TextMap, un service de news basé sur l'extraction d'entités nommées

Retour au sommaire de la lettre

TextMap est un nouveau moteur de recherche qui exploite la notion d'entités nommées (noms de personnes, de lieux, d'organismes, etc.) citées dans un texte. Déjà utilisée par Exalead et d'autres sites, TextMap, dont la devise est "Monitorer le monde pour que vous n'ayez pas à le faire", exploite ce concept pour nous aider à mieux chercher une information ciblée au travers de nombreuses fonctionnalités connexes et innovantes...

Les entités nommées désignent l'ensemble des noms de personnes, de lieux, d'organismes ou d'institutions contenus dans un texte ainsi que les dates et autres données chiffrées.
Les identifier et les extraire pour multiplier les points d'entrée vers un même texte, mais aussi additionner les occurences pour en faire émerger des tendances peut donc s'avérer particulièrement utile en notre époque d' "infobésité" galopante. Exalead l'utilise pour présenter les résultats des recherches lancées dans la Wikipedia (http://www.exalead.fr/wikipedia/homepage) et ClearForest en a tiré un plugin Firefox permettant l'analyse en temps réel du texte des pages web que vous visitez (http://sws.clearforest.com/Blog/?page_id=32/).

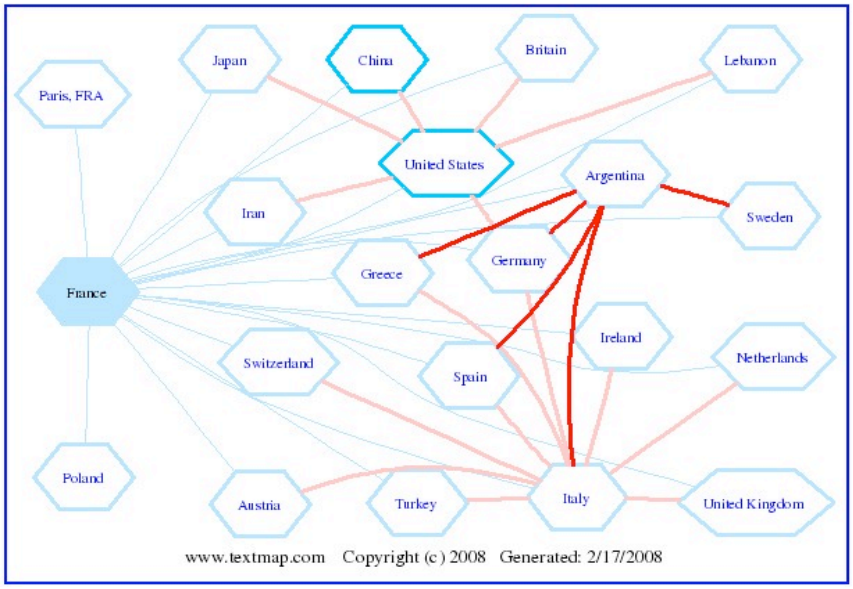
C'est sur ce principe qu'a été développé le service de recherche d'actualités **TextMap** (http://www.textmap.com/) et ses dérivés : TextMed, pour l'information médicale, TextBlg pour les blogs et TextBiz pour l'actualité économique.

La page d'accueil de TextMap (dont la devise est "*Monitorer le monde pour que vous n'ayez pas à le faire*", tout un programme...) présente par défaut les entités les plus visibles du moment classées dans huit catégories : personnes, villes, pays, compagnie, université, médicament, site web et entités citées dans les titres de l'actualité. Si vous ne trouvez pas celle que vous cherchez, il suffit de cliquer sur "*More*" pour obtenir une liste plus complète (TextMap en reconnaît plus d'un million !). Autre possibilité, utiliser la barre de recherche pour lancer votre requête. Pour information TextMap ne travaille que sur l'actualité provenant de sources US (plusieurs centaines de journaux en ligne). C'est donc un parfait point de départ pour chercher à comprendre le point de vue américain sur l'information internationale. Pour notre exemple, nous avons d'ailleurs choisi d'examiner le traitement reservé à l'entité "France".

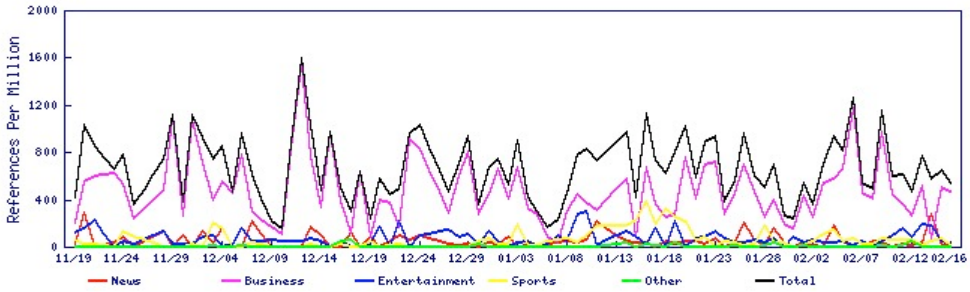
Une fois cliqué sur ce mot-clé vous arrivez sur une page dynamique qui lui est consacrée et vous propose plusieurs modes de traitements graphiques et statistiques du corpus d'actualités le concernant :

- ***Articles referencing France*** : le premier et le plus simple puisqu'il s'agit de la liste des articles dans lesquels le mot "France" est présent.

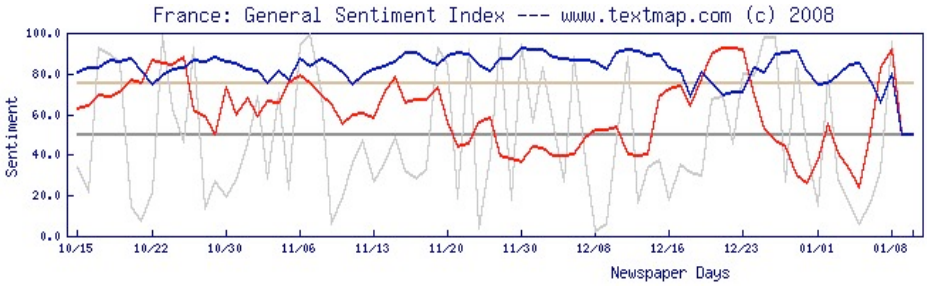
- ***Relational Network*** : permet de visualiser les entités de même nature (pays) qui sont le plus associées à l'entité-cible dans l'actualité du jour (les couleurs des lignes indiquent la force des associations)



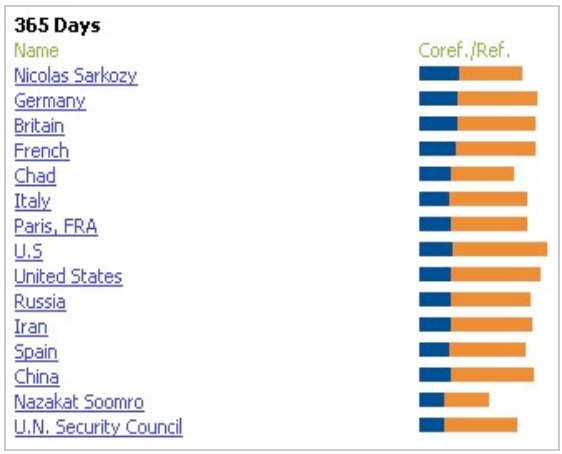
- *Popularity time serie* : il s'agit de deux graphiques permettant de suivre les thématiques dans lesquelles l'entité est présente dans le temps (News, Business, Entertainment, Sports,...) afin de voir où elle apparaît le plus souvent.



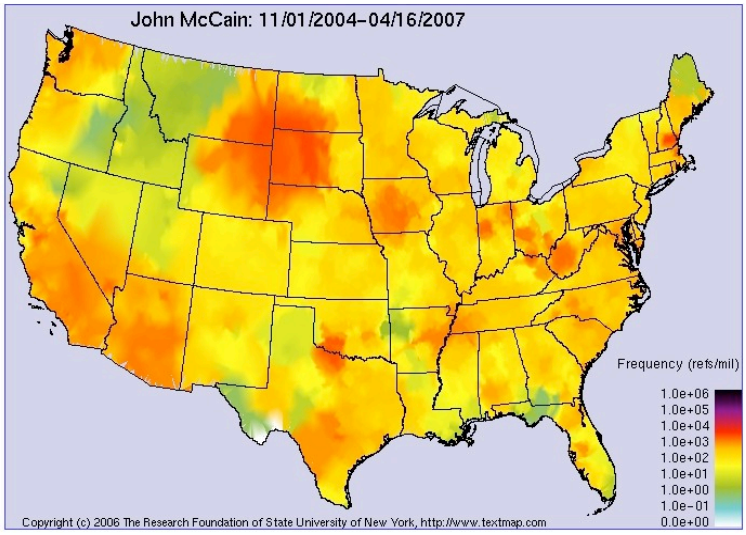
- *Sentiment analysis* : ce graphique permet de suivre le regard positif ou négatif porté sur l'entité dans l'actualité à travers les semaines. La courbe rouge mesure cette côte d'amour tandis que la bleue indique si elle laisse ou non indifférente. Il serait intéressant de savoir plus précisement comment sont calculés ces indicateurs, on sait en effet que le traitement statistique n'est pas forcément approprié au traitement qualitatif d'un corpus textuel. Quoiqu'il en soit on voit bien l'intérêt que cela peut avoir si l'on est chargé d'effectuer la veille image d'un homme politique ou d'une organisation.



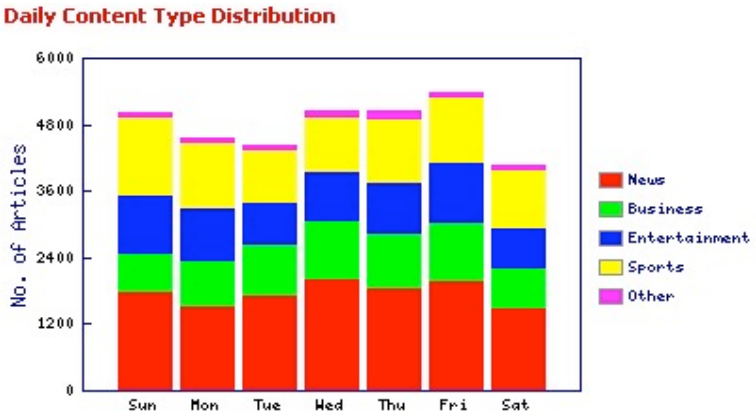
- *Juxtapositions* : cette fonctionnalité présente les entités les plus souvent associées à l'entité-cible (co-occurrences) tous types confondus. La taille des barres colorées (voir ci-dessous) reflète la popularité de l'entité (orange) et la force de l'association avec la cible (bleu). Cliquer sur une barre vous amène sur la liste des articles associant les entités. Les juxtapositions sont exprimées sur 30 jours, un an ou depuis le début de l'indexation des sources par TextMap, c'est-à-dire 2004.



- *Heatmap* : il s'agit d'une carte des Etats-Unis qui indique quelles zones du pays (et donc quels journaux régionaux) citent l'entité-cible (calcul de fréquence) depuis 2004. Afin que l'exemple soit plus significatif nous abandonnons un instant l'entité "France" pour l'entité "John McCain" :



Il faut noter que chacune des sources utilisées fait elle-même l'objet d'un traitement statistique et graphique permettant de savoir quelles entités y sont le plus souvent citées et quels types de contenus elle distribue. Cela revient finalement à générer un profil dynamique de chaque source utilisée :



Contenu du New-York Daily News (http://www.nydailynews.com/)

En revenant sur la page d'accueil on peut trouver en bas à droite d'autres fonctionnalités intéressantes qui redirigent vers des pages dynamiques spécifiques telles que :

- Daily sentiment report : page présentant la "côte d'amour" des entités marquantes de l'actualité et permet d'en voir émerger de nouvelles.
- Daily sentiment map : même chose que ci-dessus mais projeté sur une carte des Etats-Unis
- Daily heatmap report : page présentant une "heatmap" (voir ci-dessus) pour chaque entités populaire

TextMap, on l'a vu, est donc beaucoup plus qu'un simple service d'agrégation d'actualités, il s'agit d'un véritable outil de text-mining en ligne qui, une fois maîtrisé, permet de surveiller l'information de manière extrêmement pointue. Il n'est d'ailleurs pas sans nous rappeller Silobreaker, un autre excellent service d'actualité (*voir l'article à son sujet dans la Lettre Recherche et Référencement de novembre 2007*) ou encore **Inform** (http://inform.com/) désormais privé. Plus globalement de tels services sont la démonstration même de ce que le traitement statistique de l'information textuelle est en mesure d'apporter à tous ceux pour qui elle est la matière première.

Précision : TextMap est un service gratuit développé par le "Computer Science Department" de l'Université de Stony Brook (Etat de New York). Plus d'infos : http://www.cs.stonybrook.edu/

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/2008/02/textmap-un-service-de-news-bas-sur.html