Regroupement par popularité pour la RI semi-supervisée centrée sur les entités

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

[*] *Kware, 565 Rue Marcelin Berthelot, Aix-en-Provence, France*
[**] *Aix-Marseille Université, CNRS, LSIS UMR 7296, Marseille, France*

RÉSUMÉ. *Filtrer des documents web à propos d'entité (personne, entreprise ...) pour que seuls les documents d'intérêt soient gardés est un réel challenge. L'intérêt peut être qualifié de différente manière comme la nouveauté ou le fait qu'une information soit récente. Nous avons pu voir au cours des dernières années que des systèmes s'entraînent à détecter l'intérêt d'un document au regard d'une entité. Pour des raisons de passage à l'échelle, il n'est pas pensable d'avoir des données annotées manuellement pour chaque entité recherchée. Les approches obtiennent de bonnes performances, mais nous montrons que celles-ci peuvent être améliorées. Les entités peuvent différer sur certains aspects qui peuvent être mieux exploités grâce au regroupement (clustering). Cet article a pour but de montrer la valeur ajoutée que le regroupement peut avoir sur ce type de problème en utilisant une méthode de regroupement basique. Nous testons notre approche sur la tâche Knowledge Base Acceleration (KBA) de TREC 2013 et 2014 et nous obtenons des résultats significativement meilleurs.*

ABSTRACT. *Filtering pages about an entity (person, company, ...) so that only documents being of interest are kept is a real challenge. The interest can be qualified using criteria such as recency, novelty. In the last decade, we have seen classification systems trained to detect the interest for a document regarding an entity. Some approaches strive to build entity independent classification systems. Those approaches obtain good performances, but we show that they can be improved. The entities may differ on certain aspects that we think can be caught using clustering. Thus, instead of having one model per entity or one model for all entities, we propose an approach that uses one model per cluster of entities. We also introduce different strategies for automatic classification model selection. In this article, we detail the different aspects of our approach and we test it on the Knowledge Base Acceleration framework from the Text REtrieval Conference. We show that our approach brings significant improvements over a non-cluster based method.*

MOTS-CLÉS : *regroupement d'entités, classification, détection d'informations importantes*

KEYWORDS: *entity clustering, classification, important information detection*

1. Introduction

Filtrer des documents issus du Web de manière à garder seulement les informations qui sont importantes au regard d'une entité est un vrai challenge. Une entité peut être une personne, un groupe de musique, une organisation. Des piques de documents peuvent arriver lorsqu'un évènement important se produit pour une entité. Cela reste cependant très dépendant de l'entité. Si on considère une entité populaire, quand un évènement important se produit pour elle, il est très probable qu'une rafale de documents apparaisse à ce moment-là (figure 1). Lorsque l'entité est moins populaire, l'annonce d'un évènement peut n'avoir que très peu voir pas d'effet du tout. Cela pose donc un réel problème lorsqu'un système cherche à effectuer des statistiques en se basant sur le phénomène de rafale sans prendre en compte les disparités qui existent entre les entités.

Une solution complètement supervisée pourrait permettre de s'abstraire de ces contraintes, cependant d'autres contraintes plus fortes apparaissent comme le fait d'avoir de nouvelles données d'entraînement pour chaque nouvelle entité que l'on recherche. Nous proposons de regrouper les entités pour lesquels des comportements similaires sont observés en terme de nombre documents publiés dans le temps afin de vérifier notre hypothèse sur les critères de temporalité. L'idée générale de cet article est de voir si le regroupement d'entités peut avoir un impact positif sur la classification des documents vitaux. Nous évaluons notre approche en utilisant les données de la tâche KBA de 2013 et 2014.

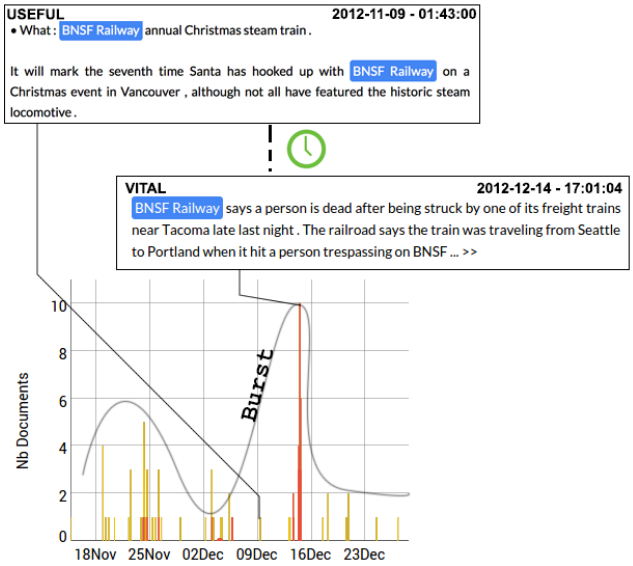

Figure 1. *Effet de rafale (burst) pour l'entité BNSF Railway.*

2. Travaux connexes

Tous les jours, énormément de documents textuels apparaissent sur le Web. Certains d'entre eux relatent les mêmes faits et même parfois avec les mêmes mots. Détecter qu'une information est importante est un réel challenge et peu représenté un besoin. (Frank *et al.*, 2012) montrent dans une étude que les bases de connaissance de grande envergure (comme Wikipedia) peuvent souffrir d'une grande latence pour les mises à jour d'articles. Le but de la tâche KBA de TREC consiste à filtrer un flux de documentsg de manière automatique afin de ne garder que les documents qui contiennent une information importante pour une des entités recherchées. Les documents ainsi filtrés pourraient servir à mettre à jour de telles bases de connaissance soit de manière automatique soit en suggérant le document à un contributeur. On pourrait également voir d'autres applications comme la veille informationnelle ou même le marketing qui sont deux domaines où la réactivité est importante. Les organisateurs de KBA ont défini quatre classes de document *garbage, neutral, useful* et *vital*. Un document est associé aux classes garbage ou neutral s'il ne se focalise pas sur une des entités recherchées. La classe du document sera useful si le document relate un fait déjà connu ou pas important concernant une des entités recherchées. La classe *vital* indique que le document contient une information importante concernant une des entités recherchées. Pour pouvoir simuler un système réel, les organisateurs ont construit un corpus, qui contient plus d'un milliard de documents datés provenant de fils d'actualités, de forums et blogs(Frank *et al.*, 2012). Ainsi il est possible de parcourir les documents de manière chronologique pour simuler un flux de document. Chaque année un ensemble d'entités est également sélectionné pour la non-popularité ou leur ambiguïté rendant la tâche encore plus compliquée.

Dans le cadre de la tâche KBA, (Bonnefoy *et al.*, 2013 ; Bellogín et Gebremeskel, 2014 ; Bouvier et Bellot, 2014 ; Balog *et al.*, 2013 ; Efron, 2014) utilisent un système de classification à base de méta critères pour évaluer les documents issus du flux de documents. Le fait d'utiliser des méta critères rend les systèmes indépendants des entités sur lesquels ils s'entraînent. Cela offre la possibilité d'évaluer des documents par rapport à de nouvelles entités sans avoir besoin de données d'entraînement supplémentaires. Ils utilisent les profils d'entités pour faciliter la détection de celles-ci dans les documents et pour garder une trace de leurs actualités. (Dietz et Dalton, 2014) utilisent une méthode d'ordonnancement (ranking) et d'expansion de requêtes pour évaluer les documents cependant, la méthode présentée ne comprend pas de notion d'importance de l'information par rapport à l'entité.

Notre approche s'inscrit dans la continuité des travaux présentés lors de KBA 2013. Plutôt que d'utiliser un système de classification avec un modèle unique, l'idée est d'utiliser une méthode de regroupement et de calculer un modèle par groupes d'entités. Dans un premier temps nous cherchons à savoir si le regroupement d'entités pour la classification apporte une valeur ajoutée. Nous utiliserons une méthode de regroupement supervisée en utilisant des critères temporels. Pour autant qu'on sache, le regroupement d'entités pour la classification de documents vitaux n'a pas été exploré. Cependant, les hypothèses que nous formulons (les paramètres et

les critères du modèle de classification utilisés pour filtrer les documents) peuvent être comparé aux travaux sur l'extraction des célébrités (Forestier *et al.*, 2012) ou sur le regroupement de profils utilisateurs sur les réseaux sociaux (Cantador et Castells, 2006).

3. Regrouper les entités sur les aspects temporels

Pour regrouper les entités, nous utilisons différents coefficients de corrélation calculés sur un ensemble de critères temporels. À chaque entité e est associé un ensemble de coefficients de corrélation $\rho_{e,c_1}, ..., \rho_{e,c_n}$. La corrélation de Pearson (équation 1) utilise deux vecteurs X et $,Y$, covariance $cov(X,Y)$ et les écarts type σ_X et σ_Y. Chaque coefficient de corrélation $p_{e,c}$ est calculé à l'aide de deux vecteurs : $V_{e,c}$ et A_e. Le premier vecteur $V_{e,c}$ contient les différentes valeurs calculées $v_1, ..., v_i \in V_{e,c}$ pour un critère c sur chacun des documents $d_1, ..., d_i$ annotés pour l'entité. Le vecteur A_e contient les différentes valeurs d'annotations $a_1, ..., a_i$ pour chacun des documents $d_1, ..., d_i$.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad [1]$$

Considérons un flux de documents D où tous les documents d ont une date t_d. les différents critères temporels que nous utilisons considèrent une série temporelle X basée sur l'apparition des documents. Les critères temporels que nous utilisons pour le calcul des valeurs des vecteurs $V_{e,c}$ sont les suivants :

La détection de rafale de Kleinberg : est un algorithme qui permet la détection d'un phénomène anormal sur une série temporelle (exemple d'un buzz). Lorsqu'un nombre anormal de documents apparaissent, on dit qu'il y a une rafale (burst) de documents. L'algorithme de Kleinberg (Kleinberg, 2002) permet de détecter la force d'une rafale à un instant t. La valeur associée à un document correspond à la force de la rafale à la date où le document apparaît. Pour calculer la valeur, toute la série temporelle jusqu'à la date du document est utilisée.

Le kurtosis : donne une idée de l'aplatissement d'une courbe. Cette mesure est très utilisée dans l'analyse temporelle pour détecter des anormalités sur la série. Plutôt que de considérer toute la série temporelle, nous considérons seulement une partie correspondant à la dernière semaine d'observation avant l'apparition du document.

Nombre de documents moyen par jour : donne la valeur moyenne du nombre de documents qui apparaissent chaque jour jusqu'a la date d'apparition du document analysé.

Nous utilisons ensuite l'algorithme *k-means* pour estimer les groupes pour chacune des entités. Cet algorithme permet de regrouper, en k groupes, un ensemble d'éléments en se basant sur des observations statistiques. Il corresponds parfaitement à notre besoin pour cette étude. Selon les entités, les phénomènes temporels peuvent varier. Les différentes catégories d'entité que l'on peut imaginer correspondent par exemple à : très populaire, populaire, commune et inconnu. On déduit intuitivement 3 ou 4 catégories qui seraient une première estimation pour le nombre de groupes k utilisé dans l'algorithme. (Chiang et Mirkin, 2010) montrent différentes manières d'estimer k. La règle du pouce (Rule of Thumb) défini par l'équation 2 est proportionnelle au nombre de points n. Ici on peut difficilement dire que le concept de chaque groupe est corrélé au nombre d'entités. La méthode du coude (elbow method) consiste à tracer un graphique avec en abscisse le nombre expérimental de clusters, et en ordonnée la valeur des sommes des carrés de chacun des groupes. Le meilleur k estimé par cette méthode se trouve à l'endroit de la courbe où il se forme un coude. La figure 2 montre un exemple sur les entités de KBA 2013. Le coude n'est pas clair, bien qu'il est possible de voir qu'après $k = 8$ la courbe est relativement stable. Nous n'utiliserons pas cette méthode pour estimer le nombre de clusters puisque nous ne trouvons pas de signification pour chaque cluster. Par ailleurs, avoir trop de groupes peut impliquer de ne pas avoir assez de données d'entrainement par groupe d'entités.

$$k \approx \sqrt{n/2} \qquad [2]$$

Pour les expérimentations, nous définissons, de manière empirique, le nombre de groupes pour les entités de KBA 2013 et KBA 2014. Pour connaître la cohérence des groupes, nous utilisons la représentation silhouette. Une mesure est calculée pour chaque observation pour voir si l'entité est cohérente par rapport aux autres entités de son groupe. Plus les valeurs sont hautes et uniformes, plus le groupe est cohérent. Pour les entités issues de KBA 2013, on peut voir très clairement sur la figure 3 que les groupes 2 et 4 sont les plus cohérents. Sur la figure 4, nous avons tracé la sortie de l'algorithme k-means pour ces mêmes entités. La visualisation graphique confirme bien les 2 groupes cohérents d'après la densité de points autour du centroid. Le groupe 4 semble plus épars, il est aussi moins peuplé.

Pour les entités de KBA 2014, nous utilisons $k = 3$ plutôt que $k = 4$ car le graphique de silhouette donne de meilleurs scores moyens de cohérence pour chacun des groupes ($0,680,32$ et $0,89$). Nous voyons sur la figure 5 que le groupe 3 semble avoir une forte inertie. Les deux autres groupes sont plus épars.

4. Filtrer les documents centrés sur une entité à l'aide de stratégie de regroupement

Les groupes d'entités ne sont pas distribués de manière équitable en terme de nombre d'entités, et de nombre de documents annotés. Par ailleurs, certains groupes sont plus cohérents que d'autres. Avoir de la cohérence entre les différentes entités

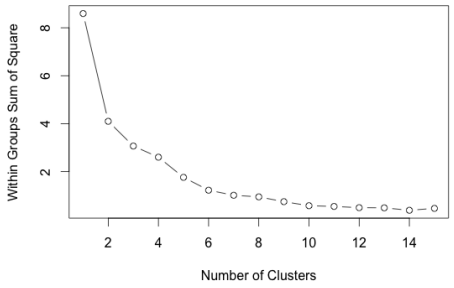

Figure 2. *Graphique résultant de la méthode du coude (Elbow Method) pour les entités de KBA 2013.*

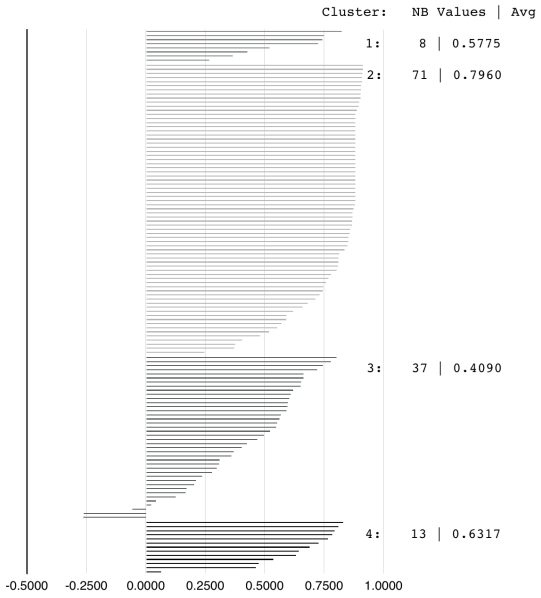

Figure 3. *Représentation silhouette avec $k = 4$ nombre de groupes pour les entités de KBA 2013. Plus les valeurs sont hautes et uniformes, plus le groupe est cohérent.*

d'un groupe est très important afin de construire un modèle de classification plus robuste. Nous proposons alors différentes stratégies pour la classification :

- 1 modèle par groupe : consiste à construire un modèle de classification pour chacun des groupes. Chaque nouveau document découvert pour une entité sera classé selon le modèle du groupe de l'entité.

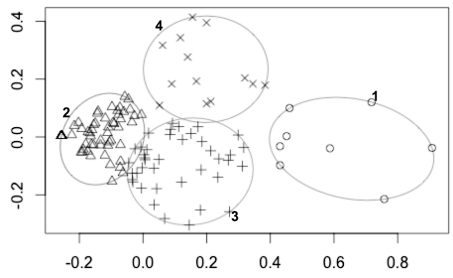

Figure 4. *Sortie de l'algorithme k-means pour les entités de KBA 2013.*

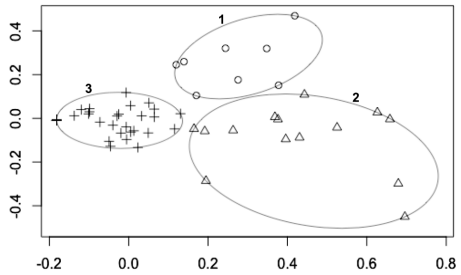

Figure 5. *Sortie de l'algorithme k-means pour les entités de KBA 2014.*

- Sélection Hybrid du Model : consiste à construire un modèle de classification global (qui utilise toutes les données d'entraînement) qui sera utilisé pour les groupes les moins cohérents. Pour les groupes les plus cohérents, un modèle de classification est déduit à partir des données annotées du groupe seulement. Pour savoir si un groupe est cohérent ou non, nous proposons d'utiliser la représentation silhouette et un seuil qui devra être estimé. Nous proposons une estimation dans le tableau 1.

Range	**Interpretation**
0.71-1.0	fortement cohérent
0.51-0.70	Cohérent
0.26-0.50	la cohérence est faible voir artificiel
< 0.25	pas cohérent du tout

Tableau 1. *Interprétation des valeurs de la représentation Silhouette*

5. Expérimentations

Pour les expérimentations, nous avons utilisé un système qui s'inspire des travaux de (Bouvier et Bellot, 2014) pour la classification des documents. Nous avons ajouté une couche supplémentaire qui permet la gestion des groupes d'entités. Afin de pouvoir observer si le regroupement d'entités est utile ou non, nous confronterons les résultats que nous obtenons sans utiliser de regroupement et avec regroupement (selon les deux stratégies énoncées en section 4).

Dans un premier temps nous comparons les résultats de notre système par rapport à ceux présentés à KBA en 2013. Nous voyons dans le tableau 2 que nous surpassons significativement les scores de KBA 2013 par $10,27\%$ pour notre meilleur système. À l'heure actuelle nous ne pouvons pas nous comparer aux systèmes de KBA 2014 puisque les résultats officiels ne sont pas disponibles. Concernant l'apport du regroupement sur l'approche sans regroupement, nous obtenons une augmentation de $+4,47\%$. Nous considérons une amélioration comme significative à partir de 5%. On observe ici que la méthode hybride semble être plus efficace que la méthode qui consiste à avoir un modèle pour chacun des groupes.

Systèmes	**F mesure**
Sans Cluster	0,380
1 modèle par groupe	0,392
Sélection hybride du modèle	**0,397**
Median KBA 2013	0,201
Meilleur KBA 2013	0,360

Tableau 2. *Scores obtenus sur les entités de KBA 2013*

Concernant les entités de KBA 2014, l'apport du regroupement semble plus important encore. En effet, un apport de $+8,07\%$ est observé pour la stratégie de regroupement hybride sur la stratégie sans cluster (cf., tableau 3).

Systèmes	**F mesure**
Sans Cluster	0,347
1 modèle par groupe	0,355
Sélection hybride du modèle	**0,375**

Tableau 3. *Scores obtenus sur les entités de KBA 2014*

De manière globale, les scores obtenus confortent l'idée qu'il est important d'avoir une cohérence entre les entités regroupées pour la classification. Par ailleurs, à défaut d'avoir assez d'entraînement pour un groupe, il est préférable d'utiliser un modèle global, plutôt que d'essayer de faire un modèle pour un groupe qui n'est pas cohérent.

Le système de classification donne pour chacun des documents une classe mais également un score de confiance pour cette classe. Pour calculer le score global du

système, le scorer officiel calcule un score de f-mesure moyen à différent pallier de confiance. Le meilleur score correspond au score officiel. Cette stratégie est tout à fait valable lorsque les entités ne sont pas dissociées, en revanche, dans notre système nous pouvons imaginer avoir différents seuils de confiance pour maximiser le score de chacun des groupes et ainsi maximiser les scores globaux du système. Ainsi nous obtenons les scores répertoriés dans le tableau 4 les entités KBA 2013 et 2014.

Sur les données de KBA 2013, l'utilisation d'un score de confiance par cluster apporte une amélioration significative de $+32,89\%$ sur la stratégie sans cluster pour le système 1 modèle par groupe. Sur les données de KBA 2014, cette stratégie apporte une amélioration de $+20,75\%$ sur le même système. Cette fois-ci, on voit que le système qui consiste à avoir un modèle par groupe semble plus efficace.

Systèmes KBA 2013	**F mesure**
1 modèle par groupe	**0,505**
Sélection hybride du modèle	0,465
Systèmes KBA 2014	**F mesure**
1 modèle par groupe	**0,419**
Sélection hybride du modèle	0,385

Tableau 4. *Scores obtenus en maximisant le score de confiance par groupe d'entités pour les entités 2013 et 2014*

6. Conclusion et Perspectives

Pour conclure, nous avons proposé d'utiliser un système de regroupement d'entités qui permet d'exploiter mieux les aspects temporels utilisés dans le système de classification de documents centrés sur les entités. Nous avons montré que notre méthode permet d'augmenter significativement les performances des systèmes en plus de dépasser les performances établies dans l'état de l'art sur la tâche KBA 2013 de la campagne d'évaluation de TREC. Cette étude a pour but de montrer que le regroupement d'entité permet d'apporter une amélioration significative dans le contexte de filtrage de documents vitaux pour une entité.

Maintenant que nous connaissons l'apport que peut engendrer une telle méthode, nous aimerions travailler sur une méthode de regroupement qui permet de s'abstraire de la supervision imposée dans notre proposition. Nous aimerions pouvoir affecter une entité à un groupe en temps réel et cela de manière automatique et non supervisée.

7. Bibliographie

Balog K., Ramampiaro H., Takhirov N., Nørvåg K., « Multi-step classification approaches to cumulative citation recommendation », *Proceedings of OAIR '13*, p. 121-128, 2013.

Bellogín A., Gebremeskel G., « CWI and TU Delft Notebook TREC 2013 : Contextual Suggestion, Federated Web Search, KBA, and Web Tracks », *The Twenty-Second Text REtrieval Conference (TREC 2013) Proceedings*, p. SP 500-302, 2014.

Bonnefoy L., Bouvier V., Bellot P., « A weakly-supervised detection of entity central documents in a stream », *The 36th International ACM SIGIR '13, Dublin*, ACM, p. 769-772, 2013.

Bouvier V., Bellot P., « Filtering Entity Centric Documents using Profile Update and Random Forest Classification », *The Twenty-Second Text REtrieval Conference (TREC 2013) Proceedings*, p. SP 500-302, 2014.

Cantador I., Castells P., « Multilayered semantic social network modeling by ontology-based user profiles clustering : Application to collaborative filtering », *Managing Knowledge in a World of Networks*, Springer, p. 334-349, 2006.

Chiang M. M., Mirkin B., « Intelligent Choice of the Number of Clusters in *K*-Means Clustering : An Experimental Study with Different Cluster Spreads », *J. Classification*, vol. 27, n° 1, p. 3-40, 2010.

Dietz L., Dalton J., « UMass at TREC 2013 Knowledge Base Acceleration Track », p. SP 500-302, 2014.

Efron M., « The University of Illinois' Graduate School of Library and Information Science at TREC 2013 », *The Twenty-Second Text REtrieval Conference (TREC 2013) Proceedings*, p. SP 500-302, 2014.

Forestier M., Velcin J., Stavrianou A., Zighed D., « Extracting Celebrities from Online Discussions », *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, IEEE Computer Society, p. 322-326, 2012.

Frank J. R., Kleiman-Weiner M., Roberts D. A., Niu F., Zhang C., Ré C., Soboroff I., « Building an entity-centric stream filtering test collection for TREC 2012 », *The Twenty-First Text REtrieval Conference (TREC 2012) Proceedings*, p. SP 500-298, 2012.

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.