Le filtrage automatique de l'information multilingue : une évaluation inspirée des vérités terrain

CHAUDIRON Stéphane [*]
TIMIMI Ismaïl [*]
BESANCON Romaric [**]
MOSTEFA Djamel [+]
LAIB Meriama [**]
CHOUKRI Khalid [+]

[*]Université de Lille 3 – GERiiCO
[**]CEA LIST
[+]ELDA

Résumé :
Le projet InFile (INformation, FILtrage, Evaluation) vise à organiser une campagne d'évaluation de logiciels adaptatifs de filtrage interlingue d'information et dès lors à mener une réflexion sur la question de l'évaluation de technologies linguistiques. Il est soutenu par l'Agence Nationale de la Recherche (ANR) et est conduit par le CEA-LIST, ELDA et l'Université de Lille 3 (laboratoire GERiiCO). Il est une campagne pilote de CLEF 2008. La collection de test est constituée d'un corpus de 1,4 million de dépêches (environ 10 Go) fourni par l'Agence France-Presse (AFP) en trois langues, l'arabe, l'anglais et le français sur une période de 3 ans. Le corpus de requêtes est constitué de 30 profils généraux (événements nationaux et internationaux d'ordre géo-politique, socio-économique, culturel, sportif…) et de 20 profils concernant des sujets scientifiques et techniques écrits dans un style d'information et de vulgarisation.

1. Introduction

La campagne d'évaluation InFile vise à mesurer la capacité de systèmes de filtrage d'information à filtrer de manière pertinente des dépêches de presse arrivant dans un flux ininterrompu d'information textuelle, en fonction de profils d'intérêt. Selon Belkin et Croft (Belkin, 1992), un système de filtrage d'information est un logiciel conçu pour gérer des données non-structurées ou semi-structurées. Les systèmes de filtrage d'information s'intéresse essentiellement à l'information textuelle et suppose l'existence de flux importants d'information tels que des fils de dépêches de presse. Le filtrage s'effectue sur la base de profils d'information individuels ou collectifs supposés représenter des besoins informationnels valides sur le long terme. Du point de vue de l'usager, le processus de filtrage signifie extraire l'information jugée pertinente des flux de données en fonction des profils qui ont été précédemment définis.

Les systèmes de filtrage d'information sont utilisés dans différents contextes professionnels. Le routage de textes implique d'envoyer des informations entrantes vers des individus ou des groupes d'individus (diffusion sélective), la classification de textes regroupe des documents entrants dans des sets homogènes pré-définis (catégorisation) ou post-définis (clusterisation), les techniques d'anti-spam visent à éliminer des fichiers jugés dangereux ou indésirés du courrier électronique, le repérage de thèmes émergeants (signe d'alerte précoce)

Dans le projet *InFile*, nous considérons le contexte de la veille informationnelle dans laquelle le filtrage d'information est une tâche spécifique de la gestion de l'information (Bouthillier, 2003). Dans notre approche, la tâche de filtrage est semblable à la DSI (Diffusion sélective de l'information), une des fonctions traditionnelles assurées par les documentalistes ou, plus

récemment, par d'autres intermédiaires de l'information comme les veilleurs ou les spécialistes de l'intelligence économique. Pour la définition du protocole d'évaluation, le projet accordera une attention particulière au contexte d'usage des systèmes de filtrage par des vrais professionnels. Même si la campagne d'évaluation reste essentiellement une campagne orientée système, nous adapterons autant que possible le protocole et les métriques à la manière dont un véritable usager procèderait, y compris en ce qui concerne l'interaction avec le système.

Des campagnes d'évaluation ont déjà été organisées dans le passé pour des systèmes de filtrage adaptatifs, en particulier dans le cadre de TREC (*Text REtrieval Conferences*) entre 2000 et 2002 (Robertson, 2002) et dans les campagnes TDT (*Topic Detection and Tracking*) entre 1998 et 2004 (Fiscus, 2004). Les caractéristiques de la campagne *InFile* par rapport aux précédentes campagnes sont présentées dans les sections suivantes.

2. Objectifs et caractéristiques de la campagne

Le projet vise trois objectifs : le premier et le plus important est d'organiser une campagne d'évaluation afin de comparer les performances des logiciels développés par l'industrie et le monde académique sur la tâche de filtrage interlingue et de définir ainsi l'état de l'art actuel. Le deuxième objectif est de définir un protocole d'évaluation fondé sur les pratiques de filtrage réelles des veilleurs et des spécialistes de l'information. En ce sens, une caractéristique forte du projet est de tenir compte, tout au long du projet, de la « vérité terrain ». Le troisième objectif est de construire une collection de test comprenant un corpus de dépêches, un corpus de profils et un référentiel formé des documents pertinents pour chacun des profils. A ces ressources textuelles, nous ajoutons des outils informatiques pour le formatage des données, le calcul de métriques, la visualisation des

résultats… Ce « package d'évaluation » sera disponible à des fins de recherche.
La campagne *InFile* se présente essentiellement comme la suite de la tâche de filtrage adaptatif de TREC 11, avec une attention particulière pour le protocole qui s'attachera à représenter au mieux la vérité terrain des professionnels de la veille informationnelle.

Dans les campagnes TDT, l'attention était principalement portée sur les « topics » définis comme des évènements, avec un niveau fin de granularité, souvent très circonscrits dans le temps, alors que dans *InFile* (de même que dans TREC 11), les « topics » couvrent des besoins informationnels stables dans le temps.

Les principales caractéristiques de la campagne InFile sont les suivantes :

- interlingue : l'anglais, le français et l'arabe sont pris en compte même si les participants peuvent être évalués sur des « runs » monolingues ou bilingues ;
- un corpus de dépêches de presse couvrant trois années récentes est fourni par l'AFP (Agence France-Presse) ;
- le corpus de profils (aussi appelés « topics ») est composé de deux sous-ensembles : un sous-ensemble constitué de sujets d'information générale et un sous-ensemble constitué de sujets scientifiques et techniques ;
- la tâche d'évaluation est assurée en interrogation automatique des systèmes participants et en simulant le *feedback* de l'usager ;
- les systèmes peuvent utiliser le *feedback* pour améliorer leur performance (une des métriques est conçue pour le calcul des degrés d'adaptation) ;
- une décision booléenne de pertinence est attribuée à chaque document en fonction de chaque profil ;

- les jugements de pertinence sont essentiellement fournis par des assesseurs humains (référentiels) ;
- à la fin de la campagne, il est demandé aux participants de remplir un formulaire pour préciser les langues utilisées par le système, les champs des profils (voir ci-dessous) ainsi que la technologie du système.

3. Les collections de test

3.1 Le corpus AFP

Le corpus *InFile* est fourni par l'Agence France-Presse uniquement à des fins de recherche. L'AFP est la plus ancienne agence de presse du monde et l'une des trois plus importantes avec *Associated Press* et *Reuters*. Bien que l'AFP soit la plus importante agence de presse française, elle diffuse des informations dans d'autres langues, comme l'anglais, l'arabe, l'espagnol, l'allemand et le portugais.

Pour *InFile*, nous avons choisi trois langues (l'arabe, l'anglais et le français) sur une période de trois ans (2004-2006), ce qui correspond à environ 1,4 million de dépêches dans les trois langues (10 Go). Les dépêches sont fournies dans les trois langues mais ne sont pas nécessairement des traductions d'une langue vers l'autre. Le volume des dépêches par année et par langue est indiqué dans la figure 1:

	2004	**2005**	**2006**	***Total***
ARA	85k	81k	87k	254k
FRE	154k	139k	154k	448k
ENG	268k	245k	244k	758k
Total	508k	467k	486k	**1 462k**

Figure 1 Statistiques du corpus AFP

Pour la campagne, seulement 100 000 documents dans chaque langue sont utilisés pour les tests de filtrage compte tenu des craintes de surcharges cognitives imposées aux assesseurs et des contraintes du temps dues au processus de filtrage interactif décrit dans la section 4.1. Ces documents correspondent au référentiel humain composé des documents pertinents correspondant aux profils, complété par un ensemble de documents considérés non pertinents par un outil d'extraction.
Les dépêches sont codées au format XML et suivent les spécifications NewsML (*News Markup Language*)[1]. NewML est un standard XML conçu pour fournir un cadre de description d'informations multimédia indépendant des médias. NewsML a été développé par l'IPTC (*International Press Telecommunications Council*) qui est un consortium réunissant les principales agences de presse.

3.2 La collection de profils

Une collection de 50 profils répartis en deux catégories a été préparée. Un premier ensemble de 30 dépêches correspond au domaine de l'information générale et des évènements nationaux et internationaux (sports, politique, économie…) et un deuxième ensemble de 20 dépêches correspond au domaine des sciences et techniques. Pour s'approcher au plus près de la vérité terrain, les profils ont été réalisés par des professionnels de la veille travaillant à l'INIST[2] (*Institut de l'information scientifique et technique*), à l'ARIST Nord Pas-de-Calais[3] (*Agence régionale d'information stratégique et technologique*), à Digiport (*un centre de services expert en TIC*)[4], à l'ONERA (*Office national*

[1] http://www.aristnpdc.org/
[2] http://international.inist.fr/
[3] http://www.aristnpdc.org/
[4] http://www.digiport.org

d'études et recherches aérospatiales)[5] et pour la société OTO Research[6].
Les profils ont été conçus selon la structure suivante, réellement utilisée dans le domaine de la veille :

- un identifiant unique (pour identifier le profileur),
- un titre (6 mots max.),
- une description sommaire (20 mots max.),
- un texte d'explication (60 mots max.),
- un maximum de 5 mots-clés,
- un exemple de texte pertinent (120 mots max.).

3.3 La traduction de profils

Pour faciliter la rédaction des profils, ceux-ci ont été rédigés en français ou en anglais puis traduits, à l'exception du champ <Exemple>, dans les deux autres langues (français ou anglais et arabe). Nous avons recommandé que la traduction vers l'anglais (ou vers le français si le profil a été rédigé directement en anglais) soit effectuée par le rédacteur du profil lui-même. Cela permettrait de conserver de manière optimale la sémantique du texte saisi dans le profil initial et de respecter la terminologie en vigueur dans les différents domaines et langues.
Pour des raisons évidentes, la traduction vers l'arabe a été effectuée autrement. Elle a été accordée à des traducteurs arabophones mais habitués des activités de veille et de recherche d'information multilingue. Vu l'importance de l'exercice, nous avons privilégié des traducteurs arabophones ayant une bonne maîtrise des deux langues sources (français et anglais), cependant à défaut d'un référentiel terminologique arabe couvrant les termes scientifiques et techniques émergeants, des difficultés de traduction de certains concepts ont vite apparu. Pour se rapprocher de la vérité terrain, nous avons fait appel à plusieurs traducteurs universitaires et

[5] http://www.onera.fr
[6] http://www.otoresearch.fr/

professionnels exerçant dans trois pays du monde arabe, de traditions linguistiques différentes (Liban, Egypte et Maroc) ; la pratique des trois langues dans ces pays étant assez répandue.
Face à des concepts peu connus ou des termes non encore officialisés, certains traducteurs ont eu recours à des outils et techniques variés (usage de certains glossaires peu spécialisés, étude de la représentation des termes sur le web, comparaison des traductions disponibles au sein de l'encyclopédie Wikipédia…).
Pour réduire l'impact de la différence terminologique entre différentes sources, on a décidé enfin de conserver dans le descriptif du profil le terme le plus répandu sur le web et de lui annexer entre parenthèses ses termes synonymes utilisés dans d'autres contextes ou sources.
Ci-dessous un extrait des cas (de profils) où la traduction vers l'arabe n'était pas simple :

- des sigles peu utilisés en arabe : *DGN, CIO*
- des termes parfois traduits, parfois transcrits : *technologies* (تكنولوجيات - تقنيات)
- des concepts à usage local : *travail à temps partiel*
- des concepts émergeants : *m-commerce* (التجارة الإلكترونية - الجوّالة - الخلوية)
- des concepts inconnus : *cosmétofoods*

Malgré ces embarras de traduction vers la langue arabe, il est à noter que seule dans cette langue où on trouve chacun des profils assorti d'un exemple extrait du web.

3.4 La collection de documents pertinents

La collection de documents pertinents est construite en deux phases, une phase de jugement avant le début de l'évaluation (phase de pré-soumission) et une phase de jugement après l'évaluation (phase de post-soumission). Pour coller à la vérité terrain, et parce que le *feedback*

doit être envoyé immédiatement après chaque soumission de résultat, l'utilisation de la « pooling method »[7] stricte n'est pas possible. En conséquence, l'évaluation est faite par rapport à une collection de documents pertinents construite par des experts humains. Néanmoins, à la fin de l'évaluation, un contrôle limité utilisant la « pooling method » est appliqué aux documents jugés pertinents par au moins deux systèmes, pour chaque profil. Ce qui permet d'enrichir de manière automatique le référentiel humain.

Dans la phase de pré-soumission, les concepteurs des profils ainsi que d'autres assesseurs ont utilisé différents systèmes de recherche d'information (SRI) pour fournir les jugements de pertinence. Dans la phase de post-soumission, de nouveaux jugements de pertinence ont complété les précédents afin d'ajuster la collection de documents pertinents, d'améliorer la qualité du feedback fourni aux participants et de vérifier les différentes mesures (de performance, d'évolutivité, d'anticipation, d'originalité…).

Enfin, pour éviter le biais méthodologique dans la construction du référentiel, dû généralement à la subjectivité des assesseurs, au degré de tolérance et à leur connaissance du sujet… nous avons procédé au calcul des taux de corrélation entre les différents assesseurs participant à la campagne. Quelques métriques du calcul de la variabilité ont été introduites (test de khi2, indice de kappa), les résultats obtenus nous ont rassurés de la convergence des décisions des assesseurs.

[7] Dans cette approche, un document est considéré pertinent quand il l'est notifié par la majorité des outils participants ou à l'unanimité.

4. Description du protocole

4.1 Le processus dôévaluation

Le protocole de la campagne InFile est conçu afin d'être une tâche réaliste pour un logiciel de filtrage. En particulier, l'idée est d'éviter que l'intégralité du corpus soit disponible pour les participants avant la campagne et donc de fournir les documents un à un (flux entrant), simulant le comportement d'un fil de presse. Le protocole force donc les systèmes à être évalués document par document.
Le protocole est interactif et se décompose de la manière suivante :

- le système se connecte au serveur d'où il obtient un identifiant pour le *run* en question. Si un participant veut soumettre plusieurs runs, le système doit se connecter plusieurs fois pour obtenir plusieurs identifiants ;
- le système reçoit un document ;
- le système filtre le document, c'est-à-dire l'associe à un ou plusieurs profils, ou le rejette ;
- pour les systèmes adaptatifs, un *feedback* de pertinence peut être fourni pour les documents filtrés ;
- une fois qu'un document a été filtré, le système reçoit un nouveau document et le processus recommence.

Un feedback de pertinence est simulé pour les systèmes adaptatifs. L'idée est à nouveau de simuler le comportement d'un veilleur professionnel. Dans un processus de veille réel, le professionnel reçoit en effet les documents jugés pertinents par le système, lit le document et décide ensuite de le conserver ou de le jeter s'il s'agit d'une erreur de filtrage. Dans *InFile*, il s'agit du seul *feedback* autorisé : le *feedback* de pertinence ne peut être demandé par un système que pour un document

qui a été associé à un profil ; il n'y a pas de *feedback* pour les documents écartés.
De plus, nous faisons l'hypothèse que les professionnels n'ont pas une patience infinie, le *feedback* n'est fourni que pour 50 documents, nombre qui a été conseillé par les professionnels participant au projet. Ce choix peut éventuellement avantager les systèmes qui s'adaptent rapidement par rapport à ceux qui ont besoin de nombreux documents pour s'entraîner mais il a semblé judicieux de placer les systèmes dans un contexte réaliste.
Une architecture client-serveur a été mise en place pour gérer la communication entre les systèmes participants et le serveur de dépêches *InFile*. Celui-ci utilise le port http à travers une architecture de *Web Service* afin de pouvoir gérer les problèmes posés par les *firewalls* des participants.
Pour vérifier le bon fonctionnement du protocole et de l'architecture, un test à blanc a été organisé avec deux profils et 50 documents à filtrer. Ces profils et documents ont été mis à disposition des participants peu de temps avant le début de la campagne réelle afin de leur permettre d'adapter leurs systèmes au format des profils et des documents.

4.2 Les métriques

Les résultats fournis par les participants dépendant d'un jugement binaire[8] sur la base de l'association d'un document avec un profil. Pour un profil donné, les résultats peuvent donc être résumés sous la forme d'une table de contingence du type:

[8] Pour des raisons de commodité (contraintes techniques), nous nous sommes contentés de « jugement binaire » aussi bien dans les réponses des systèmes que dans l'élaboration de référentiels, même si dans la pratique courante de veille, les usagers ont des comportements différents et leur décision est souvent « graduée ».

	Doc. du référentiel	Doc. hors du référentiel
Doc. jugés Pertinents	a	b
Doc. jugés Non pertinents	c	d

Pour chaque profil, deux nombres sont connus en amant :
- *(a+c) : nbr total des documents pertinents (effectif du référentiel)*
- *(a+b) : nbr total des documents reçus (cardinal du flux envoyé).*

Il suffit alors de connaître la valeur de a (nombre de doc. jugés correctement pertinents par le système) pour en tirer les trois métriques classiques qui suivent.
A partir de cette table, un ensemble de mesures classiques sont calculés pour chaque profil:

- Précision, définie par $P = a / (a+b)$
- Rappel, défini par $R = a / (a+c)$
- F-mesure, une métrique médiane qui tient compte simultanément des deux métriques précitées, c'est une combinaison de la précision et du rappel pondérés avec un paramètre α (Van Rijsbergen, 1979) ; $F_\alpha = [(1+ \alpha) * PR] / (\alpha * P + R)$[9].

Pour calculer une moyenne, les valeurs sont d'abord calculées pour chaque profil. Une autre façon de calculer une moyenne serait de d'additionner les valeurs pour tous les profils en fonction de la table de contingence et de calculer les scores qui en résultent. La première méthode est préférée parce qu'elle permet d'équilibrer l'impact de chaque profil, dont on suppose que les différences sont la source principale de variation dans le calcul des mesures.
Afin de mesurer l'*adaptativité* (évolutivité) des systèmes, les mesures sont également calculées à différents moments du processus de filtrage (tous les 10 000

[9] Quand α=1, la même importance est donnée à la précision et au rappel et la F-mesure est la moyenne harmonique des deux valeurs.

documents), et une courbe d'évolution des différentes valeurs dans le temps est fournie.
De plus, deux mesures expérimentales sont utilisées. La première est la mesure d'*originalité* qui est définie par la capacité d'un système à être le seul à sélectionner des documents pertinents. Cette mesure donne plus d'importance aux systèmes qui utilisent des technologies innovantes permettant de retrouver des documents « difficiles ».
La seconde est une mesure d'*anticipation*, conçue pour identifier les systèmes qui sélectionnent le premier document correspondant à un profil donné. Cette mesure est particulièrement justifiée dans le domaine de la veille par l'intérêt à retrouver le plus vite possible les premières informations du secteur qui est surveillé (alerte précoce). Ce score est mesuré par le rang inverse des premiers documents pertinents retrouvés (dans une liste de documents) sur la moyenne de l'ensemble des profils. Cette mesure est semblable à la mesure MMR (*mean reciprocal rank*) utilisée pour l'évaluation des systèmes de Question-Réponses (Voorhees, 1999), bien qu'elle ne soit pas calculée sur la liste ordonnée des documents retrouvés mais sur la liste chronologique des documents pertinents.

5. Conclusion

Au moment d'écrire cet article, la campagne *InFile* n'est pas achevée et nous ne pouvons donc pas présenter les résultats de l'évaluation des systèmes. Néanmoins, des résultats significatifs ont déjà été obtenus, en particulier la construction d'une collection de test composée d'un corpus trilingue structuré (français, anglais, arabe) de dépêches d'agence de presse, un corpus trilingue structuré de profils validés par des experts de la veille et un ensemble de documents pertinents correspondant à ces profils.
D'autres travaux à l'issue de l'organisation de la campagne sont en cours : étude de métriques nouvelles

employées dans le projet, observation de pratiques chez les assesseurs…
Les deux autres objectifs sont la campagne d'évaluation proprement dite (qui sera lancée en juillet 2008) et la modélisation de la pratique de veille par les professionnels, qui est un objectif à long terme dépassant le cadre du projet *InFile*.

Bibliographie

Belkin N., Croft B. (1992). Information filtering and information retrieval : two sides of the same coin. In *Communications of the ACM*, vol. 35, n°12, pp. 29-38.
Bouthillier F., Shearer K. (2003). *Assessing Competitive Intelligence Software : A Guide to Evaluating CI Technology*. Medford, Information Today Inc.
Fiscus J.G, Wheatley B. (2004). Overview of the TDT 2004 evaluation and results. In *TDT–02*, NIST.
Robertson S., Soboroff I. (2002). The TREC 2002 Filtering Track Report. In *Proceedings of The Eleventh Text Retrieval Conference* (*TREC–2002*). NIST Special Publication : 500-251, http://trec.nist.gov/pubs/trec11/papers/OVER.FILTERING.pdf
Soboroff I., Robertson S. (2002). Building a Filtering Test Collection for TREC 2002. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR–03)*
Van Rijsbergen, C.J. (1979) Information Retrieval. Butterworths, London.
Voorhees, E.M (1999) The TREC-8 Question Answering Track Report, in Proceedings of the Eighth Text REtrieval Conference (TREC-8).