Newssift, un nouvel outil de détection et de traitement des entitées nommées

Retour au sommaire de la lettre

Domaine :	**Recherche**	Référencement
Niveau :	**Pour tous**	Avancé

Dans la lignée de notre article du mois précédent sur iSeek, voici un nouvel outil permettant de traiter les entités nommées (noms de personnes, d'organismes, de lieux, etc.) dans les contenus du Web. Lancé par le Financial Times il y a peu, NewsSift permet d'explorer de nombreuses sources d'information et, par exemple, de mettre en place une veille sur l'e-réputation d'une entreprise. Exploration...

Nous évoquions le mois dernier l'outil iSeek (http://www.iseek.com/), un moteur de recherche généraliste qui innovait en détectant les entités nommées et en permettant de les croiser entre elles afin d'en faire émerger des résultats pertinents. Il n'aura pas fallu longtemps pour qu'un second moteur du même type, tout aussi intéressant en terme de fonctionnalités et d'ergonomie, apparaisse. A croire que l'idée était dans l'air...

Ce nouvel arrivant s'appelle Newssift (http://www.newssift.com/). Il a été lancé le mois dernier par le Financial Times et permet, comme son nom l'indique, de rechercher dans l'actualité. Il ne se contente toutefois pas des seules informations proposées par le FT (déjà conséquentes), mais indexe, d'après ses concepteurs, plusieurs millions d'articles en provenance de sources d'actualités "business" internationales.

A l'instar d'iSeek, Newssift traite cette actualité grâce à un filtre sémantique qui lui permet de détecter des entités nommées telles que :
- les organisations ;
- les lieux ;
- les personnes ;
- les thèmes ;
- le vocabulaire "business" (ex : "*executive structure*", "*board of directors*", "*workforce employement*",...)

Toutefois le traitement automatique n'est pas tout, puisque ce moteur est également paramétré par des consultants experts et des membres de l'équipe éditoriale qui aident, semble t-il, à détecter, évaluer et valider les sources à prendre en compte.

Concrètement, vous lancez une recherche dans Newssift comme dans n'importe quel autre moteur. C'est ensuite que la différence se fait...

Nous utiliserons pour ce test le mot-clé "peugeot" dont on est sûr qu'il apparaît dans l'actualité récente... Dés le moment où l'on commence à taper un mot, l'interface fait des propositions "*as-you-type*" de mots et concepts associés à ajouter à votre recherche. Yahoo fait aussi cela très bien mais dans le cas de Newssift les mots-clés apparaissent déjà classés par thèmes :

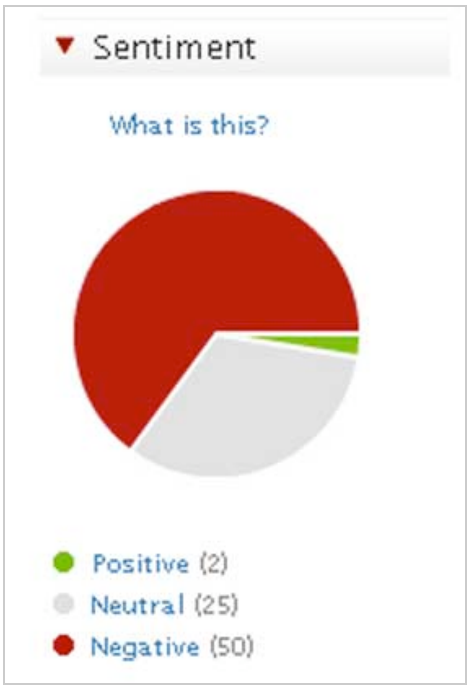
Vous pouvez donc lancer directement la recherche sur votre mot-clé initial ou bien la modifier en cours de frappe. La page qui apparaît alors propose, pour cette requête, 1 509 résultats ainsi que les entités nommées associés à notre mot-clé. Comme avec iSeek, l'intérêt est ici que l'on est certain d'obtenir des résultats en croisant ces termes entre eux puisqu'apparaissent uniquement ceux qui sont liés aux nôtres dans le texte des articles traités. Nous allons ainsi pouvoir affiner notre requête au fur et à mesure de nos sélections successives, jusqu'à cibler les quelques articles répondant à l'ensemble de nos critères.

Dans notre exemple nous allons maintenant chercher à savoir comment Peugeot est associé à la Slovaquie puisqu'une usine de production y est installée. Ce lieu géographique n'apparaissant pas dans la fenêtre "*Places*", nous allons cliquer sur "*More*" pour faire apparaître plus de choix, puis sur "*Slovakia*".

Nous n'avons plus alors que 82 résultats classés par pertinence, que nous allons classer par dates pour plus de commodité.

Intéressons-nous maintenant à cette page de résultats. Deux représentations distinguent Newssift d'iSeek, il s'agit de graphiques en "camembert". Le premier est intitulé "*Sentiment*" et le second "*Article Sources*".

Le graphique des "*Sentiments*" permet de se faire une idée de la tonalité des articles présents dans les résultats.



Dans notre cas, 65% des articles sont considérés par le système comme négatifs et 2% comme positifs, le reste étant neutre. Il faut évidemment être prudent avec ces interprétations généralement réalisées par des outils de traitement sémantique. S'ils identifient bien les phrases à connotation positive ou négative, ils sont en revanche totalement imperméables au cynisme, au second degré, aux euphémismes et plus globalement à tout ce qu'un rédacteur peut vouloir cacher derrière des mots.

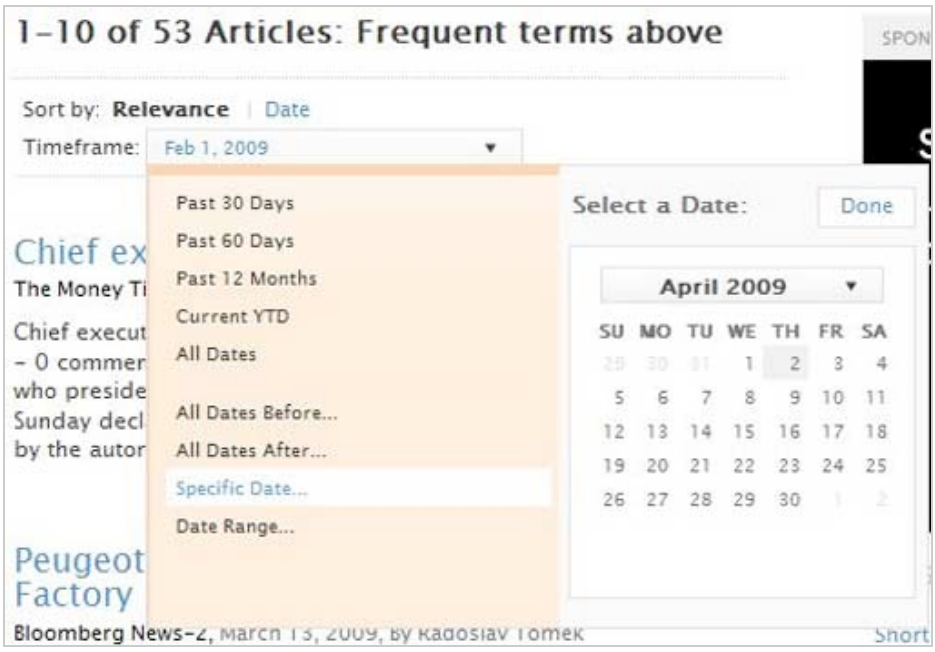
Malgré une marge d'erreur certaine, cet outil peut toutefois s'avérer utile à un premier niveau de filtrage. Il s'agit d'un effort notable pour traiter l'information issue du web de manière

qualitative, effort qui devrait aller croissant à l'avenir et donner des résultats toujours plus pointus. Du fait de l'augmentation quasi exponentielle des sources potentielles de données, phénomène en partie lié au web social (Facebook, LinkedIn, Twitter,...), il va en effet devenir de plus en plus complexe pour une organisation de mener une veille "image". Il sera donc nécessaire de déléguer ce travail à des algorithmes et à des machines.

Pour en revenir à ce graphique en "camembert", il faut noter qu'il présente l'intérêt d'être cliquable. En cliquant sur la part rouge, on pourra donc ajouter un filtre à notre requête et la limiter ainsi aux seuls articles à tonalité négative. Le nombre de résultats tombe alors à 53.

Nous allons maintenant pouvoir utiliser le second graphique de la même manière afin d'affiner votre requête en fonction du type de sources qui nous intéresse. On peut ainsi choisir de n'afficher que les articles du Financial Times ou encore ceux tirés d'autres journaux, de la radio/télévision, de magazines en ligne, de portails d'actualité,...

Mais ce n'est pas tout. Une fonctionnalité baptisée *Timeframe* va en effet nous permettre de sélectionner les articles en précisant une période donnée ou une durée.



En limitant, dans notre exemple, cette durée à un mois, on obtient donc au final quatre articles traitant de Peugeot et de la Slovaquie, classés dans l'ordre antéchronologique, et ayant une tonalité négative. Il ne reste plus qu'à les lire ou, pourquoi pas, à les affiner encore avec, par exemple, le nom d'organisation "*Moody's Corporation*" ou celui de son nouveau PDG, *Christian Varin*.

Si les résultats de la requête définie par les différentes facettes que nous avons ajoutées les unes aux autres ne nous satisfont toujours pas, nous pouvons évidemment en supprimer certaines et en ajouter de nouvelles. Nous pouvons aussi voir si un concept proche de celui présent dans une facette ne serait pas plus adapté encore. Il suffit pour cela de positionner le curseur sur les deux petites flèches rouges lorsqu'elles apparaissent. Un double menu déroulant apparaît alors, vous proposant des concepts qui vous permettront soit d'étendre ("*Expand*") la requête grâce à un mot-clé qui "l'englobe", soit de la préciser d'un mot-clé qui la restreint ("*Refine*").

Notre exemple est ici très éclairant, puisque nous obtenons pour la facette « Slovakia » les mots Europe ("*Expand*") et Bratislavsky, une déclinaison du nom de la capitale, Bratislava ("*Refine*").

Notez enfin qu'il est possible de "forcer" un mot-clé en l'ajoutant *via* la barre de recherche et que Newssift permet d'enregistrer des requêtes ("*Saved search*") et conserve par défaut un historique des recherches (qu'il est possible de supprimer).

La détection d'entités nommées n'est pas un concept nouveau, surtout pour les moteurs de recherche d'actualité (cf. Silobreaker, Newsbrief, Newstin ou encore Evri), toutefois la manière dont Newssift la met en œuvre est innovante et s'avèrera utile pour qui veut affiner ses recherches sur un thème déjà connu ou, au contraire, découvrir de nouvelles pistes de travail.

Christophe Deschamps
Consultant et formateur en gestion de l'information.
Responsable du blog Outils Froids (http://www.outilsfroids.net/)

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2009/04/newssift-un-nouvel-outil-de-detection.html