Veille d'Information sur le Web avec Re-Watch

Christophe Brouard*, Christian Pomot**

*Université Grenoble Alpes, LIG UMR 5217/Equipe AMA, France
Christophe.Brouard@imag.fr,
http://ama.liglab.fr/ brouard/
**Société Com&Net, 155 Cours Berriat 38028 Grenoble Cedex 1
cpomot@com-et-net.com
http://com-et-net.com

Résumé. Les algorithmes d'apprentissage automatique peuvent être utilisés pour créer des outils de recommandation qui permettent de prédire la pertinence d'un document pour une thématique de veille donnée en se basant sur les précédents jugements de pertinence donnés pour cette thématique pour d'autres documents. Ces outils de recommandation permettent de filtrer dans un flux entrant de documents ceux qui sont susceptibles d'être pertinents sans que l'utilisateur ait besoin de déterminer lui-même les mots clefs marquant l'adéquation d'un document pour un sujet de la veille. Bien que cette problématique de recherche ait été abondamment abordée, les outils de veille d'information pour le web intégrant un apprentissage en sont encore à leur balbutiements. Nous présentons ici l'application web Re-Watch permettant la définition d'un thème de veille, la sélection de sources d'information sur le web relatives à ce thème et l'adaptation des scores de pertinence des documents aux retours de l'utilisateur. L'application permet aussi, pour chaque thème, une auto-évaluation de la qualité du filtrage et une interrogation du moteur de recherche Google. Cette application encore en cours de développement est néanmoins actuellement fonctionnelle et accessible sur le web à l'url suivante : http://www.specific search.com.

1 Introduction

La veille d'information peut-être définie selon Cacaly et al. (2008) comme « un processus continu et dynamique faisant l'objet d'une mise à disposition personnalisée et périodique de données ou d'informations, traitées selon une finalité propre au destinataire, faisant appel à une expertise en rapport avec le sujet ou la nature de l'information collectée ». Plusieurs aspects importants ressortent de cette définition. D'une part, le processus est continu, c'est-à-dire qu'il sétend sur une certaine durée (typiquement plusieurs semaines, mois ou années). Or les moteurs de recherche traditionnels comme Google, Bing, Yahoo, pour citer les plus utilisés, sont conçus pour des besoins d'information ponctuels. Les fonctions de sauvegarde des informations pertinentes et des sources associées y sont notamment très rudimentaires et la récupération des résultats des précédentes sessions de recherche n'est pas aisée. D'autre part, le processus est dynamique, des informations apparaissent et disparaissent tous les jours. Or

les moteurs de recherche traditionnels sont des outils dits « PULL ». Cela signifie que l'utilisateur devra inlassablement se connecter aux différents outils de recherche pour se tenir informé. Enfin, l'information est pertinente relativement à une finalité propre au destinataire qui ne correspond pas nécessairement totalement à une thématique parfaitement identifiée et associée à des mots clefs évidents. Or dans les moteurs de recherche traditionnels, tout le travail de formulation de la requête consistant à trouver les bons mots clés par reformulations successives en fonction des résultats retournés par le moteur de recherche n'est pas facilement capitalisé. Le développement d'outils dédiés à la veille d'information répondant à ces différents besoins, en mesure d'apprendre les préférences de l'utilisateur et accessibles à tous comme le sont les moteurs de recherche en est encore à ses prémices (Katakis et al., 2009), (Nanas et al., 2010).

2 Présentation générale de l'outil

L'interface de l'outil ressemble à celle d'un webmail. A gauche, on trouve les différents thèmes de veille (appelés recherches). La sélection d'une recherche déclenche l'affichage dans la partie centrale des nouveaux résultats, des résultats stockés et du paramétrage pour la recherche sélectionnée. La sélection du titre d'une nouvelle dans la liste des résultats déclenche l'affichage de son contenu dans la partie basse de l'interface (voir la figure 1).

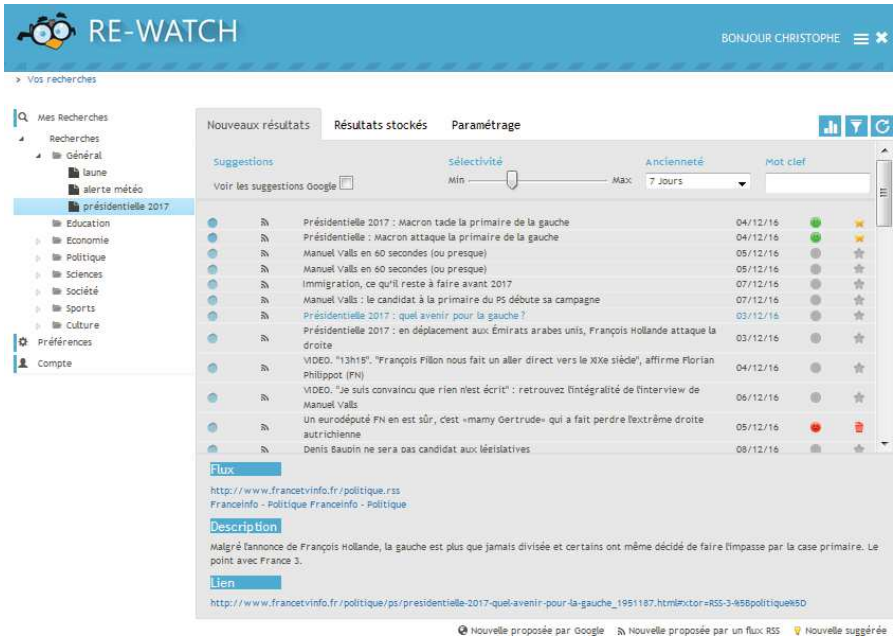

FIG. 1 – *Interface de visualisation des nouveaux résultats avec Re-Watch. L'utilisateur peut filtrer les nouvelles sur le score, le niveau d'ancienneté ou la présence de mots-clefs*

Le mode opératoire pour l'utilisation de l'outil est le suivant : l'utilisateur commence par créer une recherche en tapant quelques mots clefs et en sélectionnant parmi les sources pro-

posées (des flux rss) celles qui lui semblent les plus pertinentes. Il indique ensuite, parmi les nouvelles informations recueillies par l'outil, celles qui selon lui, sont pertinentes et celles qui ne le sont pas. L'outil utilise cette information pour recalculer le score de pertinence des différentes nouvelles et affiche les nouvelles par ordre de pertinence.

L'utilisateur peut ensuite choisir de sauvegarder/supprimer différentes nouvelles indépendamment de ses jugements de pertinence. Les nouvelles peuvent être filtrées en indiquant une probabilité de pertinence seuil, en précisant un degré d'ancienneté (limité à un mois) et en saisissant une chaîne à trouver dans la nouvelle. A tout moment, l'utilisateur peut ajouter/supprimer des sources d'informations. L'utilisateur peut aussi demander à tout moment de voir les résultats issus d'une recherche Google et s'appuyant sur ses retours de pertinence. Enfin, il peut afficher, pour la recherche sélectionnée, le niveau de qualité de filtrage atteint après la prise en compte de ses retours de pertinence.

En plus du serveur web et du serveur de base de données présents dans toute application web, Re-Watch intègre deux autres serveurs en charge de l'apprentissage automatique. Cet apprentissage est basé sur l'algorithme Echo (Brouard, 2012). Enfin, les sources d'information surveillées ont la forme de flux RSS (tout site web peut potentiellement être transformé en un flux RSS) et des scripts exécutés régulièrement et automatiquement récupèrent les nouvelles des différents flux et mettent à jour un index.

3 Interfaçage avec l'algorithme Echo

L'application Re-Watch repose sur l'algorithme Echo (Brouard, 2012). Cet algorithme peut être appliqué à différents types de problèmes de sélection d'information, comme la recherche d'information (sélection d'un document pour une requête), la classification supervisée (sélection d'une classe pour un document) ou encore l'extension de requête (sélection d'un terme pour un ensemble de termes). Echo peut être décrit comme un système de construction et d'exploitation de réseau associatif s'appuyant sur des mécanismes neuronaux simples. La construction du réseau repose sur la règle de Hebb renforçant la connexion entre deux informations survenant simultanément (deux termes présents dans le même document par exemple), son exploitation repose sur une méthode de propagation d'activation et la mesure d'une quantité d'activation rétro-propagée vers les sources d'activation (correspondant à la notion d'écho). Il est basé sur une formalisation de la notion de pertinence qui combine les notions de spécificité et d'exhaustivité qui sont au cœur des modèles de pertinence en recherche d'information (Brouard et Nie, 2004).

L'application Re-Watch s'appuie sur deux serveurs intégrant l'algorithme Echo. Un premier serveur construit un index des sources d'information sur la base de toutes les nouvelles contenues dans la base de données (actuellement plusieurs centaines de milliers et potentiellement plusieurs millions) et le met à jour régulièrement et incrémentalement avec les nouvelles informations recueillies. Cet index permet de proposer à l'utilisateur pour un ensemble de mots-clefs saisis, les sources d'informations susceptibles de l'intéresser. Un second serveur permet, sur la base des jugements de pertinence donnés pour les nouvelles d'un thème par l'utilisateur, de calculer des scores et probabilités de pertinence du thème pour les nouvelles auxquelles aucun jugement de pertinence n'a été associé. Il permet aussi, à la demande de l'utilisateur, une évaluation de la capacité de l'algorithme à séparer les documents pertinents des non pertinents pour le thème. Il permet enfin de déterminer les meilleurs termes associés

au thème, de faire une requête au moteur de recherche Google, et de calculer les scores de pertinence des résultats du moteur pour ces termes en tenant compte des retours de pertinence.

4 Conclusion

Une interface homme-machine intégrant des fonctionnalités permettant de faciliter l'activité de veille sur le web a été proposée. La réalisation d'expérimentations en conditions réelles avec différents utilisateurs véritablement engagés dans un processus de veille semble incontournable. Re-Watch est actuellement en bêta-test et nous comptons sur des retours nombreux et pertinents pour l'améliorer. Des évolutions sont en cours. Certaines concerneront notamment la mise en place de fonctionnalités relatives à l'étape de diffusion d'information facilitant le partage des résultats de la veille.

Références

Brouard, C. (2012). Document classification by computing an echo in a very simple neural network. In *IEEE International Conference on Tools with Artificial Intelligence*, pp. 735–741.

Brouard, C. et J.-Y. Nie (2004). Relevance as resonance: a new theoretical perspective and a practical utilization in information filtering. *Information Processing and Management 40*, 1–19.

Cacaly, S., Y.-F. LeCoadic, P.-D. Pomart, et E. Sutter (2008). *Dictionnaire de l'information*. Paris : A. Colin.

Katakis, I., G. Tsoumakas, E. Banos, N. Bassiliades, et I. Vlahavas (2009). An adaptive personalized news dissemination system. *Journal of Intelligent Information Systems 32(2)*, 191–212.

Nanas, N., V. Manolis, et H. Elias (2010). Personalised news and scientific literature aggregation. *Information Processing and Management 46*, 268–283.

Summary

Machine learning algorithms can be used to build recommendation tools which allow to predict document relevance for a particular topic considering previous relevance judgements given for other documents for the same topic. Although this research domain has been often studied, tools for monitoring information on the web are very rare. Here, we present the web application called Re-Watch allowing topic definition, information source selection related to the topic and relevance score adaptation to user relevance feebacks. The application provides also, for each topic, an auto-evaluation of the filter quality and an adapted query to the Google search engine. The development of the application is still ongoing however it is currently available on the web at the following url : http://www.specific search.com.