Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques

Tulechki Nikola[1,2] Tanguy Ludovic[1]
(1) CLLE-ERSS : CNRS et Université de Toulouse 2, 5 allées Antonio Machado, 31058 Toulouse CEDEX 9
(2) Conseil en Facteurs Humains, 4 impasse Montcabrier. 31500 Toulouse
{tanguy,tulechki}@univ-tlse2.fr

RÉSUMÉ

Cet article étudie le lien entre la similarité textuelle et une classification extrinsèque dans des collections de rapports d'incidents aéronautiques. Nous cherchons à compléter les stratégies d'analyse de ces collections en établissant automatiquement des liens de similarité entre les documents de façon à ce qu'ils ne reflètent pas l'organisation des schémas de codification utilisés pour leur classement. Afin de mettre en évidence les dimensions de variation transversales à la classification, nous calculons un score de dépendance entre les termes et les classes et excluons du calcul de similarité les termes les plus corrélés à une classe donnée. Nous montrons par une application sur 500 documents que cette méthode permet effectivement de dégager des thématiques qui seraient passées inaperçues au vu de la trop grande saillance des similarités de haut niveau.

ABSTRACT

Deletion of dimensions of textual similarity for the exploration of collections of accident reports in aviation

In this paper we study the relationship between external classification and textual similarity in collections of incident reports. Our goal is to complement the existing classification-based analysis strategies by automatically establishing similarity links between documents in such a way that they do not reflect the dominant organisation of the classification schemas. In order to discover such transversal dimensions of similarity, we compute association scores between terms and classes and exlude the most correlated terms from the similarity calculation. We demonstrate on a 500 document corpus that by using this method, we can isolate topics that would otherwise have been masked by the dominant dimensions of similarity in the collection.

MOTS-CLÉS : similarité textuelle, classification de documents, corpus spécialisé.

KEYWORDS: textual simliarity, document classification, specialised corpora.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 439–446,
Grenoble, 4 au 8 juin 2012.

1 Introduction et contexte applicatif

Dans toute industrie à risque, le retour d'expérience (REX) occupe une place capitale dans les mécanismes de gestion de la sûreté. Des politiques de recueil, d'analyse et de stockage sont mises en place afin de garder une trace de tout évènement qui s'écarte de la norme, de tout incident ou accident qui survient lors des opérations. Les informations ainsi recueillies servent ensuite de support aux experts de sûreté pour mettre à jour les règles et les procédures d'exploitation en les adaptant à un contexte en perpétuelle évolution.

1.1 Texte et codification des rapports d'incidents aéronautiques

L'objet de notre étude est un sous-ensemble particulier de REX, les rapports de type Aircraft Safety Report (ASR) recueillis dans le service de sécurité de la compagnie aérienne Air France. Les ASR sont des textes relativement courts (105 mots en moyenne) rédigés par les pilotes eux-mêmes immédiatement ou peu après qu'un incident s'est produit, et décrivant celui-ci en langage libre. Lorsqu'ils sont soumis, ces rapports sont saisis dans la base de données de la compagnie et enrichis d'un certain nombre d'informations factuelles, telles que le modèle de l'avion, les conditions météo, la localisation ou encore le poids de l'appareil le jour de l'incident.

Ensuite les rapports subissent une première analyse visant à "coder" l'évènement suivant un schéma préétabli. Un schéma de codification est une abstraction d'un scénario d'accident, composée de plusieurs taxonomies de codes en rapport avec différents aspects d'un accident. En pratique, l'expert en charge du codage doit décrire l'évènement en utilisant quelques centaines de codes, à partir de listes fermées. (Voir (Ponvert, 2009) pour les détails de l'élaboration et la mise en place du schéma de codification actuel d'Air France). Une fois codés, les rapports sont stockés dans la base de données et peuvent être interrogés *via* des requêtes portant sur les informations factuelles et la codification. Un expert peut ainsi, par exemple, extraire de la base l'ensemble d'incidents, où il y a eu une panne du radar météo, dans un Boeing 747 survenue lorsque l'avion était en phase de montée initiale.

1.2 Limites de la codification

Avec du recul, on peut voir le procédé de codification comme un effort visant à maîtriser la variation inhérente des rapports afin d'atteindre un niveau d'abstraction suffisamment stable pour une exploitation informatisée d'une base de REX. Sans rentrer dans les détails, nous dirons que cet effort est nécessairement accompagné d'un appauvrissement du contenu informationnel directement accessible aux experts. Le fait de réduire un texte à un squelette prédéterminé a pour effet de ne garder que les éléments les plus saillants de l'évènement au détriment de subtilités qui, tout en étant présentes dans le texte original, ne trouvent pas leur place dans la codification.

Une autre limite de ces stratégies est leur caractère intrinsèquement réactif. Un schéma de codification est une représentation de la réalité figée à un instant précis, alors que la réalité qu'elle reflète est en perpétuelle évolution. Toute changement majeur du contexte doit être reflété dans le schéma, ce qui correspond à un effort considérable et prends un temps précieux aux experts, pendant lequel un risque nouvellement apparu peut se trouver sans code associé.

1.3 Objectifs applicatifs

Conscients des limites des stratégies d'analyse des REX par codification, notre objectif est de concevoir des techniques et outils venant en complément de ces stratégies et permettant aux experts d'explorer les collections de rapports en fonction des particularités de leur contenu textuel et de leur distribution chronologique. S'affranchissant de la rigidité de la codification, dans l'idéal ces outils devront être capables d'alerter leurs usagers de configurations particulières d'évènements, de tendances émergentes ou encore d'évènements anormaux (Tulechki, 2011).

2 Similarité textuelle

Dans un premier temps nous avons cherché à rapprocher les textes en fonction de leur contenu en utilisant des méthodes classiques en recherche d'information (RI) : la similarité cosinus (Salton *et al.*, 1975), une mesure du recouvrement lexical qui attribue un score compris entre 0 et 1 à chaque paire de documents dans la collection. Un score de 0 signifie une absence de termes en commun et un score de 1 une identité complète du contenu lexical des deux textes. Ce score est obtenu en calculant le cosinus entre deux vecteurs dans un espace à n dimensions correspondant aux termes présents dans la collection.

En plus de son utilisation immédiate dans des applications de type moteur de recherche, ce calcul permet de superposer automatiquement une couche de structure sur une collection et transformer un matériau symbolique et qualitatif en des données numériques et ouvre la voie à d'autres traitements comme l'apprentissage non supervisé (Steinbach *et al.*, 2000), ou encore la détection d'anomalies (Chandola *et al.*, 2009) pour en citer quelques uns. Toutefois à l'heure actuelle, nous avons préféré tout d'abord évaluer l'apport *per se* de la similarité textuelle en développant un outil utilisant ce calcul et en le soumettant aux experts de sûreté.

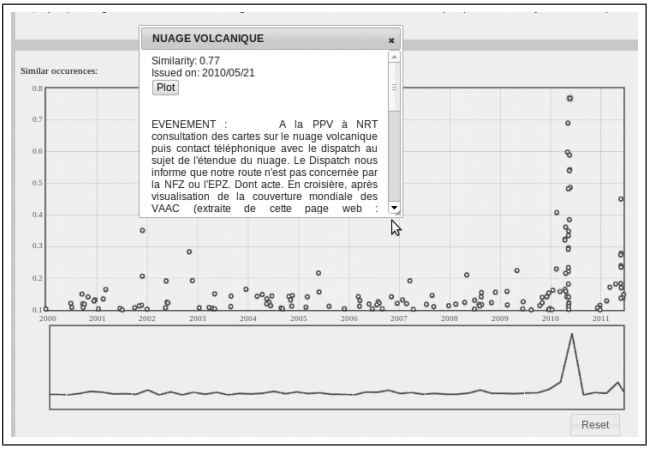

FIGURE 1 – Similarité de rapports d'incidents de sur un axe chronologique

L'outil *timePlot* présenté en figure 1 permet, à partir d'un rapport pivot, de visualiser les rapports similaires et leur distribution dans le temps. Les rapports sont présentés à l'utilisateur sur un graphique interactif qui permet un accès direct à leur contenu. Différentes configurations chronologiques peuvent apparaître, comme ici le pic de rapports associés à l'éruption volcanique du printemps 2010, ou encore des phénomènes saisonniers, comme par exemple des incidents liés à la neige qui, naturellement, apparaissent dans les périodes de grand froid.

2.1 Limites de la similarité

L'approche par similarité textuelle a connu un succès auprès des experts qui ont apprécié son coté intuitif et le potentiel de rapprochement de rapports dont le lien n'est nullement reflété par un codage commun. La logique d'utilisation simplifiée et l'intuitivité de l'interface, conçue d'emblée comme un support à une exploration de la collection sans a priori, ont aussi contribué à la validation de cette approche par ses usagers.

Néanmoins, nous nous sommes vite aperçus que le calcul de similarité, dans notre contexte précis et compte tenu des spécificités du matériau textuel auquel nous avons affaire, souffrait d'un manque de finesse évident. Étant donné que l'ensemble des textes sont issus d'un domaine très circonscrit, celui de l'aviation civile, ils sont tous plus ou moins similaires, la plupart parlant d'"avions", de "pilotes" et de "vols". Ce fond lexical commun est en partie géré par les techniques de pondération, comme le TF/IDF (Spärck-Jones, 2004), mais compte tenu de la variation lexicale inhérente du domaine, notamment la multitude de termes [1] désignant un même objet qui sont employés par les rédacteurs, des rapprochements sont faits sans pour autant designer des facteurs de similarité pertinents pour une analyse.

Une autre limite directement liée à la visée applicative de nos travaux apparaît aussi. On s'aperçoit de l'existence d'un fort recouvrement entre le codage des rapports et les regroupements mises à l'évidence par le calcul de similarité textuelle. Un lien entre le termes du texte et leur codage existe dans le corpus. Il est clair que dans les rapports parlant de chocs aviaires [2], on trouve des termes comme "oiseau", "mouette", et "aviaire". Un rapprochement basé sur ces termes retrouvera plus ou moins la catégorie *choc aviaire* qui est déjà mise en évidence dans le codage. Dit autrement, la similarité textuelle à tendance de retrouver les dimensions les plus saillantes dans le corpus, dimensions qui sont pour la plupart déjà bien identifiées et reflétées dans les schémas de codification.

Par ailleurs, un système de classification automatique de ces données, permet déjà, sur la base du contenu textuel, de proposer des codes aux experts (Hermann *et al.*, 2008). Or, un de nos objectifs est notamment de chercher des facteurs communs plus subtils, pouvant rapprocher des incidents sur des critères différents et transversaux au codage.

3 Dimensions de la similarité textuelle

La similarité textuelle, telle qu'elle est calculée, représente toute parenté qui peut exister entre deux textes sur une dimension unique, sans tenir compte des multiples facteurs qui peuvent contribuer à cette parenté.

1. Pour parler du pilote, on trouve dans les textes le terme "pilote", mais aussi un ensemble de termes et d'acronymes spécifiques au domaine comme "cdb" (commandant de bord), "opl" (officier pilote de ligne), "copilote", "copi", "pf" (pilot flying), "pnf" (pilote not flying) etc.
2. Il arrive très fréquemment que les avions percutent des oiseaux.

Une des thématiques actuelles en recherche d'information est de regrouper les résultats des moteurs de recherche par thème en utilisant des méthodes de *clustering*, afin de mettre en évidence les différentes thématiques qui sont présentes dans la liste des résultats. Une requête comme "japon" par exemple, peut ramener des documents traitant du tourisme au japon et de gastronomie japonaise (Navarro *et al.*, 2011). En se basant sur les similitudes entre ces documents un système d'apprentissage non supervisé regroupe ensuite ces résultats en deux paquets et permet à l'utilisateur de focaliser sa recherche sur le sous-ensemble qui l'intéresse. Ces méthodes, tout en raffinant et classant les résultats ne peuvent pas encore gérer des collections où les thématiques varient simultanément sur plusieurs dimensions. Les résultats sur le japon peuvent concerner des localisations différentes ("Tokyo" et "Osaka", par exemple) sans qu'une localisation soit particulièrement associée à un des thèmes. Un système de clustering peinera à isoler ces deux dimensions de variation et à proposer un découpage des résultats selon les deux critères (thème et localisation) simultanément. De travaux sont en cours visant à développer des méthodes efficaces de clustering avec recouvrement [3], notamment pour répondre à l'unidimensionalité des techniques actuelles.

Principalement orientées vers un usage dans un moteur de recherche "classique" et sur des collections larges de textes hétérogènes, ces méthodes n'assument pas une organisation à priori de la collection. Or dans un corpus spécialisé, comme les bases de rapports d'incidents, les schémas de codification visent justement à organiser la collection, de façon pertinente compte tenu de spécificités de son contexte d'utilisation, tout en intégrant l'hétérogénéité des facteurs de similarité et en représentant. Illustrons ceci par les trois exemples suivants que nous avons construit en nous inspirant de textes réels intitulés comme suit :

```
1) Choc aviaire au décollage.
2) Turbulences au décollage.
3) Choc aviaire à l'atterrissage.
```

Entre ces trois textes un score de similarité comparable sera calculé entre 1) et 2) et entre 1) et 3). Pourtant, les raisons de ce rapprochement sont différentes dans les deux cas. 1) et 3) traitent d'un même type d'incident, alors que 1) et 2) partagent les circonstances dans lesquels sont survenus des incidents différents. Ces deux aspects sont pris en compte dans le schéma de codification, grâce aux champs "type d'incident" et "phase de vol". Le "type d'incident" pour 1) et 3) sera *choc aviaire* et *turbulences* pour 2). La "phase de vol" sera *décollage* pour 1) et 2) et *atterrissage* pour 3). Nous allons donc regarder de près comment mettre en évidence le lien entre le codage des rapports et leur contenu textuel.

3.1 Lien entre codage et contenu

Nous avons déjà vu que certains termes des textes étaient fortement liés à certaines classes du schéma de codification et que ces mêmes termes font en sorte que la similarité textuelle retrouve souvent les classes du schéma.

Afin d'étudier ce lien, nous avons constitué un corpus de test en prenant des rapports traitant de chocs aviaires et de turbulences, survenus lors de l'atterrissage et lors du décollage, de manière à avoir une collection équilibrée que nous savons varier sur deux dimensions, la phase de vol et le type d'incident. Le corpus est constitué de 482 rapports que nous avons choisis en nous basant sur le codage de leur champs *type d'incident* et *phase de vol* :

3. Concrètement,une telle méthode doit être capable de classer un même document dans plusieurs classes, en fonction de critères de rapprochement différents.

	Turbulences	Choc aviaire	Total
Atterrissage	118	133	251
Décollage	107	124	231
Total	225	257	482

Le premier test a consisté à mesurer le degré de recouvrement entre similarité et catégorisation dans le corpus. Pour cela nous avons, pour chaque document, automatiquement sélectionné les 30 documents les plus similaires et, pour chacun de ces documents, testé s'il partage la même valeur pour les champs *type d'incident* et *phase de vol*. En moyenne, 89% des documents partagent la catégorie et 75% des documents partagent la phase de vol, alors que si aucun lien entre codage et similarité n'existait, nous nous attendrions à ce que ces valeurs avoisinent les 50%.

3.2 Effacement des dimensions principales

Afin d'isoler les dimensions de similarité, nous avons tout d'abord cherché les termes qui sont les plus liés à chacune d'entre elles en utilisant une mesure d'interdépendance statistique : l'information mutuelle (IM), en nous inspirant des techniques de sélection de traits utilisées en recherche d'information (voir [4] (Manning *et al.*, 2008, Section 13.5.1) pour l'algorithme utilisé). En RI cette technique permet, pour une collection catégorisée, de réduire l'espace des termes en ne sélectionnant que ceux qui sont statistiquement corrélés à une classe donnée. L'IM est aussi communément utilisée en classification automatique. Étant donné un terme t et une classe C, plus l'information mutuelle $IM(t,C)$ est élevée, plus t permet de correctement prédire C.

Pour la RI, l'hypothèse sous-jacente qui justifie ce procédé est que ce sont typiquement les termes décrivant au mieux la variation relative à une organisation particulière repérée par un humain *via* une classification donnée qui seront aussi les plus performants pour l'indexation de la même collection. Notre objectif est exactement inverse. Nous allons exclure ces termes du calcul de similarité, afin qu'il ne reflète pas l'organisation déjà présente dans le schéma de codification.

Nous avons calculé l'IM entre tous les termes d'un corpus de 4450 rapports, et les 4 classes que nous avons isolées. Voici les 5 termes les plus corrélés par catégorie.

	Turbulences	**Choc aviaire**	**Atterrissage**	**Décollage**
1	vent	aviaire	approche	décollage
2	turbulence	collision	finale	poussée
3	gaz	oiseau	atterrissage	rotation
4	arrière	impact	stabilisation	t/o [5]
5	windshear [6] [7]	bird [7]	arrondir	vr [8]

Nous avons de nouveau mesuré la moyenne de recouvrement (MR) entre similarité et codification, mais cette fois ci en excluant soit les 50 termes les plus associés aux deux phases de vol (phVol), soit les 50 termes les plus associés aux types d'évènement (typEve). Nous avons aussi calculé un taux de perturbation (TP) en regardant, pour chaque document le nombre de documents qui apparaissent dans les 30 documents les plus similaires lors de l'application d'un filtrage.

4. version disponible en ligne à `http://nlp.stanford.edu/IR-book/`
5. Take-off (Décollage)
6. Cisaillement du vent
7. Il est très courant que des termes anglais soient employés dans ces textes pourtant écrits en français.
8. Vitesse de rotation

	MR phVol	MR typEve	TP [9]
Sans filtre	75%	89%	-
Filtre sur phVol	64%	84%	9,8
Filtre sut typEve	73%	69%	13,6

On peut voir que le recouvrement entre la similarité textuelle et une dimension donnée varie en fonction du filtrage des termes associés à cette même dimension, alors que le recouvrement sur l'autre dimension est moins affecté. Après filtrage on trouve, en moyenne, respectivement 9,8 et 13,6 nouveaux documents dans la liste des 30 premiers, ce qui témoigne de l'effet du filtrage sur le classement des résultats.

Concrètement ceci signifie que, pour un rapport traitant de turbulences au décollage, un filtrage des termes associés avec les phases de vol privilégiera les rapports traitant de turbulences alors qu'un filtrage des termes associés avec le type d'évènement privilégiera les rapports traitant d'évènements survenus lors du décollage.

3.3 Dimensions transversales

En effaçant ces dimensions de similarité, le filtrage des termes associés possède la capacité de mettre en évidence des facteurs de similarité secondaires. Voici un exemple d'une telle dimension qui a émergé de notre corpus. Le rapport suivant traite de turbulences à l'atterrissage, mais mentionne en plus un double pilotage [10] :

> INCURSION VFE SUITE CISAILLEMENT EN FINALE. [REPORT]. Fort cisaillement en finale reporté par les avions précédents. La soudaineté du phénomène surprend l'OPL PF. Légère incursion dans la VFE (3 ou 4 kts). Réponse des commandes par CDB (double pilotage pendant 1 à 2 s.). Avion stabilisé, l'OPL reprend les commandes. Atterrissage sans problème. -FIN-

Lorsque nous regardons la liste des rapports similaires sans filtre, au premiers rangs nous trouvons ceux qui évoquent des turbulences à l'atterrissage, comme par exemple :

> FORT CISAILLEMENT DE VENT EN FINALE 26R CDG. [REPORT]. FORT CISAILLEMENT DE VENT EN FINALE. -FIN-

Pour le même document, lorsque l'on filtre les termes associés avec la phase de vol et le type d'évènement, les rapports parlant uniquement de turbulences à l'atterrissage apparaissent plus bas dans la liste des rapports similaires et on retrouvera leur place ceux qui partagent des termes non associés avec les phases et les types d'évènement, notamment des rapports traitant de double pilotage, information qui n'est pas reflétée dans le codage. Ce facteur commun permet d'établir un lien entre ces deux rapports, qui dans certains cas peut s'avérer pertinent pour un expert.

> BREF DOUBLE PILOTAGE AU DECOLLAGE. [REPORT]. OPL PF au décollage. Vent travers avec rafales. Brève action réflexe en latéral du CDB pour contrer rafale et début d'inclinaison à droite. Prise de priorité peu pertinente pour effet immédiat. -FIN-

9. Les valeurs ici sont des moyennes pour les 482 documents.
10. Double pilotage signifie que les deux pilotes agissent simultanément sur les commandes de l'avion.

4 Conclusion et perspectives

La technique que nous avons présentée s'inscrit dans un effort global dont l'objectif est de proposer des outils d'exploration de collections de documents et la mise en évidence de liens de similarité "faibles" entre les documents qui seraient autrement masqués par la dimension de similarité la plus saillante. D'emblée conçues pour un usage par un utilisateur averti, notre intention est d'en évaluer l'apport applicatif en les proposant à des experts en sûreté aérienne sous forme d'un outil de visualisation et d'exploration permettant dynamiquement à l'utilisateur de choisir les dimensions à ne pas prendre en compte lors du calcul à partir d'une liste des dimensions les plus saillantes pour le sous-ensemble en cours d'analyse. Le fait de se baser sur la codification assure que les choix de filtres qu'auront les experts reflètent des concepts qu'ils ont l'habitude de manipuler dans leur activité d'analyse.

Disposant à l'heure actuelle d'une preuve de concept, nous comptons, dans les mois qui viennent, passer à l'échelle en prenant en compte l'intégralité de la codification des collections. S'agissant de techniques exploratoires et fortement dépendantes du domaine et de leur objectif applicatif précis, nous ne sommes pas en mesure de proposer un protocole d'évaluation classique et comptons sur un évaluation par l'usage et un échange constant avec les usagers pour pouvoir juger de la pertinence de ces méthodes.

Références

Chandola, V., Banerjee, A. et Kumar, V. (2009). Anomaly detection : A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.

Hermann, E., Leblois, S., Mazeau, M., Bourigault, D., Fabre, C., Travadel, S., Durgeat, P. et Nouvel, D. (2008). Outils de Traitement Automatique des Langues appliqués aux comptes rendus d'incidents et d'accidents. *In 16e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Avignon*.

Manning, C. D., Raghavan, P. et Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press, New York.

Navarro, E., Chudy, Y., Gaume, B., Cabanac, G. et Pinel-Sauvagnat, K. (2011). Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? *In Actes de Coria 2011 : Conférence en Recherche d'Information et Applications*.

Ponvert, M. (2009). Définition des besoins nécessaires à la mise en place d'un Data WareHouse dans le cadre du SGS Air France. Mémoire de D.E.A., École Nationale de l'Aviation Civile.

Salton, G., Wong, A. et Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Spärck-Jones, K. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 60(5):493–502.

Steinbach, M., Karypis, G., Kumar, V. *et al.* (2000). A comparison of document clustering techniques. *In KDD workshop on text mining*, volume 400, pages 525–526. Boston.

Tulechki, N. (2011). Des outils de TAL en support aux experts de sûreté industrielle pour l'exploitation de bases de données de retour d'expérience. *In Actes des 13èmes Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2011)*.