Critères numériques et temporels pour la détection de documents vitaux dans un flux

Vincent Bouvier[*,**] — **Patrice Bellot**[**]

[*] *Kware, Aix-en-Provence, France;*
[**] *Aix-Marseille Université, CNRS, LSIS UMR 7296, Marseille, France;*

RÉSUMÉ. *Cet article s'intéresse, au travers de la construction de critères numériques et temporels, à la problématique de classification de documents dans un processus de filtrage de documents vitaux. La classification a pour but de différencier les documents "vitaux" des autres documents. Un document est défini comme vital s'il concerne l'entité pour laquelle ce dernier a été sélectionné et surtout s'il contient une importante nouveauté sur une entité. Nous présentons les différents critères que nous mettons en place, dans le but de créer un modèle adaptatif qui ne dépend pas des entités sur lesquels le modèle est entraîné. La méthode est donc semi-supervisée. Ensuite, nous évaluons les résultats obtenus et nous les confrontons aux résultats de la campagne d'évaluation TREC KBA 2013 (Knowledge Base Acceleration).*

ABSTRACT. *This paper addresses to a classification challenge in a filtering task. We use different kind of features to depict vital documents and filter them out from the stream. A vital document has to be relevant for a particular entity and has to relate a new story about it. We introduce different features that uses time as well as entity profil to perform classification. We evaluate our method on framework from TREC KBA 2013 (Knowledge Base Acceleration).*

MOTS-CLÉS : *Filtrage, modèle adaptatif, profil d'entité, Random Forest,*

KEYWORDS: *Filtering, Adaptive Model, Entity Profil, Random Forest*

1. Introduction

Il existe aujourd'hui plusieurs alternatives pour suivre des informations sur le Web. Suivant la nature de l'information recherchée, il est possible d'utiliser des sites spécialisés dans l'actualité, dans les réseaux sociaux et microblogues, dans les blogues ou forums. Parcourir toutes ces sources et sélectionner les informations pertinentes est un processus long et fastidieux. Pourtant, ces informations constituent un réel besoin dans certains domaines d'applications comme la veille technologique ou encore le suivi informationnel (revue de presse). Par ailleurs, il a été observé dans l'étude de (Frank *et al.*, 2012), que les bases de données collaboratives, comme Wikipédia, souffrent de leurs grandes envergures qui ne permet pas un maintien à jour toujours efficace des informations qu'elles contiennent. Cette étude montre qu'il existe un temps médian de 356 jours entre le moment où une information concernant un sujet ou une entité relativement peu populaire apparait sur le Web et le moment où elle est relayée sur Wikipédia.

Ces problématiques soulèvent un réel besoin de veille informationnelle, qui permet de détecter de nouvelles informations sur un sujet en particulier à partir d'un certain nombre de sources. Le fait de traiter, de manière séquentielle, les documents provenant d'une ou plusieurs sources afin d'extraire ceux qui sont pertinents par rapport à un sujet correspond à la notion de *filtrage de documents* (Belkin et Croft, 1992). À la différence d'un système de RI classique, s'appuyant sur un index, le filtrage traite à la volée les documents dès leurs apparitions sur ce que l'on appelle "un flux".

Il est possible de définir un sujet à l'aide de simples mots clés, de phrases ou encore d'un ensemble de documents. Nous proposons de nous intéresser au processus de filtrage de documents lié à une entité en passant par la construction de profils d'entités. Un profil d'entité a pour rôle de représenter une entité et ses caractéristiques intrinsèques au travers de structures comme des modèles de langue, des graphiques de relation entre entités... Nous utilisons ces profils pour calculer les valeurs associées aux critères numériques et temporels pour chacun des documents filtrés, afin de définir leurs degrés de pertinence à l'aide d'une méthode de classification.

La méthode proposée a la particularité d'être indépendante des entités sur lesquelles le modèle de classification est créé. Il est alors possible d'utiliser ce modèle pour classer les documents qui traitent d'entités non présentes lors de la phase d'entrainement.

Pour tester notre méthode, nous avons utilisé l'environnement fourni pour la tâche KBA 2013 (Knowledge Base Acceleration) de la campagne d'évaluation TREC (Text RETrieval Conference).

La suite de cet article se compose d'une partie qui décrit la tâche KBA 2013 afin de prendre connaissance des particularités de la tâche et des différentes classes de documents que l'on devra filtrer. Ensuite, nous étudierons, dans une partie état de l'art, des méthodes de filtrages et de classification de documents selon des critères particuliers. Nous détaillerons les différents critères mis en place et nous étudierons

leurs impacts sur la classification de documents. Enfin, nous comparerons nos résultats à ceux soumis par les participants de la tâche KBA 2013 à TREC. Finalement, nous donnerons nos conclusions et perspectives.

2. Description de la tâche KBA 2013

La tâche KBA est directement liée au problème de mise à jour de bases de connaissances énoncé précédemment. Cette tâche a été lancée pour la première fois en 2012, où un simple filtrage de documents était demandé aux participants sans inclure de notion de nouveauté. Cependant, il ne s'agit pas d'une nouvelle tâche ad hoc de RI. En effet, l'originalité réside en partie dans le corpus appelé *"stream-corpus"* que l'on traduira par flux de documents.

Ce corpus contient environ un milliard de documents (en 2013) issus du Web et, plus spécifiquement de site d'actualité, des forums ou encore des blogues. Le point commun entre toutes ces catégories de site réside dans le fait que les documents sont datés. Les participants doivent trouver et filtrer des documents qui concernent un ensemble d'entités (sélectionnées par les organisateurs) en parcourant le corpus de manière chronologique. Il est également demandé de donner une classe aux documents filtrés parmi les quatre classes suivantes : vide/inutile, neutre, utile et vitale. Nous nous intéresserons en particulier aux classes "utile" et "vitale" qui respectivement définissent un document contenant une information connue et un document contenant une nouvelle information. La différence entre ces deux classes est très mince, ce qui ne rend pas la tâche de différenciation triviale. La prise de décision concernant la classe d'un document doit être immédiate et donc il n'est pas possible de prendre de décision a posteriori en observant des documents apparaissant plus tard dans la chronologie.

3. État de l'art

Le meilleur système de KBA 2012 (Kjersten et McNamee, 2013) utilise un système de classification simplement en créant un modèle par entité pour catégoriser les documents. Ce système impose d'avoir des données d'entrainement pour chaque nouvelle entité que l'on souhaite suivre. Cette contrainte est très forte et donc il faudra trouver un moyen de généraliser la classification de documents afin que le modèle puisse s'appliquer à n'importe quelles entités.

Nous avons, au travers d'une précédente étude (Bonnefoy *et al.*, 2013a ; Bonnefoy *et al.*, 2013b), réalisé des travaux relativement proches de ceux de (Zhou et Chang, 2013) dans le sens où nous avons essayé de généraliser le problème plutôt que d'y répondre spécifiquement pour l'ensemble d'entités fournies. Nous avons utilisé l'algorithme de classification Random Forest pour tenter de corréler un ensemble de valeurs de caractéristiques (équivalentes aux méta caractéristiques) afin de déterminer la pertinence d'un document. Il a été montré dans une étude de (Huang *et al.*, 2003), que les algorithmes de classification Naive Bayes, Random Forest ou SVM offrent

des performances proches sur plusieurs corpus. Nous avons décidé d'utiliser le Random Forest qui permet, à l'aide des Variables d'Importances, d'étudier les critères prédominant dans la classification (Breiman, 2001). Nous étions donc en mesure de catégoriser n'importe quelles entités sans pour autant avoir un corpus d'entrainement pour celles-ci. Les limites de cette méthode résident dans le choix des caractéristiques qui ne permettaient pas toujours de répondre au problème de manière efficace. Nous avons également tenté d'ajouter des caractéristiques liées à la temporalité, mais celles-ci dégradaient (en partie) les résultats.

Le filtrage de documents concernant une entité est une tâche difficile. Pour cela, il faut déjà que l'entité soit *"bien définie"* de manière à permettre une détection sur le flux de documents. Nous nous sommes alors intéressés à la construction de profils d'entités. (Li *et al.*, 2003) proposent une méthode pour extraire des profils d'entités en utilisant des patrons d'extraction. Pour cela, ils définissent la notion de profil d'entité comme étant une matrice d'attributs-valeurs en ne prenant pas en compte les possibles relations entres attributs, ni le contexte dans lequel l'entité apparait.

L'article de (Sehgal et Srinivasan, 2007) montre la construction de profil en utilisant, des documents résultant de requêtes émises sur l'API "Google Search". Il relève ensuite la similarité du profil avec la page Wikipédia de l'entité recherchée pour évaluer sa méthode. Il obtient des similarités plus élevées lorsque le profil est construit en utilisant l'intégralité des tops n documents trouvés. Cet article montre qu'il est possible d'obtenir une représentation d'une entité à partir de documents issus du Web, en utilisant une page Wikipédia comme support de comparaison. En considérant le problème à l'inverse, on peut faire l'hypothèse qu'une page issue de Wikipédia et focalisée sur une entité peut permettre d'avoir une représentation globale de ce qu'est l'entité. À minima, cela permet d'avoir une connaissance sur le contexte dans lequel l'entité est susceptible d'apparaitre.

En 2013, sur la tâche KBA, certains systèmes se sont fortement inspirés des meilleurs modèles de l'année précédente (Bellogín et Gebremeskel, 2014 ; Wang *et al.*, 2014). En revanche, la méthode de (Efron, 2014) montre un nouvel aspect dans les profils : le dynamisme. Jusqu'alors les profils qui étaient construits n'avaient pas pour vocation d'évoluer. Un profil est constitué d'un modèle de langue auquel sont ajoutés les mots présents dans les documents jugés pertinents. Cette méthode constitue une première étape dans la réalisation de notre but, mais elle est trop arbitraire. En effet, en cas d'erreur sur le jugement du document, tout le modèle de langue est altéré. Par ailleurs, pour éviter de faire grossir le modèle de langue trop fortement, ils limitent le nombre de documents qui permettent de constituer le modèle en créant une fenêtre glissante sur les documents jugés pertinents.

4. Critères de classification pour le filtrage

Un des principaux challenges consiste à trouver un ensemble de critères qui permet de caractériser au mieux la notion de document utile et vital sans tenir compte de

l'entité qui est concernée. Il sera alors tout à fait possible de calculer ces critères pour une entité pour laquelle aucune donnée d'entrainement n'a été fournie. Il y aura donc qu'un seul modèle.

Pour avoir une idée des critères pouvant être pertinents pour la détection de documents vitaux, nous avons tout d'abord procédé à une analyse des documents disponibles dans le corpus d'entrainement de KBA. Nous avons distingué ainsi trois types de critères que nous allons détailler dans les sous-parties suivantes :

- Les critères de correspondance avec l'entité ;
- Les critères du document ;
- Les critères temporels.

4.1. *Les critères de correspondance avec l'entité*

Une entité ou un sujet peut être décrit de plusieurs manières. Nous parlerons à présent de la notion de profil d'entité comme étant une structure qui permet de rassembler des informations sur une entité. Dans le cadre de la tâche KBA, 150 entités ont été sélectionnées par les organisateurs. Nous ferons référence à ces entités en parlant des entités "Twitter" et des entités "Wikipedia" suivant leur provenance. Par ailleurs, les participants ont le droit d'utiliser une version d'une base Wikipédia antérieure au début du flux de documents (début 2012). Cependant, pour certaines entités (ex. des utilisateurs Twitter), aucune page d'information n'est présente.

Nous présentons dans la suite de cette section les différents composants du profil.

4.1.1. *Les variantes de nom*

Une entité peut être mentionnée dans un document sous différents noms que nous appelons variantes. Par exemple, Elvis Presley était souvent appelé *"le King"*. Les variantes de noms permettent le plus souvent d'augmenter le rappel des documents retrouvés.

Pour obtenir ces variantes de manière non supervisée, les observations de (Cucerzan, 2007) proposent, pour une entité pour laquelle la page Wikipédia est connue, de sélectionner à la fois tous les mots en gras dans le premier paragraphe de la page ainsi que toutes les légendes des liens qui pointent vers la page Wikipédia correspondant à l'entité.

Pour les entités pour lesquelles aucune page Wikipédia ne peut être associée, il peut exister d'autres alternatives selon le type d'entité. Par exemple, pour des entités issues des réseaux sociaux, il est possible d'utiliser l'identifiant de l'entité et/ou les valeurs des champs "nom" et "prénom" lorsque ces derniers sont indiqués. Sur Twitter par exemple, l'identifiant correspond au nom d'utilisateur commençant par `@`. Par ailleurs, il est aussi possible de connaitre le nom et prénom de la personne directement sur la page du profil de l'utilisateur en question.

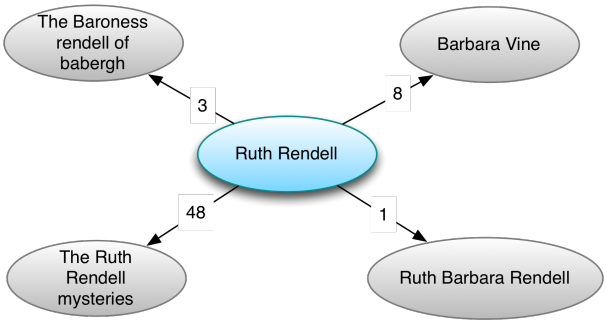

Figure 1. *Graphe des variantes pour l'entité cible Ruth Rendell avec sur chaque lien un poids qui correspond au nombre de fois où la variante a été trouvée.*

Le calcul des différents critères sur un document est un processus couteux en terme de temps de calcul. C'est pourquoi, nous proposons d'effectuer un premier filtrage de documents en se basant sur les mentions des variantes de noms à l'intérieur des documents. La manière la plus naïve et permissive (dans une optique d'obtenir un rappel plus important) consiste à considérer tous les documents ayant au moins une mention d'une entité. Les documents ainsi filtrés sont alors soumis au processus de classification.

Dans le processus de classification, différents critères sont calculés. Nous proposons d'utiliser les variantes de noms comme un critère en utilisant la mesure *TF* (Term Frequency). Le *TF.IDF* est une mesure souvent utilisée en RI. Cette mesure permet de calculer l'importance d'un mot dans un document (TF) en tenant compte de la rareté de ce mot dans un corpus de documents (IDF : Inverse Document Frequency). Le calcul de l'IDF est problématique sur un flux de documents, car ce dernier change au fur et à mesure que le flux de documents se déroule. Nous proposons dans un premier temps de n'utiliser que le TF calculé pour une entité e ayant un ensemble de variantes de noms V_e où toutes les mentions des variantes $v \in V_e$ trouvées par la fonction $f(v, D)$ sont sommées puis normalisés par la taille $|D|$ d'un document D (equation 1).

$$tf(V_e, D) = \frac{\sum_{v \in V_e} f(v, D)}{|D|} \quad [1]$$

Les critères qui utilisent les variantes de noms sont répertoriés dans le tableau 1. Les variantes de noms sont également utilisées pour construire un snippet (extrait) du document. Il est possible de construire un tel extrait en agglomérant tous les paragraphes dans lesquels il existe une mention de l'entité (voir figure 2).

tf_title	# mentions dans le titre
$tf\_document$	# mentions dans le document

Tableau 1. *Critères liés aux mentions dans le document*

Figure 2. *Snippet (extrait) de documents contenant des mentions de l'entité Edgar M. Bronfman*

4.1.2. *Le modèle de langue de l'entité*

Nous souhaitons mettre en place des critères qui permettent de mesurer la similarité d'un document avec un profil. Pour cela, un modèle de langue par entité doit être construit.

Pour les entités "Wikipédia", le modèle sera constitué à l'aide de la page associée à l'entité. Concernant les entités Twitter, le modèle sera vide. En effet, il serait tout à fait possible d'utiliser la description présente sur le profil, mais pour des raisons de cohérence dans le temps, nous ne pouvons pas les utiliser pour notre expérimentation.

Le modèle de langue que nous avons choisi est une représentation en sac de mots du document. Nous effectuons un pré-traitement sur les mots : nous ignorons les mots récurrents de la langue anglaise (utilisation d'une liste) ; les mots sont lemmatisés et mis en minuscule.

La similarité Cosine (équation 2) peut être utilisée directement sur le vecteur représenté par le sac de mots. Étant donnés deux vecteurs de mots θ et D représentant respectivement le modèle de langue d'une entité et le document, il est possible de

calculer la similarité cosinus en utilisant le produit scalaire de θ et D divisé par le produit des normes.

$$cos(\theta, D) = \frac{\theta \cdot D}{||\theta|| \cdot ||D||} = \frac{\sum_{i=1}^{n} \theta_i \times D_i}{\sqrt{\sum_{i=1}^{n} (\theta_i)^2} \times \sqrt{\sum_{i=1}^{n} (D_i)^2}} \quad [2]$$

La représentation sac de mots permet également d'utiliser des mesures telles que la divergence $js(P_\theta, P_D)$ de Jensen-Shannon (Endres et Schindelin, 2003) à condition d'utiliser des distributions de probabilités P_θ et P_D pour représenter respectivement le modèle de langue de l'entité θ et le document D.

$cos(\theta_e, D)$	similarité : modèle θ_e vs document D
$cos(\theta_e, D_{snippet})$	similarité : modèle θ_e vs snippet
$js(P_{\theta e}, P_D)$	divergence : modèle θ_e vs document
$js(P_{\theta e}, P_{D\,snippet})$	divergence : modèle θ_e vs snippet
$size(\theta_e)$	taille du modèle θ_e

Tableau 2. *Critères liés au modèle de langue θ_e d'une entité e*

4.1.3. *Les liens entre entités*

Il est possible de représenter les bases de connaissances, ou encore les réseaux sociaux comme un ensemble de graphes orientés où les noeuds correspondent à une information et les liens correspondent aux liens qui existent entre les informations. L'information dans le cas de Wikipédia correspond à une page. Dans le cas d'un réseau social, l'information concerne le plus souvent l'utilisateur. Il est possible de distinguer plusieurs types de liens par rapport à leurs orientations :

liens entrants : correspond au lien qui arrive sur le noeud ;

liens sortants : correspond au lien qui part vers un noeud différent ;

liens réciproques : lorsqu'il existe un lien entrant et sortant entre deux mêmes noeuds.

Pour établir les critères c_{lien} pour chacun des types de liens, nous calculons la fréquence d'apparition $tf(e, D)$ dans le document D de l'entité e liée au profil. Cette fréquence est normalisée par le nombre d'entités appartenant au type de lien $|L_{type}|$:

$$c_{lien}(L_{type}, D) = \frac{\sum_{e}^{L_{type}} tf(e, D)}{|L_{type}|} \quad [3]$$

4.2. *Les critères du document*

Il est important de prendre en compte des critères intrinsèques au document afin d'apporter une information supplémentaire sur la richesse d'information contenue dans le document indépendamment d'une entité. En effet, en utilisant l'entropie, il est possible d'avoir un indicateur sur la quantité d'information délivrée par un document.

Dans les critères présentés dans la table 1, un critère sur le nombre de mentions dans le titre est calculé. Cependant, nous souhaitons être capables de distinguer les cas où un document n'a pas de titre, des cas où un document à un titre. Pour cela, nous ajoutons un critère de type booléen.

4.3. *Les critères temporels*

Nous avons observé la distribution des documents par classes sur un axe temporel. La figure 3 montre que le nombre de documents (axe des ordonnées) peut varier très fortement d'un jour à l'autre (axe des abscisses), caractérisant ainsi le phénomène de rafale (ou "burst" en anglais). Ce phénomène peut apparaitre sur des documents utiles comme sur des documents vitaux. Ce critère peut être un bon indicateur sur la pertinence du document.

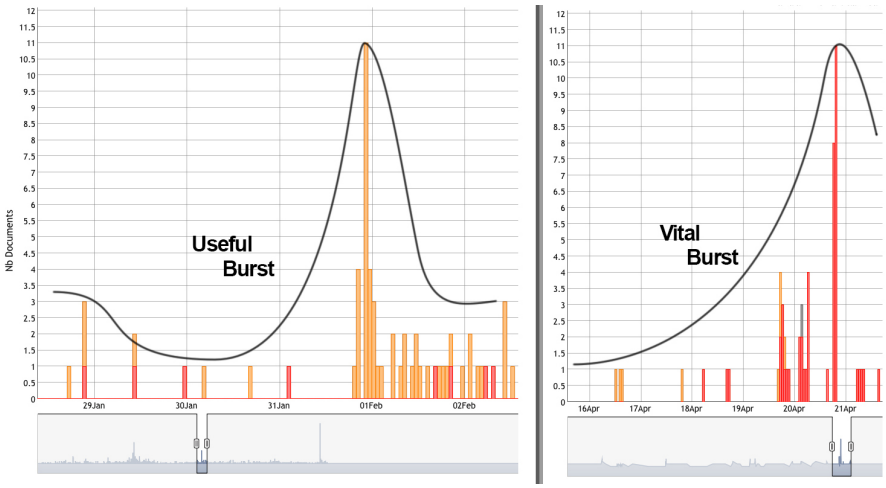

Figure 3. *Le phénomène de rafale, constaté sur deux entités différentes, ne donne pas forcément lieu à des documents dits vitaux. À gauche documents utiles, à droite documents vitaux.*

Pour caractériser l'effet rafale, nous avons utilisé une implémentation de l'algorithme de (Kleinberg, 2002) qui permet de déterminer la "force" de la rafale

d'après une analyse de série chronologique (time series). Il est possible d'utiliser différentes échelles pour calculer la série chronologique. Nous utilisons une échelle basée sur une heure pour calculer la série chronologique. L'algorithme permet également de donner la tendance de la rafale (montante ou descendante). Nous utiliserons enfin cette caractéristique directement sur le score de la force en appliquant un coefficient de -1 lorsque la rafale est descendante.

Aussi nous calculons une caractéristique basée sur le nombre de documents dans lesquels il y a une mention de l'entité les 24 heures précédant l'apparition du document évalué (voir tableau 3).

Nous avions constaté, à la suite d'expérimentations supplémentaires sur nos études (Bonnefoy *et al.*, 2013a ; Bonnefoy *et al.*, 2013b), que certains critères temporels tendaient à dégrader les résultats. Nous avons finalement choisi de ne garder que les critères en correspondance avec le nombre de mentions trouvés les dernières 24 heures et celui sur la force (et direction) d'un éventuel effet de rafale par rapport aux observations décrites précédemment.

$kleinberg_{1h}$	force et direction de la rafale
$match_{24h}$	nombre de documents trouvés les dernières 24h

Tableau 3. *Critères liés à la date d'apparition du document*

5. Expérimentations

Nous avons utilisé le corpus d'entrainement fourni par les organisateurs de TREC KBA pour entraîner des algorithmes de classification de types *"Random Forest"* sur les documents annotés. Nous parcourons le corpus de manière chronologique afin de simuler un flux de documents. Nous pouvons ainsi calculer toutes les caractéristiques précédemment présentées.

Nous avons testé quatre systèmes différents :

- le premier *"Single"* entraine un algorithme de classification qui connait les quatre classes de documents de KBA (Garbage, Neutral, Useful, Vital). La réponse de ce dernier donne la classe du document ;

- le second système *"TwoStepClassifier"* est composé de deux algorithmes de classification utilisés en cascade. Chacun s'entraîne sur 2 classes. Le premier donne une classe parmi : *"Garbage/Neutral"* et *"Useful/Vital"*. Le second algorithme de classification donne une classe parmi : *"Useful"* et *"Vital"* ;

- le troisième système *"VitalvsOthers"* entraine un algorithme de classification qui apprend à distinguer seulement la classe *"Vital"* contre toutes les autres classes (*"Others"*).

Le quatrième système *"Combine"* est quant à lui un peu différent puisqu'il va apprendre les scores issus des trois précédents systèmes afin de trouver la meilleure combinaison possible.

6. Résultats

Nous avons utilisé l'outil d'évaluation officiel pour pouvoir comparer nos résultats à ceux des autres participants de KBA 2013 sur la partie test du corpus. La mesure utilisée pour l'étude est la f-mesure ($f_1(p, r)$) qui est une moyenne harmonique entre la précision p et le rappel r (voir équation 4).

$$f_1(p, r) = \frac{2 * p * r}{p + r} \quad [4]$$

Dans l'évaluation de KBA 2013, il y a deux aspects pris en compte : la recherche d'information et la classe donnée à un document. Dans notre analyse, nous comparons nos scores obtenus avec l'outil d'évaluation officiel afin de servir de référence. Nous analyserons aussi d'autre part l'efficacité de la classification seule en prenant en compte un système de RI parfait.

6.1. *Analyse des résultats dans le cadre de KBA 13*

Nous pouvons voir d'après le tableau 4 que nous obtenons des résultats supérieurs au meilleur résultat présenté lors de la campagne d'évaluation pour la majorité de nos systèmes sur la classification de documents vitaux. De plus, nous remarquons que d'apprendre à combiner les scores des différents systèmes s'avère être une stratégie payante.

Système	f_1
Best KBA	.360
Mean KBA	.193
Single	.395
TwoStep	.388
VitalVsOther	.346
Combine	**.403**

Tableau 4. *Récapitulatif des résultats présentés à TREC KBA et de ceux de nos différents systèmes*

Les scores de KBA sont observés en utilisant plusieurs seuils suivant le score de confiance donné par l'algorithme de classification (score allant de 0 à 1000). Le

graphique 4 montre une cohérence dans le score de confiance donné à l'issue de la classification. En effet, plus le score est haut (proche de l'axe des ordonnées) plus la précision monte. En revanche, pour les autres systèmes, les pourcentages de précision, rappel et f-mesure s'écroulent lorsque le score de confiance augmente.

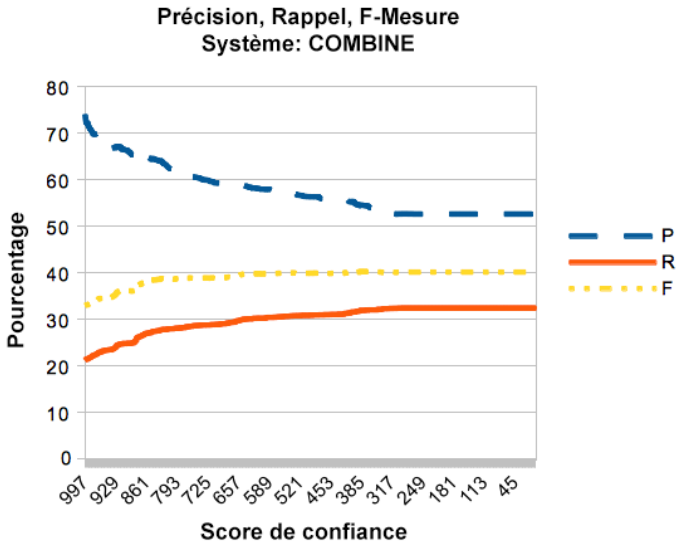

Figure 4. *Variation de la Précision (P), du Rappel (R) et de la F-mesure (F) en fonction du score de confiance du système Combine.*

Si l'on regarde d'un point de vue seulement de classification de documents (table 5), on remarque que les scores de classifications sont plutôt bons, ce qui montre une lacune de notre système sur le processus de recherche d'information. Pour cette analyse, nous utilisons une matrice de confusion (Vrais Positifs, Vrais Négatifs, Faux Positifs, Faux Négatif) pour calculer les scores de précision, rappel et f-mesure.

Système	Precision	Rappel	F-Mesure
Single	.569	.406	**.474**
TwoStep	.475	**.436**	.455
VitalVsOther	**.725**	.323	.447
Combine	.619	.368	.461

Tableau 5. *Résultats de la classification sans prendre en compte le processus de RI*

Cette analyse nous permet de voir que le système *"VitalVsOther"* offre la précision la plus haute au détriment du rappel. Nous pouvons constater également que le système *"Combine"* tente de tirer parti de tous les systèmes en offrant une précision

relativement élevée et un rappel plus élevé que *"VitalVsOther"*. C'est d'ailleurs *"Combine"* qui obtient le deuxième meilleur score en f-mesure.

Nous avons voulu en savoir plus sur les critères prédominants dans le processus de décision de la classe vital. Pour cela, nous avons utilisé le logiciel R et la bibliothèque "Party" qui implémente un algorithme de classification de type *"Random-Forest"* pour lequel il est possible de calculer les Variables d'Importances (VI). Les VI sont calculées à l'aide d'une permutation aléatoire des valeurs des différents critères pour finalement, calculer la différence de précision "avant/après" révélant ainsi l'importance du critère (Breiman, 2001).

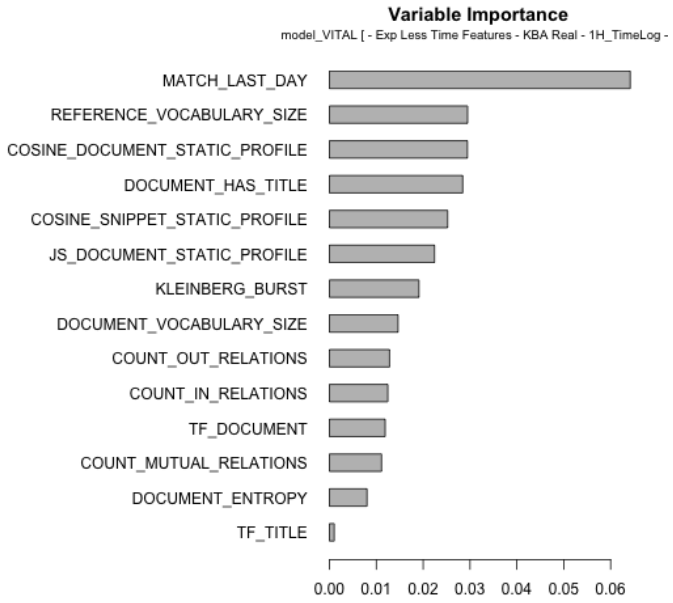

Figure 5. *Classement des critères en fonction de l'importance dans la classification, avec en abscisse l'importance relative des critères.*

La figure 5 montre que le critère le plus important dans la classification correspond au critère qui regarde le nombre de documents apparus les 24 heures précédentes et portant au moins une mention de l'entité. Cela montre qu'effectivement le temps peut jouer un rôle très important dans la manière dont est publiée une nouvelle information. Cependant, il faut tout de même être prudent sur l'utilisation de tels critères. En effet, les valeurs calculées pour certains critères peuvent être très variables en fonction des entités. Par le fait, ces critères ne sont peut être pas adaptés à un modèle généraliste.

On notera cependant que le critère mesurant la rafale de Kleinberg ne joue pas un rôle important. En observant l'allure du critère sur les modèles, les scores sont tous négatifs. Autrement dit, il ne capte que des rafales descendantes. Nous projetons de faire une étude un peu plus poussée afin de trouver comment utiliser de manière efficace les différents paramètres de l'algorithme.

Le deuxième critère correspond à la taille du modèle de langue de l'entité. Le modèle a très bien su s'adapter aux différences qu'il peut y avoir entre les entités de type Wikipédia et Twitter où, pour ces dernières, le modèle de langue est vide.

7. Conclusion et Perspectives

Nous avons présenté dans cet article trois différents types de critères permettant la classification de documents vitaux pour une entité. Nous avons également montré que le modèle généré à l'issue de l'entrainement n'est pas dépendant des entités sur lesquelles il est créé. Finalement, nous avons montré les performances de notre système en nous comparant aux résultats de la campagne d'évaluation TREC KBA 2013.

Nous planifions de revoir nos paramètres sur le critère de rafale afin de voir s'il est possible de trouver un bon réglage afin de détecter la rafale au moment de la montée, ce qui nous semble plus logique pour la découverte de documents vitaux. Nous avons vu dans l'état de l'art une étude (Efron, 2014) sur l'évolution du profil d'entité en fonction du temps. Nous aimerions nous investir dans cette voie afin de voir s'il est possible d'adapter notre système pour prendre en compte le dynamisme des entités. Enfin, nous voudrions retravailler notre méthode préliminaire de recherche d'information, afin d'améliorer le rappel de notre système. Ici, nous utilisons les variantes de nom issues de Wikipédia, cependant nous ne les mettons jamais à jour. Par exemple, il serait intéressant de prendre en considération un système d'extraction d'entités nommées afin de permettre d'extraire de nouvelles variantes de noms.

8. Bibliographie

Belkin N. J., Croft W. B., « Information filtering and information retrieval : two sides of the same coin ? », *Communications of the ACM*, vol. 35, nº 12, p. 29-38, December, 1992.

Bellogín A., Gebremeskel G., « CWI and TU Delft Notebook TREC 2013 : Contextual Suggestion, Federated Web Search, KBA, and Web Tracks », *proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013) Gaithersburg, Maryland, November 19–22, 2013*, National Institute of Standards and Technology (NIST), p. 500-302, 2014.

Bonnefoy L., Bouvier V., Bellot P., « LSIS/LIA at TREC 2012 knowledge base acceleration », *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012)*, p. 500-298, 2013a.

Bonnefoy L., Bouvier V., Bellot P., « A weakly-supervised detection of entity central documents in a stream », *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, p. 769-772, 2013b.

Breiman L., « Random Forests », *Machine Learning*, vol. 45, n° 1, p. 5-32, 2001.

Cucerzan S., « Large-Scale Named Entity Disambiguation Based on Wikipedia Data », *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,(EMNLP-CoNLL 2007)*, ACL, p. 708-716, 2007.

Efron M., « The University of Illinois' Graduate School of Library and Information Science at TREC 2013 », *proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013) Gaithersburg, Maryland, November 19–22, 2013*, National Institute of Standards and Technology (NIST), p. 500-302, 2014.

Endres D. M., Schindelin J. E., « A new metric for probability distributions », *IEEE Transactions on Information Theory*, vol. 49, n° 7, p. 1858-1860, 2003.

Frank J., Kleiman-Weiner M., Roberts D. A., Niu F., Zhang C., « Building an Entity-Centric Stream Filtering Test Collection for TREC 2012 », *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012) Gaithersburg, Maryland, November 6-9, 2012*, National Institute of Standards and Technology (NIST), p. 500-298, 2012.

Huang J., Lu J., Ling C. X., « Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy. », *ICDM*, IEEE Computer Society, p. 553-556, 2003.

Kjersten B., McNamee P., « The HLTCOE approach to the TREC 2012 KBA track », *proceedings of the Twenty-First Text REtrieval Conference (TREC 2012) Gaithersburg, Maryland, November 6-9, 2012*, National Institute of Standards and Technology (NIST), p. 500-298, 2013.

Kleinberg J., « Bursty and hierarchical structure in streams », *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, p. 91-101, 2002.

Li W., Srihari R., Niu C., « Entity profile extraction from large corpora », *Proceedings of Pacific Association for Computational Linguistics (PACLING 2003)*, 2003.

Sehgal A. K., Srinivasan P., « Profiling Topics on the Web », *Proceedings of the WWW2007 Workshop I^3 : Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007*, p. 1-8, 2007.

Wang J., Song D., Lin C.-Y., Liao L., « BIT and MSRA at TREC KBA CCR Track 2013 », *proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013) Gaithersburg, Maryland, November 19–22, 2013*, National Institute of Standards and Technology (NIST), p. 500-302, 2014.

Zhou M., Chang K. C.-C., « Entity-centric document filtering : boosting feature mapping through meta-features », *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, ACM, p. 119-128, 2013.