour http://lienverslasource → http://(.)*$

Etat des remontées avant filtrage par les expressions régulières pour supprimer les doublons



Cela permet d'appliquer le module *Unique* pour éliminer les doublons



III-4- Participation aux différentes tâches quotidiennes

En parallèles à ces missions, j'étais chargé de plusieurs tâches quotidiennes et d'études ponctuelles.

Les tâches quotidiennes consistaient à traiter les alertes reçues par KB Crawl, les abonnements presse et les outils de veille tarifaire, publicitaire et panéliste. Cela passait par leur traitement et leur classification dans des dossiers dédiés.

J'ai aussi mené des études Ad' Hoc provenant de demandes internes. J'ai profité de ces sollicitations spécifiques pour tester l'efficacité du dispositif que je devais mettre en place.

J'étais également en chargé de la rédaction de livrables veille et d'alertes alimentant le blog interne.

Pour faciliter la prise en main du nouveau dispositif, j'ai mis sous la main des collaborateurs du DVE, deux tutoriels. L'un expliquant en détails l'utilisation de Yahoo Pipes et l'intégration des flux dans Google Reader, et l'autre montrant comment créer des flux RSS pour les sites qui n'en proposent pas, grâce à l'outil Feed43.

IV- Evaluation du bilan de la méthodologie adoptée

Evaluer les résultats du dispositif mis en place revient évaluer la qualité du filtrage d'information qui est la base même du dispositif. Cela revient aussi à distinguer les documents pertinents et non pertinents.

Du fait que les informations remontées sont sélectionnées en fonction des requêtes composées de mots clés, et en considérant que ces mêmes requêtes ont été optimisées grâce aux opérateurs booléens incorporés dans les moteurs de recherche (Google Alerte) ou dans le module de filtrage de Yahoo Pipes, nous considéreront que toutes les remontées sont pertinentes de ce point de vue. Mais d'un point de vue pratique et aux yeux du Département Veille et Etudes, la pertinence a une autre connotation. Ce n'est pas parce que le système a remonté une information qui répond fidèlement à la requête qu'elle sera considérée comme pertinente. En effet, pour diverses raisons, cette pertinence ne peut être seulement examinée sous le critère de la précision de la requête.

La première raison est liée à l'algorithme du moteur de recherche de Google Alerte, qui en ignorant certains caractères, peut fausser les résultats. Pour mieux comprendre, j'illustre un exemple. Nous surveillons toutes les nouveautés relatives à un concurrent, *Banque Accord* par exemple. Lors du paramétrage, je compose cette requête : **"banque accord".** Je fais signifier au moteur via les deux crochets, de rechercher exactement la suite de caractères **banque accord.** Mais il se trouve que Google ignore lors de ses recherches certaines caractères tels les signes de ponctuation. Ainsi, lors de son crawl, s'il retrouve une suite de caractères telle **banque : accord** il remontera l'information correspondante. Ainsi nous aurons un article parlant d'un accord d'une quelconque banque et non pas de **banque accord**.

L'autre raison est le degré de proximité des mots clés. Il devient difficile lors de requêtes comportant l'opérateur booléen « AND » de remonter une information qui soit de qualité. Si nous recherchons tout article traitant des paiements mobiles en relation avec un acteur du crédit à la consommation, il n'est pas rare d'être confronté à un article où les deux expressions soient suffisamment éloignées pour considérer la remontée non pertinente. Ainsi, malgré l'existence d'opérateurs de proximité, ceux-là ne sont pas suffisamment fiables.

La troisième raison provient des mots ayant plusieurs sens. Le mot « orange » est une marque de télécommunications qui peut renvoyer aussi vers un article parlant des oranges en tant que fruit ou en tant que couleur.

S'il existe des métriques dans le cadre de TREC pour évaluer la qualité des systèmes de recherche d'information, elles ne sont toujours pas applicables et adaptées à l'évaluation des systèmes de filtrage d'information **[8]**. En effet, les métriques de la recherche d'information se basent sur le rappel et la précision.

Précision = nombre documents pertinents retrouvés / nombre documents retrouvés

Rappel = nombre documents pertinents retrouvés / nombre documents pertinents dans la base

Le taux de rappel mesure la capacité des systèmes évalués à retrouver tous les documents pertinents répondant à une requête donnée, alors que le taux de précision mesure la qualité des réponses fournies par le système **[8].** Ainsi, si on veut calculer le taux de rappel, il faut au préalable connaître le nombre de documents pertinents dans la base interrogée, ce qui ne peut être envisageable car cette base dans notre cas est le web. Nous ne pourrons donc pas évaluer

ce taux. Pour y remédier, différentes propositions de TEC ont été lancées en introduisant dans les évaluations des systèmes de filtrage d'information *le critère d'utilité* **[8]**. Ce critère a été introduit au cours de TREC-4 qui a proposé pour toute expérience **Ri** évaluant la capacité des systèmes de filtrage à trier un ensemble de documents en deux catégories A et B, cette métrique : **Ui = uaiAi + ubiBi [28].**

Où **Ai** est le nombre de documents pertinents trouvés, et **Bi** le nombre de documents non pertinents. **Uai** et **ubi** sont des constantes correspondant aux valeurs d'utilité données par l'utilisateur. Ce sont ces dernières qui varient selon que l'on privilégiera la qualité des réponses, leur quantité ou un équilibre entre les deux. Ces trois scénarios sont identifiés dans le tableau suivant :

Expériences	Valeurs des constantes d'utilité	Mesure du SFI
R1	ua1 = 1 , ub1 = - 3	u1 = A1 - 3B1
R2	ua2 = 1 , ub2 = - 1	u2 = A2 – B2
R3	ua3 = 3 , ub3 = - 1	u3 = 3A3 – B3

Tableau 6 : métrique proposée au cours de TREC-4

Il s'est avéré rapidement que cette métrique ne permettait pas d'évaluer globalement l'efficacité d'un SFI, mais seulement de comparer entre plusieurs systèmes pour une même requête. A partir de TREC-6, la notion d'utilité prend deux paramètres : décision de sélection et de pertinence **[8], [28]**.

Documents	Pertinents	Non pertinents
Sélectionnés	R_+ / λ1	S_+ / λ2
Non sélectionnés	R_- / λ3	S_- / λ4
Métrique	$U(\lambda1,\lambda2,\lambda3,\lambda4) = \lambda1R_+ + \lambda2S_+ + \lambda3R_- + \lambda4S_-$	

Tableau 6 : métrique d'utilité se basant sur les deux paramètres : *Sélection* et *Pertinence* **[8]**

- R_+ représente le nombre de documents pertinents sélectionnés par le système ;
- S_+ est le nombre de documents non pertinents sélectionnés par le système ;
- R_- représente le nombre de documents pertinents non sélectionnés par le système ;

- S_- est le nombre de documents non pertinents non sélectionnés par le système ;
- $\lambda 1, \lambda 2, \lambda 3, \lambda 4$ représentent la pondération affectée à chaque document dans sa catégorie de correspondance. Le tableau suivant montre les propositions des différents TREC pour pondérer ces paramètres :

Version TREC	Fonction	Commentaires
TREC-6	$U(3,-2,0,0) = F1 = 3*R_+ - 2*S_+$ $U(3,-1,-1,0) = F2 = 3*R_+ - S_+ - R_-$	
TREC-7	$U(4,-1,0,0) = F3 = 4*R_+ - 1*S_+$	Suppression de $\lambda 3$ car difficile d'estimer le nombre exact des documents pertinents non retrouvés.
TREC-8	$NF1 = 6*(R_+)^{0.5} - S_+$ $NF3 = 6*(R_+)^{0.8} - S_+$	
TREC-9	$U = 2*R_+ - S_+$	

Tableau 6 : les différentes fonctions proposées au cours des conférences TREC **[8]**

IV-1- Analyse des résultats

Avant l'évaluation du dispositif que j'ai mis en place, il fallait sélectionner quelle fonction TREC pouvait répondre au mieux à ce dispositif. Ce choix dépendra des paramètres à prendre en considération lors de l'évaluation :

- Le nombre de documents mis en liste de suivi de Google Alerte sont des documents considérés comme pertinents et correspondent dans la métrique TREC à $\mathbf{R_+}$ avec une pondération élevée.
- Le nombre de doublons remontés. Je distingue deux sortes de doublons : les doublons de **catégorie 1** qui sont des remontées ayant la même source et donc la même url, car même avec un système de dédoublonnage de Yahoo Pipes, il arrive et pour des raisons techniques que le filtre ne fonctionne pas correctement à tous les crawls du système. Et des doublons de **catégorie 2** qui sont des remontées n'ayant pas la même url mais traitant de la même information.
- Les remontées sélectionnées par le système mais considérées par l'expertise humaine comme non pertinentes par rapport à l'environnement de l'entreprise correspondront à $\mathbf{R_+}$ mais avec une pondération faible.

- Les remontées non pertinentes mais sélectionnées par le système et provenant de la non prise en charge de Google Alerte des chaînes de caractères tels les caractères de ponctuation, ou la non proximité des mots clés, correspondent à S_+.

C'est ainsi que le choix s'est porté sur la métrique de TREC-9 **($U = 2*R_+ - S_+$)** qui ne prend pas en compte le nombre de documents pertinents non sélectionnés car il est très difficile de l'estimer. Comme mon système de filtrage prend en compte certains paramètres relatifs aux documents R_+, cette métrique sera sous la forme suivante : $U = \lambda 1 R1_+ + \lambda 2 R2_+ + \lambda 3 R3_+ + \lambda 3 R4_+ - \lambda 5 S_+$.

Où :

$R1_+$ est le nombre de documents pertinents sélectionnés par le système et mis en liste de suivi de Google Alerte ;

$R2_+$ le nombre de doublons de catégorie 2 ;

$R3_+$ le nombre de doublons de catégorie 1 ;

$R4_+$ le nombre de remontées considérées pertinentes par le système mais pas par l'humain.

Si évaluer plusieurs systèmes de filtrage ne demande qu'à comparer ceux-ci entre eux, évaluer un seule système s'avère plus complexe à réaliser. Je pouvais calculer la précision du système en prenant en compte les documents pertinents selon l'expertise humaine (P_H). Cela ne constituerait pas un vrai indice pour évaluer la qualité de filtrage du système car ce dernier ne peut remplacer l'œil humain. C'est pour cela qu'il est primordial de calculer aussi la précision en prenant en compte les documents jugés pertinents par le système (P_S) . Pour ne prendre en compte qu'un seul paramètre, j'ai utilisé la métrique U tout en assignant des pondérations aux constantes λ1, λ2, λ3, λ4 et λ5. Cela donnera plus d'importance aux documents jugés pertinents par l'utilisateur tout en n'ignorant pas ceux jugés pertinents par le système, qui lui se verra donner une pondération moindre mais pas nulle.

Détermination des pondérations :

$\lambda 1 = 2$

$\lambda 2 = 1$

$\lambda 3 = -0.5$

$\lambda 4 = 0.5$

$\lambda 5 = -1$

Nous obtenons ainsi la métrique du système que j'ai mis en place :

$$U = 2*R1_{+} + 1*R2_{+} - 0.5*R3_{+} + 0.5*R4_{+} - 1*S_{+}.$$

Comme chaque profil a un nombre de remontées différent des autres, et pour une comparaison plus équitable avec la précision selon l'expertise humaine et celle faite par le système (car elles mêmes divisées par le nombre total des remontées), je divise également cette métrique par le nombre de remontées total (P_M).

Statistiques des remontées du système de filtrage mis en place

Pour tester le système, j'ai évalué son efficacité sur la période allant du 1er août au 30 aout 2012. Les équations considérées sont les suivantes :

P_H = ($R1_{+}$ + $R2_{+}$)/total des remontées

P_S= ($R1_{+}$ + $R1_{+}$ + $R4_{+}$)/total des remontées

P_M = U/total des remontées

Les résultats obtenus qui sont annexe 2 ont montré les graphiques suivants :

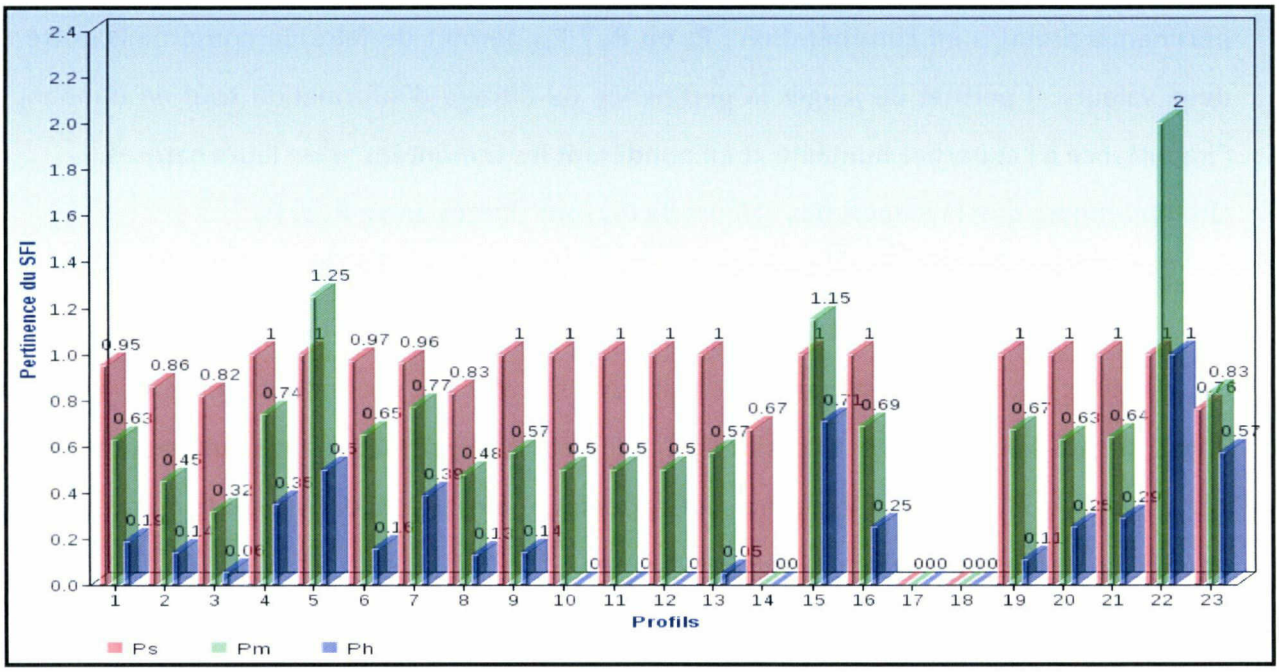


Figure 28 : corrélation entre P_M, P_H et P_M dans l'évaluation du SFI (profils de 1 à 23)

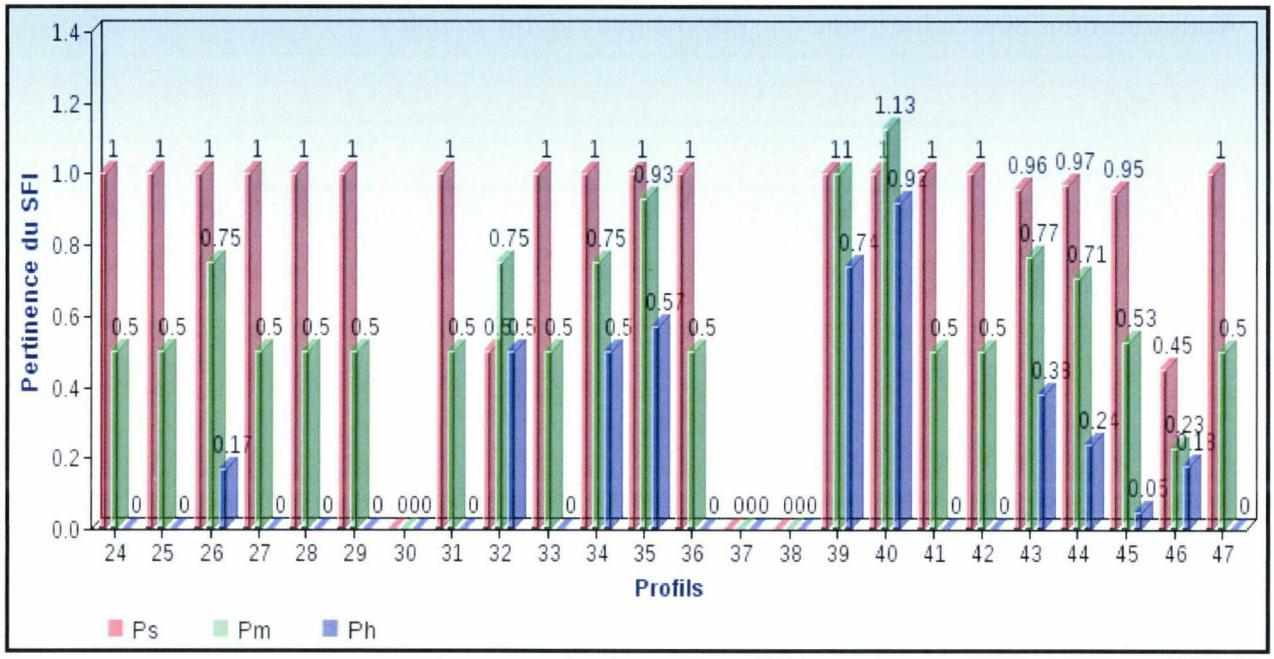


Figure 29 : corrélation entre P_M, P_H et P_M dans l'évaluation du SFI (profils de 24 à 47)

Les valeurs de $\mathbf{P_S}$ sont bien sur plus élevées car le système se base sur un filtrage sur mots-clés efficace. Mais toutes les informations ne sont pas pertinentes pour l'utilisateur, ce qui fait apparaître des valeurs de $\mathbf{P_H}$ moins élevées. La question qui se posait est la suivante : quelle pertinence prendre en considération : $\mathbf{P_S}$ ou $\mathbf{P_S}$? $\mathbf{P_M}$ permet de faire un compromis entre ces deux valeurs. Il permet de jauger la pertinence du filtrage d'information tout en donnant de l'importance à l'expertise humaine et en pondérant les remontées selon leurs natures.
On remarquera que la plupart des valeurs de $\mathbf{P_M}$ sont situées entre $\mathbf{P_S}$ et $\mathbf{P_H}$.

Ces résultats peuvent permettre de comparer les profils selon leurs pertinences et en fonction des résultats obtenus, il sera possible d'agir sur les profils qui présentent des pertinences sous un seuil minimal qui reste à déterminer. Si d'autres systèmes s'appuyant sur les mêmes paramètres de filtrage sur le contenu sont apportés, il sera possible d'utiliser la même métrique pour comparer les performances de ces systèmes.

Autre outil testé est KB Crawl. Comme déjà indiqué plus haut, cet outil n'a pas donné entière satisfaction. Pour l'évaluer, j'ai utilisé le ratio suivant.

Précision = nombre documents pertinents retrouvés / nombre documents retrouvés

Vu le nombre de sources qui n'est pas le même avec celui de Google Reader, je n'ai pas utilisé la même métrique que celle pour l'agrégateur de flux RSS. le but ici est de montrer ce que produit KB Crawl comme remontées.

Tout d'abord, il faut savoir que KB Crawl remonte les informations par paquets. Ainsi, chaque profil qui correspond à une remontée par alerte mail, contient une ou plusieurs remontées appelées remontées intra-mail. Ce sont ces dernières qui définissent le vrai nombre de remontées. Sur une période de 3 semaines, les résultats sont comme suit :

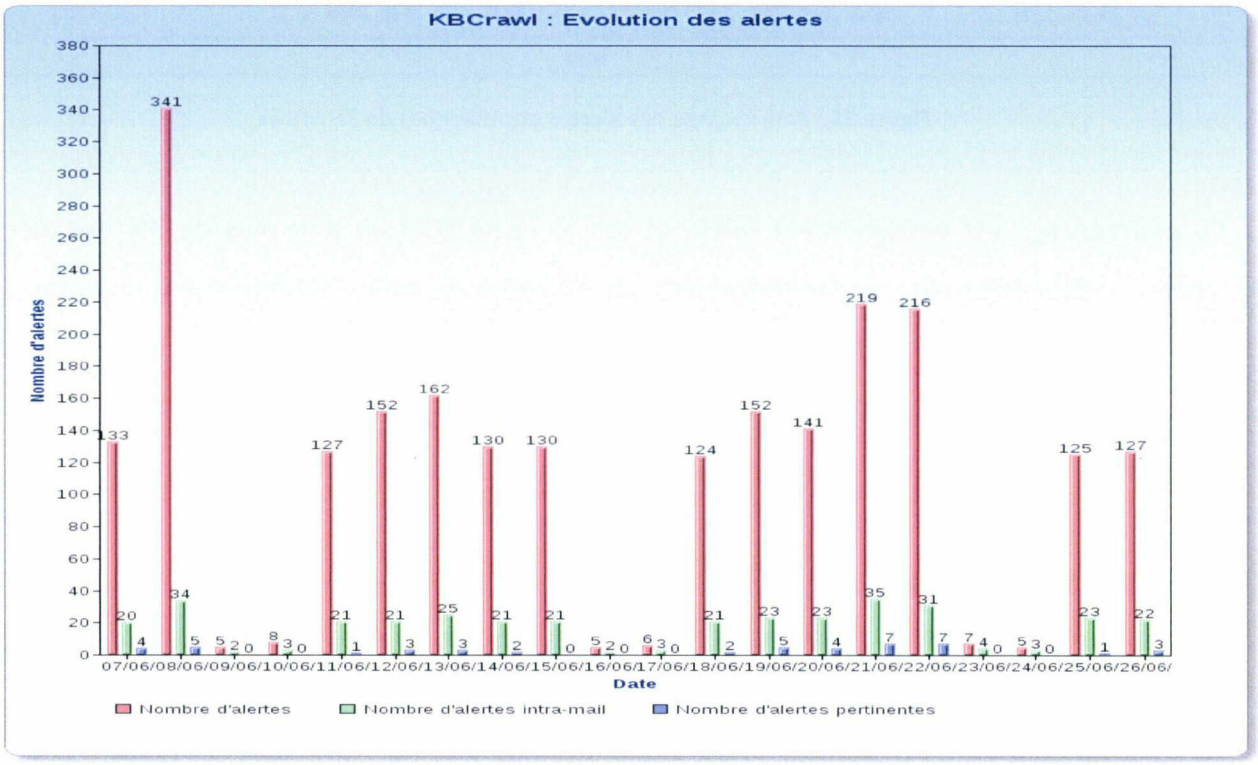


Figure 30 : nombre de remontées KB Crawl

On constate rapidement le nombre faramineux des remontées intra-mail. Elles sont en moyenne égales à 163 remontées par jour pour une moyenne des remontées pertinentes égale à 3.5 par jour. Cette faible fiabilité de l'outil remet en question son efficacité. Le pourcentage de ces alertes est illustré par le graphique suivant :

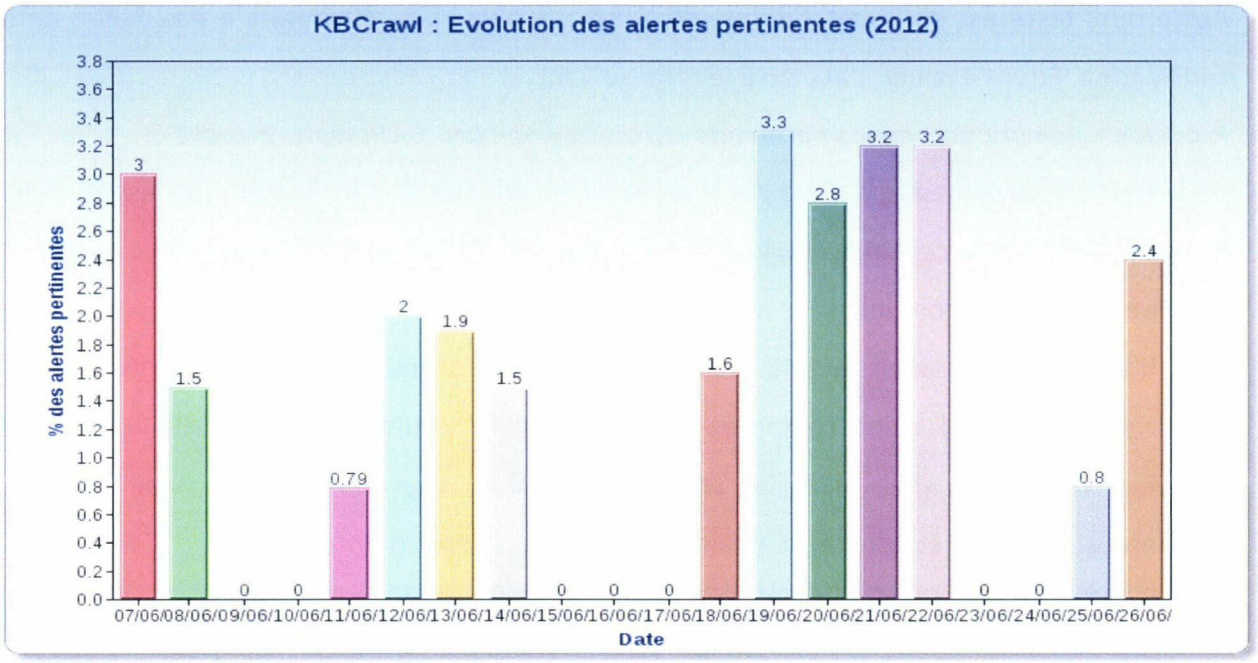


Figure 31 : Pourcentage des alertes quotidiennes de KB Crawl

Ce pourcentage est extrêmement faible et oscille entre 0 et un peu plus de 3% des alertes totales. Cela a bien sûr des conséquences sur l'activité de veille. Examiner en moyenne 163 informations par jour est lassant et prend beaucoup de temps. En passant en moyenne 1 minutes pour chaque remontées, l'utilisateur de l'outil passe près de deux heures à analyser ces remontées.

IV-2- Apports constatés

Par rapport aux attentes su-citées, il apparaît clairement que les apports attendus ont été clairement satisfaits.

Le pack de sources s'est enrichit et ce grâce aux flux RSS déjà présents ou crées par l'outil Feed43. Il permet ainsi de surveiller un périmètre plus large. Il peut être enrichi avec de nouvelles sources d'information avec une simplicité aisée. Il remonte également plus d'informations pertinentes avec des requêtes plus exhaustives couvrant beaucoup plus de produits et de filiales, ce qui diminue la probabilité de passer à côté d'une information utile ou stratégique ou annonciatrice d'un signal faible.

Le système a permis aussi de centraliser les flux d'informations. En effet, les collaborateurs peuvent se contenter de consulter Google Reader pour leur veille quotidienne, au lieu de

basculer entre les deux plates-formes déjà existantes : l'ancien Google Reader moins complet, et KB Crawl qui prenait beaucoup de temps pour des résultats médiocres.

L'unes des appréhensions de l'équipe du DVE est la possibilité que Google Reader remonte un flux d'informations trop important et ainsi faire déborder ses utilisateurs. Les statistiques sur les remontées montrent que la moyenne est de 23 remontées par jour pour tout le système, avec des valeurs pour chaque profil oscillant entre 0 et 5 remontées par jour. Cela reste largement gérable.

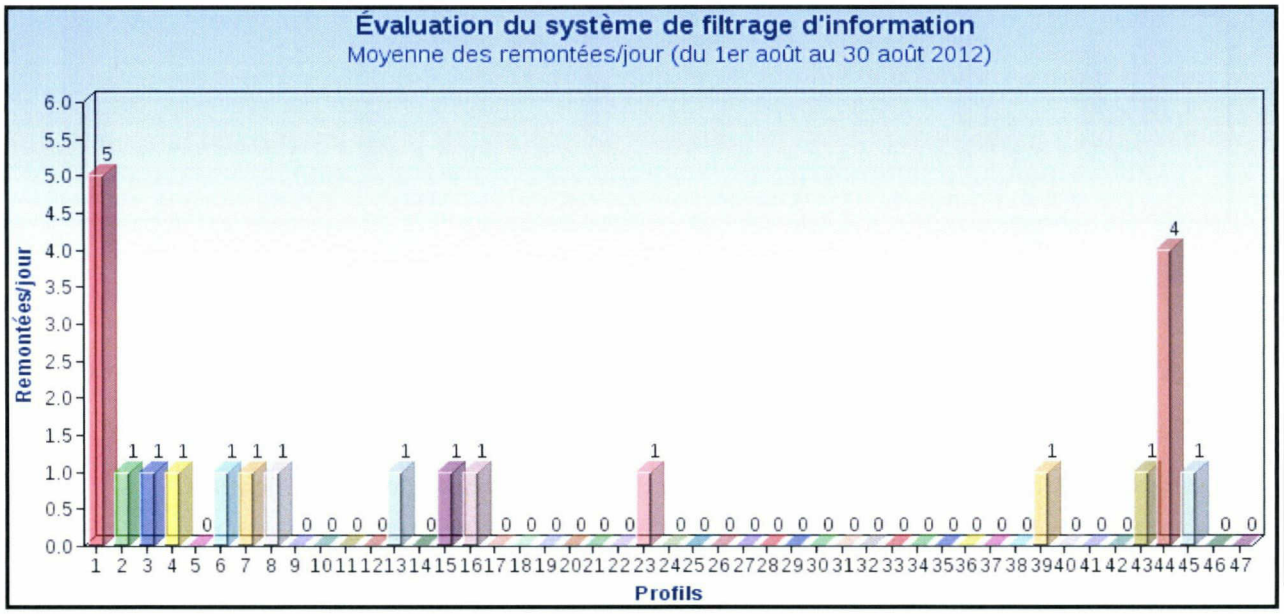


Figure 32 : moyenne des remontées quotidiennes du SFI

Cela est en grande partie dû aux requêtes précises et à la performance du système mis en place. Ce dernier a comme précédemment été indiqué, permis une précision des remontées grâce aux flux RSS et à l'élimination des doublons. Comme l'indique le graphique ci-dessous, ces derniers ont sensiblement diminué, même si on remarque comme même leur présence qui est due essentiellement à l'espacement entre les mises à jour des flux RSS qui font que Yahoo Pipes lors de sa phase d'analyse des anciens flux, ne détecte pas ces derniers et le considère comme nouveaux. Ce petit bug technique est heureusement très rare (1 doublon/jour en moyenne).

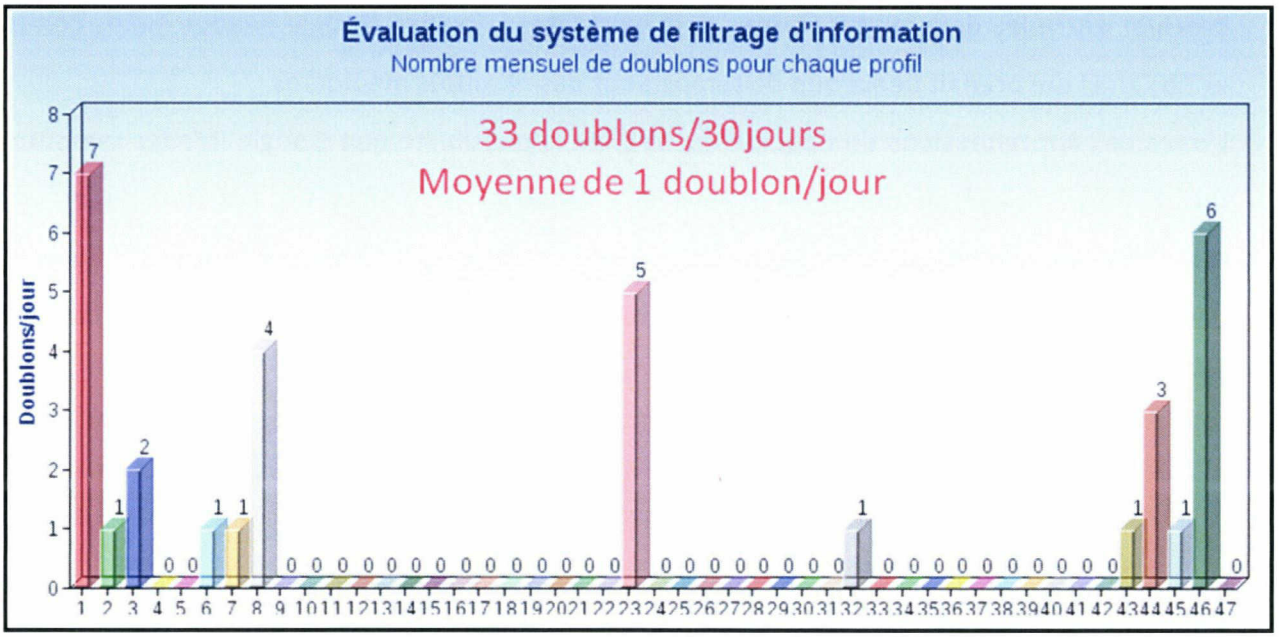


Figure 33 : Moyenne des doublons remontés par le SFI

Le DVE n'est plus dépendant des critères de remontée d'information de l'algorithme de Google. Il peut à tout moment inclure ou exclure une source donnée et appliquer un filtre par mots-clés sur les flux RSS pour toute source disposant de cette technologie de syndication.

La lecture des informations sur Google Reader est facilitée grâce aux nouveaux items rajoutés dans Yahoo Pipes où l'utilisateur peut visualiser la source et la date de publication exacte de l'article. Il peut ainsi jauger le degré d'importance de ce dernier sans même lire l'article et gagner ainsi du temps. D'autres extensions propres à Google Reader pour le navigateur Internet *Google Chrome* ont permis un visuel plus agréable et une consultation des articles à l'intérieur même de l'agrégateur.

De nouveaux profils ont été mis sous surveillance dans des dossiers ordonnés. Ils sont classés selon qu'ils concernent la concurrence directe ou indirecte, les informations du secteur du crédit et des produits et services en rapport aux activités de COFIDIS. Chaque profil contient plusieurs flux RSS centralisés en un seul grâce à leur compilation par Yahoo Pipes.

Enfin, l'optimisation de KB Crawl n'a pas donné des résultats satisfaisants. En effet, malgré un paramétrage plus poussé, il apparaît que l'outil ne permet pas de cibler la surveillance d'une manière précise. Cette optimisation a certes diminué le bruit mais pas suffisamment pour diminuer le temps passé à analyser ses alertes. A titre de comparaison, le temps passé sur Google Reader est d'environ 20 minutes contre deux heures sur KB Crawl. c'est un gain de

temps considérable sachant qu'en plus de cela, Google Reader contient plus de sources. Après une période de tests et de comparaisons avec le nouveau Google Reader, il a été décidé de ne se focaliser que sur les alertes de ce dernier, qui, avec l'outil de veille tarifaire, satisfont largement aux attentes du DVE.
Le responsable du DVE a ainsi décidé l'abandon de KB Crawl en mettant fin à leur collaboration.

IV-3- Bilan du stage

a- Bilan du travail effectué et usages et pratiques du dispositif de filtrage d'information au niveau du Département Veille et Etudes

Mettre en place un dispositif de filtrage d'information en contexte de veille a bien évidemment été du point de vue technique une bonne réussite. Cette réussite s'est remarquée sur le terrain via les résultats de l'évaluation et surtout par la satisfaction des utilisateurs de ce dispositif.
Le point de vue des utilisateurs tout au long de la mise en place, a grandement contribué à l'optimisation du dispositif de veille. J'ai accordé beaucoup d'importance à leurs demandes, à leurs exigences et à leurs remarques.
Le bilan objectif a donc tiré ses bases sur des critères eux aussi objectifs. Les résultats statistiques ont montré que les bruits ont été réduits au maximum, que les informations remontées étaient précises et ciblées et que le dispositif a permis de centraliser les informations tout en ayant un nombre de remontées acceptable et gérable. Un gain de temps et de productivité considérables ont été remarqués en se passant de KB Crawl. Cela permettait selon les déclarations des utilisateurs de se consacrer à d'autres tâches.
Consulter le dispositif est devenu une étape obligatoire et essentielle pour être tenu informé des informations relatives à l'environnement de COFIDIS. Le DVE l'utilise comme l'une des principales sources d'information. Il permet de mener une veille concurrentielle, technologique, d'innovation et de marché. Il est notamment utilisé lors des demandes Ad' Hoc pour consulter la liste de suivi et extraire les informations nécessaires à leurs réalisations. Il permet également de créer de nouvelles thématiques à surveiller.

b- Bilan personnel

Le stage m'a permis de me consacrer à une veille plutôt technique, demandant des compétences en informatiques telles les expressions régulières, les langages de programmation web et le langage Perl. C'était une opportunité de mettre en pratiques ces connaissances acquises lors de l'année universitaire et de voir en pratique leur efficacité et les résultats qu'ils pouvaient produire.

Cette expérience est donc une réussite au niveau personnel où j'ai dans mon bagage une compétence non négligeable sur la mise en place d'une veille gratuite et efficace. En ces temps de crise, cela peut s'avérer utile.

Au delà des compétences techniques, j'ai également côtoyé l'environnement de COFIDIS, qui est une grande entreprise. Cela m'a permis de découvrir le monde du crédit à la consommation et de la façon dont sont menées les stratégies de marketing, de communication et des opérations commerciales. Je me suis aperçu du rôle important joué par le relationnel humain en entreprise, du travail en équipe et des concertations continues pour une meilleure productivité.

Ce stage constitue donc pour moi une expérience professionnelle enrichissante dans le monde de la finance. Ce qui pourrait faciliter mon intégration rapide dans le monde professionnel.

Conclusion

Les techniques développées dans ce mémoire de stage contribuent à améliorer les remontées d'information au sein du Département Veille et Etudes de COFIDIS. Ces travaux s'inscrivent dans une finalité d'utilisation d'outils gratuits mais pouvant accomplir des tâches avec des performances égales ou même meilleures que celles proposées par certains outils payants. Grâce à la manipulation des flux RSS, il est possible de mettre en œuvre une surveillance presque complète sur tout type de support digital. J'ai pu via le dispositif mis en place, utiliser ce moyen d'abonnement à l'information, à remplacer KB Crawl qui est un outil de surveillance des changements survenant sur les sites Internet. Yahoo Pipes qui est l'outil principal ayant permis cette prouesse offre une mine de possibilités pour manier les flux RSS et les rendre plus performants. L'un de ses modules les plus utilisés est *Filter.* Il évite à l'utilisateur de traiter beaucoup d'information en réduisant au maximum les bruits grâce aux requêtes appliquées qui sont constituées de mots-clés. Ainsi, le filtrage d'information sur le contenu a été rendu possible. Avec ce dispositif, il est possible de cibler les parties d'un site web à surveiller même si ce dernier ne présente pas des données structurées ou semi-structurées. Cette problématique est la base même de la création des flux RSS. Ainsi, ces derniers, qui sont des documents XML, structurent l'information dans des items répétitifs.

Le côté technique qui a pris toute sa place lors de ce stage, a notamment été au cœur de l'utilisation des expressions régulières. Dans Yahoo Pipes, ces dernières ont grandement été utiles particulièrement lors des phases de « *rechercher-remplacer* ». Elles évitent ainsi d'utiliser trop d'opérations complexes.

Lors du déploiement du dispositif, l'analyse de l'existant, des besoins et des attentes du DVE ont contribué à mener des réflexions approfondies pour trouver la meilleure solution. Cela a été l'une des raisons qui m'ont mis sur la piste de création de flux RSS pour les sites qui n'en proposaient pas. Ainsi, grâce à l'outil Feed43, j'ai pu créer des flux RSS personnalisés car je pouvais extraire les informations voulues sans pour autant subir des informations non désirées. Cet outil est tellement puissant que j'ai pu remplacer le module *macro* de KB Crawl. Ce dernier pour rappel permet de saisir lui même la requête dans un champ d'un moteur de recherche. J'ai ainsi pu mettre en place un filtre sur url puisque les requêtes saisies sont reprises dans l'url des

résultats. Malheureusement Feed43 ne gère pas encore les scripts puisque ceux-ci ne font pas changer les adresses url.

Le but de ce stage était certes d'optimiser le dispositif, mais certaines pratiques et usages devaient être préservés. C'est pour ces raisons que Google Reader qui est l'agrégateur des flux entrants a été gardé, sachant qu'il est l'un des plus utilisés et l'un des plus performants grâce à ses différentes fonctionnalités.

Cette approche d'intégrer trois outils gratuits pour centraliser les informations a donné des résultats satisfaisants au niveau de la qualité et de la pertinence des remontées, et le DVE a été formé pour assurer sa maintenance et ses mise à jour.

Ce que je tire de cette expérience, est que le filtrage d'information prend au fur et à mesure que le web avance, une importance grandiose. Les entreprises recherchent de plus en plus la bonne information au bon moment, et avec la quantité de données circulant sur le web, il est devenu inimaginable de se passer des outils de filtrage d'information. Reste que l'évaluation de leurs performances reste tributaire des systèmes mis en place. Il devient ainsi difficile de mettre en place une métrique unique pouvant évaluer n'importe quel dispositif de filtrage d'information. Il faut sans cesse adapter cette évaluation à chaque système.

Liste des abréviations :

DTD : Document Type Definition

DVE : Département Veille et Etudes

FI : Filtrage d'Information

HTML : Hyper Text Markup Language

REGEX : Expressions régulières

RI : Recherche d'Information

RSS : Really Simple Syndication

SaaS : Software as a Service

SFI : Système de Filtrage d'Information

SRI : Système de Recherche d'Information

URL : Uniform Resource Locato

XML : Extensible Markup Language

ANNEXES

Annexe 1 : Liste des expressions régulières les plus utilisées

Expression régulière	Signification
.	N'importe quel caractère sauf le saut de ligne
*	Représente zéro ou plusieurs occurrences de l'expression la précédant
+	Représente une ou plusieurs occurrences de l'expression la précédant
^	Trouve un terme se trouvant uniquement au début d'un paragraphe
$	Trouve un terme se trouvant uniquement à la fin d'un paragraphe
\n	Représente le saut de ligne
\s	Représente un espace (ou séparateur)
\S	Représente tout sauf un espace
\t	Représente une tabulation
\>	Trouve la chaîne recherchée uniquement si celle-ci figure à la fin d'un mot
\<	Trouve la chaîne recherchée uniquement si celle-ci figure au début d'un mot
\d	Trouve les chiffres
\D	Trouve tout caractère sauf les chiffres
\w	Trouve tous les caractères alphanumériques (chiffres + lettre de l'alphabet)
\W	Trouve tous les caractères sauf les alphanumériques
|	Représente l'expression « OR »
\	Cette barre d'échappement permet de représenter les caractères figurant déjà dans les expressions régulières comme \ ou ^
i	Permet d'ignorer la casse
{}	Permet de préciser le nombre d'occurrences exact

Annexe 2 : résultats des l'évaluation du SFI entre le 1er août et le 30 août 2012

Profils	Total	R1+	R2+	R3+	R4+	S5+	Nombre remontées/jour	U	$P_{M\ (U/Total)}$	P_S	P_H
Profil 1	148	12	16	7	113	0	5	93	0,63	0,95	0,19
Profil 2	29	2	2	1	21	3	1	13	0,45	0,86	0,14
Profil 3	17	0	1	2	13	1	1	5,5	0,32	0,82	0,06
Profil 4	17	1	5	0	11	0	1	12,5	0,74	1,00	0,35
Profil 5	2	1	0	0	1	0	0	2,5	1,25	1,00	0,50
Profil 6	31	3	2	1	25	0	1	20	0,65	0,97	0,16
Profil 7	28	3	8	1	16	0	1	21,5	0,77	0,96	0,39
Profil 8	23	2	1	4	16	0	1	11	0,48	0,83	0,13
Profil 9	14	0	2	0	12	0	0	8	0,57	1,00	0,14
Profil 10	13	0	0	0	13	0	0	6,5	0,50	1,00	0,00
Profil 11	6	0	0	0	6	0	0	3	0,50	1,00	0,00
Profil 12	5	0	0	0	5	0	0	2,5	0,50	1,00	0,00
Profil 13	22	1	0	0	21	0	1	12,5	0,57	1,00	0,05
Profil 14	3	0	0	0	2	1	0	0	0,00	0,67	0,00
Profil 15	17	5	7	0	5	0	1	19,5	1,15	1,00	0,71
Profil 16	16	1	3	0	12	0	1	11	0,69	1,00	0,25
Profil 17	0	0	0	0	0	0	0	0	0,00	0,00	0,00
Profil 18	0	0	0	0	0	0	0	0	0,00	0,00	0,00
Profil 19	9	1	0	0	8	0	0	6	0,67	1,00	0,11
Profil 20	8	0	2	0	6	0	0	5	0,63	1,00	0,25
Profil 21	7	0	2	0	5	0	0	4,5	0,64	1,00	0,29
Profil 22	1	1	0	0	0	0	0	2	2,00	1,00	1,00
Profil 23	21	6	6	5	4	0	1	17,5	0,83	0,76	0,57
Profil 24	2	0	0	0	2	0	0	1	0,50	1,00	0,00
Profil 25	1	0	0	0	1	0	0	0,5	0,50	1,00	0,00
Profil 26	12	2	0	0	10	0	0	9	0,75	1,00	0,17
Profil 27	3	0	0	0	3	0	0	1,5	0,50	1,00	0,00
Profil 28	2	0	0	0	2	0	0	1	0,50	1,00	0,00
Profil 29	2	0	0	0	2	0	0	1	0,50	1,00	0,00
Profil 30	0	0	0	0	0	0	0	0	0,00	0,00	0,00
Profil 31	3	0	0	0	3	0	0	1,5	0,50	1,00	0,00
Profil 32	2	1	0	1	0	0	0	1,5	0,75	0,50	0,50
Profil 33	3	0	0	0	3	0	0	1,5	0,50	1,00	0,00
Profil 34	2	0	1	0	1	0	0	1,5	0,75	1,00	0,50
Profil 35	7	1	3	0	3	0	0	6,5	0,93	1,00	0,57
Profil 36	2	0	0	0	2	0	0	1	0,50	1,00	0,00
Profil 37	0	0	0	0	0	0	0	0	0,00	0,00	0,00
Profil 38	0	0	0	0	0	0	0	0	0,00	0,00	0,00
Profil 39	23	3	14	0	6	0	1	23	1,00	1,00	0,74
Profil 40	12	2	9	0	1	0	0	13,5	1,13	1,00	0,92
Profil 41	1	0	0	0	1	0	0	0,5	0,50	1,00	0,00
Profil 42	1	0	0	0	1	0	0	0,5	0,50	1,00	0,00
Profil 43	24	3	6	1	14	0	1	18,5	0,77	0,96	0,38
Profil 44	119	14	15	3	87	0	4	85	0,71	0,97	0,24
Profil 45	20	1	0	1	18	0	1	10,5	0,53	0,95	0,05
Profil 46	11	2	0	6	3	0	0	2,5	0,23	0,45	0,18
Profil 47	5	0	0	0	5	0	0	2,5	0,50	1,00	0,00
TOTAL	694	68	105	33	483	5	23		0,59	0,84	0,20

BIBLIOGRAPHIE

1. Nicholas J. Belkin et W. Bruce Croft (Dec 1992). Information filtering and information retrieval: Two sides of the same coin? Communication of the ACM. v35 n12 p29(10).
2. http://benhur.teluq.uqam.ca/SPIP/inf6460/article.php3?id_article=17&id_rubrique=4&sem=2#nb6
3. BOUGHANEM et al. (Avril 1999). Query modification based on relevance backpropagation in Adhoc environment, Information Processing and Managment. Vol 35, pages 121-139, Elsevier Science.
4. Foltz, P. W. et Dumais, S. T. (1992). Personalized information delivery: An analysis of information filtering methods. Communications of the ACM, 35(12), 51-60.
5. P.J.Denning (1982). Electronic Junk. Communication of the ACM, Vol 25, N°3.
6. Malone, T. et al. (1987). Intelligent information sharing systems. Communications of the ACM, 30(5) :390–402
7. Goldberg et al (1992). Using collaborative filtering to weave an information tapestry. ACM SIGIR Forum, 12(35) :61–70.
8. Tebri H. (2004). Formalisation et spécification d'un système de filtrage incrémental d'information. Thèse.
9. BALABANOVIC M., SHOHAM Y. (Mars 1997). «Fab: content-based collaborative recommendation» *Communications of the ACM*, vol. 40, n° 3, p. 66-72.
10. Zaïr Z. (2003). Modèle multi-agent pour le filtrage collaboratif de l'information. Projet de recherche. Université du Québec Montréal.
11. Harman, D. (1992). The darpa tipster project. ACM SIGIR Forum, 2(26) :26–28.
12. www.cofidis.fr (site du Groupe Cofidis)
13. www.cofidis.com (site du Groupe Cofidis Participations)
14. AFNOR: norme [XP X 50-053 Avril 1998]
15. http://atlas.irit.fr
16. www.digimind.com
17. http://intelligenceco.over-blog.com

18. http://intelligenceco.over-blog.com

19. www.definitions-marketing.com

20. International Journal of Infos & Com Sciences for Decision Marketing. 1er trimestre 2003. ISSN : 1265-499X

21. EUGEN-COSTIN POPOVICI (Janvier 2008). Information Retrieval of Text, Structure and Sequential Data inHeterogeneous XML Document Collections. Thèse.

22. http://www.liafa.jussieu.fr/~carton/Enseignement/XML/Cours/support.html#id3944058

23. http://www.xul.fr/xml-rss.html#what

24. http://www.eolya.fr

25. http://intelligences-connectees.fr

26. http://bibliotheques.wordpress.com

27. Lenormand P. (2007). Internet : Techniques de recherche pour les professionnels. Editions ENI.

28. Antonio Balvet (2002). Approches catégoriques et non catégoriques en linguistique des corpus spécialisés. Thèse, Paris X Nanterre.