Similarité textuelle pour l'exploration de rapports d'incidents

UE TAL – 26/11/2012

Nikola TULECHKI

Plan

- 1. Contexte
  - CFH
  - REX et sécurité
  - Pourquoi la similarité
- 2. Similarité et temps
- 3. Calculs de similarité
- 4. Dimensions de similarité
- 5. Perspectives

Contexte > CFH (1)

- Conseil en Facteurs Humains
- Domaines de compétences :
  - Ergonomie et Facteurs Humains pour la conception et l'amélioration des processus, analyse de besoins, sécurité, implantations de nouveaux moyens, informatisation…
  - Traitement Automatique des Langues (Safety Data Analysis ™) :
    - R&D et Exploitation : Analyse et traitement des informations de sécurité aéronautiques et de REX (Catégorisation & Similarité)
    - R&D : Analyse de la qualité et de la cohérence de bases de données de Procédures Opérateurs
- CIFRE depuis 2010

Contexte > CFH (2)

- R&D et Exploitation : Analyse et traitement des informations de sécurité aéronautiques et de REX

  - **EASA** – European Aviation Safety Agency
  - **DGAC** – Direction Générale de l'Aviation Civile
  - **Air France**
  - **DSNA** - Direction des Services de la Navigation Aérienne
  - **ICAO** – International Civil Aviation Agency
  - **WFP** – World Food Program (Transporteur ONU)
  - **Astrium**
  - **IMdR** - Institut pour la Maitrise des Risques (étude TAL & REX)
  - ***BEA*** *– Bureau d'Enquêtes et d'Analyses*
  - ***Total***
  - ***SNCF***
  - ***IMO*** *– International Maritime Organisation*

Contexte > REX et Sécurité (1)

- REX (**R**etour d'**EX**périence): Toute formalisation d'un événement passé

  (= Processus et Données à la fois)
- Données REX
  - Comptes rendues d'incidents (http://asrs.arc.nasa.gov/)
  - Rapports d'accidents (http://www.bea.aero)
  - *Etudes et synthèses de sécurité (https://www.easa.europa.eu → Annual Safety Review)*
  - *Rapports maintenance*
  - *Etudes de cas*
  - *Etudes in situ*
  - *Etc…*
- Processus REX
  - Collecte → Analyse initiale → Codage → Stockage → Exploitation

Contexte > REX et Sécurité (2)

- REX dans l'aviation civile
  - *Reporting* réglementé et obligatoire
    - Collecte et traitement en interne
      - Compagnies, Aéroports, Navigation Aérienne, Clubs de vol, Particuliers, etc…
    - Soumission aux autorités (DGAC)
  - Contexte européen et international
    - Synchronisation des normes
    - Echange entre agences
      - DGAC, EASA, ICAO, FAA, TC etc..
  - Flux importants
    - 600 rap. /mois pour AF
    - 5000 rap. /mois pour la DGAC

Contexte > REX et Sécurité (3)

- Utilisation des REX
  - *1. Collecte*
  - 2. Codage
    - Schémas préétablies (*taxonomie*)
    - Informations relatives à l'incident
      - *Type d'avion, Météo, Type d'accident (> 1000 val, 4 niv) , Phase de vol, Lieu, […], âge et expérience du pilote, espèce d'oiseau percuté, fabriquant des moteurs etc..*
    - Taxonomies différentes → pb. de partage
    - Taxonomies complexes → pb. de cohérence
    - Contexte évolutif → modification constante

➔ Données bruités et impropres

Contexte > REX et Sécurité (4)

- Utilisation des REX
  - *2.Codage*
  - 3. Stockage
    - Formats différentes
      - Logiciels spécialisés (**ECCAIRS**, Sentinel, eCare)
      - BD aux schémas ad-hoc
    - Formats souvent inadaptés aux taxonomies
    - Pb de partage
      - Passerelles,
      - couches de recodage,
      - exports

      → pertes d'informations

Contexte > REX et Sécurité (5)

- Utilisation des REX
  - *3. Stockage*
  - 4. Analyse
    - Analyses quantitatives (périodiquement)
    - Analyses qualitatives (ponctuellement)
      - Basés sur des requêtes dans les bases → fortement dépendante du codage
      - Basés sur la mémoire des experts

- Améliorer la qualité du codage
  - Corrections au cas par cas
  - Vérificateurs automatiques de cohérence
  - **Codification automatique basé sur les narratifs (CFH)**

Contexte > Pourquoi la similarité (1)

- Calculer la ressemblance de deux rapports en se basant sur le texte de leur narratifs
  - Indépendante du codage (et tous ses problèmes)
  - Venir complémenter les stratégies actuelles

  - Exploitable directement (problématique de RI classique)
  - Matrice de distance → base pour des traitements comme le clustering, la recherche d'anomalies etc..

Contexte > Pourquoi la similarité (2)

- Identifier des évènements similaires, composante essentielle des tactiques employées par les experts:

> **Example 8. A series of events indicating improperly secured cargo.**
> Over a couple months, investigators noticed several similar events involving pieces of cargo in the aircraft hold being found, on arrival, to be improperly fastened down, [...]. Unrestrained cargo can be a problem if it moves around and affects the trim and handling of the aircraft. Three events had been reported in the first month, and then it went up to about seven in the second month […] These events presented a clear pattern, suggesting that something was amiss in the loading of cargo. On closer examination, investigators found that all the events could be traced back to the same terminal, reinforcing and localizing their suspicion of an underlying problem with work practices there.

(Macrae 2007)

Contexte > Pourquoi la similarité (3)

- Problématique lors du passage à l'échelle

> It can become difficult to search databases of 500,000 records to determine whether similar incidents have occurred in the past.
>
> (Johnson 2003)

Contexte > Pourquoi la similarité (4)

The Agency receives data from various sources, including ICAO, national investigation authorities, manufacturers, maintainers and other parts of the industry. The data is captured and stored in various ECCAIRS repositories.

In the treatment of the reports received, an important aspect is the determination whether the occurrence received concerns an isolated event or whether this is part of a larger number of similar events. Traditionally, such analysis was done using the knowledge of the experts concerned but also a verification based on coded data. This approach as some limitations: it relies on a solid memory of persons working in the field as well as on proper correct and complete coding of occurrences which both cannot be assured.

It would therefore assist in this work, if the present approaches could be complimented by an automated review of narratives aimed at detecting similar occurrences based on their description.

(EASA 2012) – Appel d'offres 11/2012

Plan

- 1. Contexte
- 2. Similarité et temps
  - Démo *timePlot*
  - Lessons learned
- 3. Calculs de similarité
- 4. Dimensions de similarité
- 5. Perspectives

Similarité et temps > Démo *timePlot*

- Principe général:
  - Outil d'exploration de bases d'incidents
  - Un rapport *source*
  - Identifier les rapports similaires dans la base et produire un nuage de points interactif avec la date sur l'abscisse et la similarité avec le rapport source sur l'ordonnée
  - Cliquer sur les points permet de visualiser les rapports

Similarité et temps > Lessons learned

- Si c'est beau, ca marche
- Facilité d'utilisation avant tout

- Prototype vs. version de production
  - Passage à l'échelle
  - Obligations de temps de réponse
  - Contraintes de temps, livrables, maintenance ….

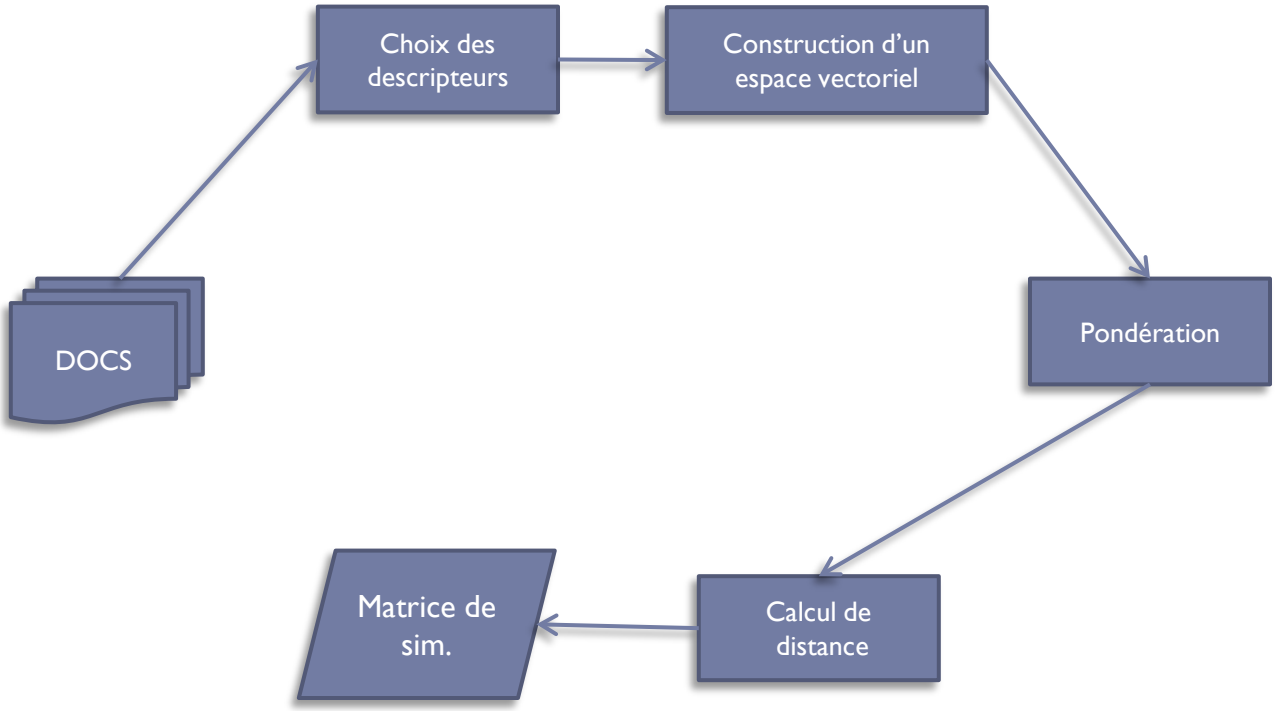
- Importance du feedback des utilisateurs (mais avec modération)

Plan

- 1. Contexte
- 2. Similarité et temps
- 3. Calculs de similarité
  - Choix des descripteurs
  - Espace vectoriel
  - Pondérations
  - Calcul de distances
- 4. Dimensions de similarité
- 5. Perspectives

Le calcul de similarité > Schéma de base

Le modèle vectoriel (Salton et al., 1975)



Le calcul de similarité > Choix des déscripteurs

- Comment représenter un document par un ensemble de traits
  - Lexique
    - Stoplists
    - Catégories pleines
  - Morphologie
    - Lemmatisation vs stemming
    - Analyseurs dérivationnels
  - Syntaxe
    - Termes complexes
    - Colocations vs analyse syntaxique
    - Termes complexes vs termes structurés
  - Sémantique
    - Indexer des "concepts"
    - Ressources externes
    - Apprentissage endogènes
- Choix conditionnent la suite

(Morreau & Sebilliot, 2005)

Le calcul de similarité > Espaces vectoriels

- Espace à n-dimensions

d1 “A fire damaged the shipment of gold”

d2 “A delivery of silver arrived in a silver truck”

d3 “The shipment of gold arrived in a truck”

Terms		d1	d2	d2
a	A =	1	1	1
arrived		0	1	1
damaged		1	0	0
delivery		0	1	0
fire		1	0	0
gold		1	0	1
in		1	1	1
of		1	1	1
shipment		1	0	1
silver		0	2	0
truck		0	1	1

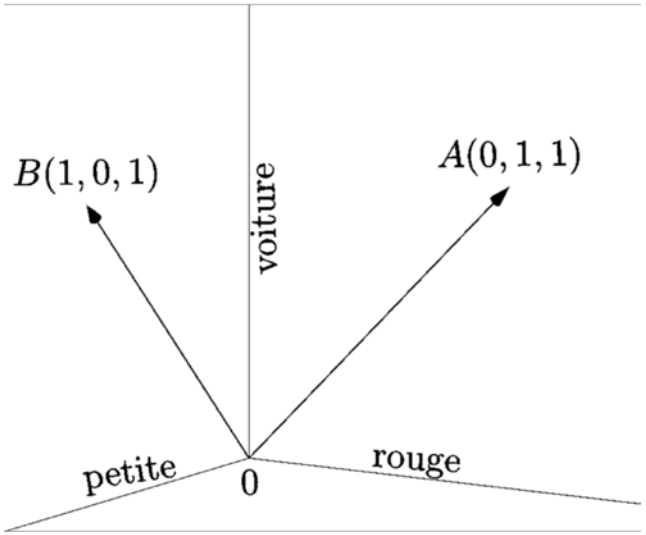
Le calcul de similarité > Pondérations

- Importance relative des descripteurs par rapport aux documents
  - Fréquence du terme dans le document
  - Fréquence du terme dans le corpus
  - Longueur du document
  - Taille du corpus
- TF/IDF
- DFR
- BM25
- IM
- etc…

(Tirilly, 2010)

Le calcul de similarité > Calcul de distance

- Cosinus



A: "petite voiture"
B: "voiture rouge"

$$sim(A,B) = \frac{x_1x_2 + y_1y_2 + z_1z_2}{\sqrt{x_1 + y_1 + z_1}\sqrt{x_2 + y_2 + z_2}} = \frac{0+0+1}{\sqrt{2}\sqrt{2}} = \frac{1}{2} = 0.5$$

Plan

- 1. Contexte
- 2. Similarité et temps
- 3. Calculs de similarité
- 4. Dimensions de similarité
  - Problématique
  - Isoler les dimensions
  - Effacer les dimensions
  - Tests en corpus
  - Conclusion
- 5. Perspectives

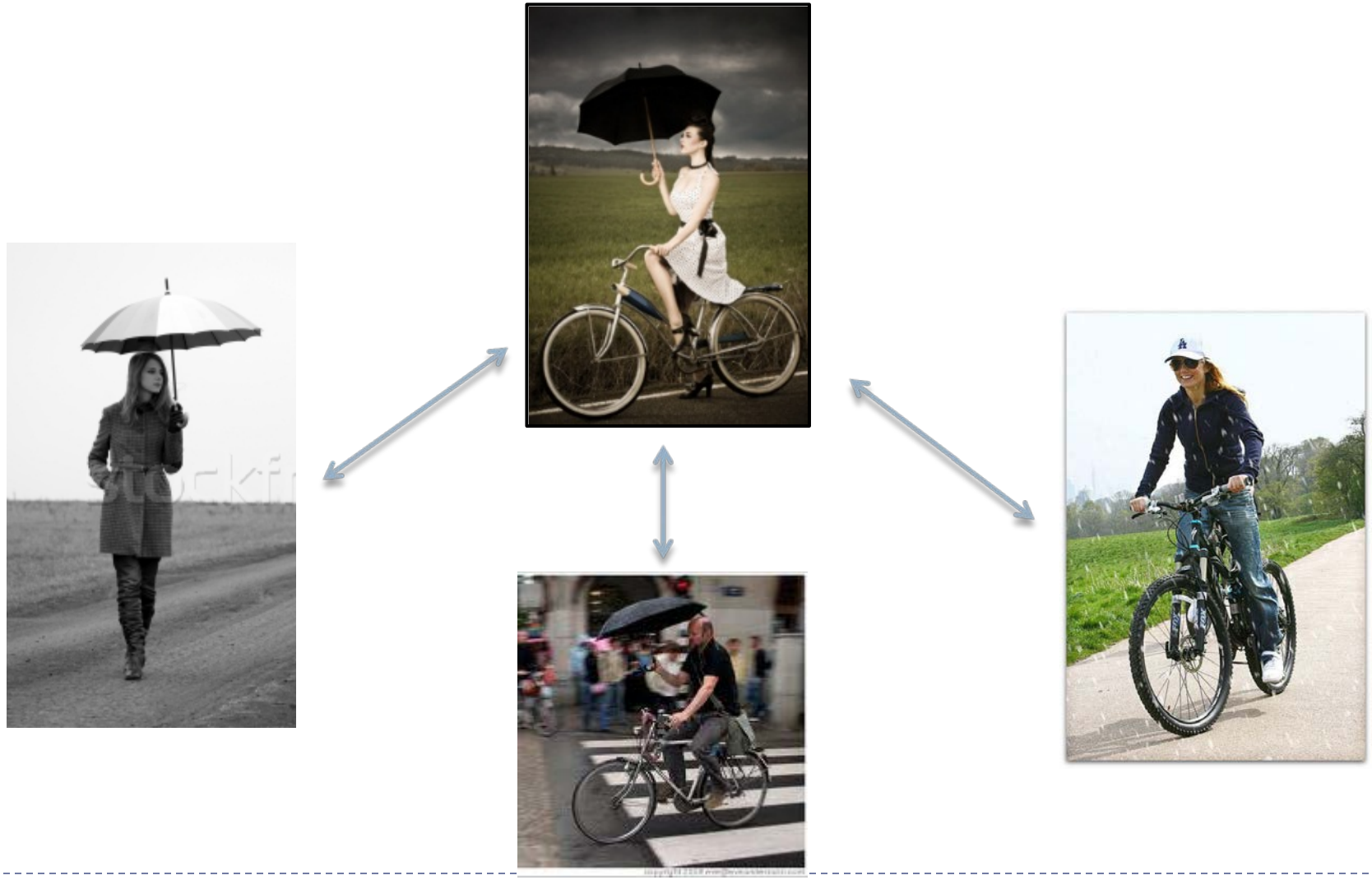
Effacement de dimensions de similarité > Problématique de base

- Score de similarité unidimensionnelle
- Réalité plus complexe, multifacette
  - Dimensions principales
  - Dimensions secondaires, transversales, moins saillantes

- Isoler et pouvoir sélectionner les dimensions à prendre en compte dans le calcul

(Tulechki et Tanguy, 2012)

Effacement de dimensions de similarité > Exemple



Effacement de dimensions de similarité > Quelles dimensions

- Rapports codifiés
  - Phase de vol (Atterrissage, Décollage)
  - Type d'évènement (Choc Aviaire, Turbulences)

- Corpus de test équilibré

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

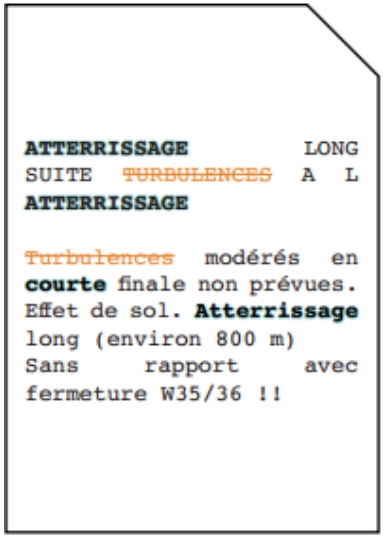
- Fort recouvrement entre similarité et codification
  - 89% Type
  - 75% Phase

Effacement de dimensions de similarité > Isoler les dimensions

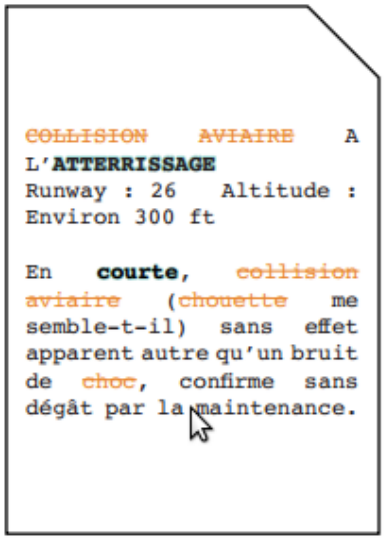
- Identifier les termes les plus associés aux classes
  - Information mutuelle

	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]

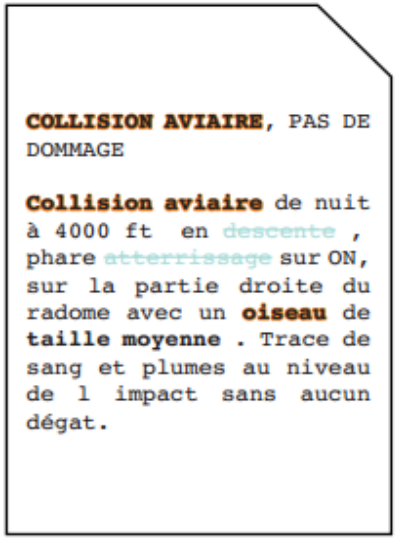
- Effacer ces termes de l'espace termes lors du calcul (stoplists)


ATTERRISSAGE LONG SUITE ~~TURBULENCES~~ A L **ATTERRISSAGE**

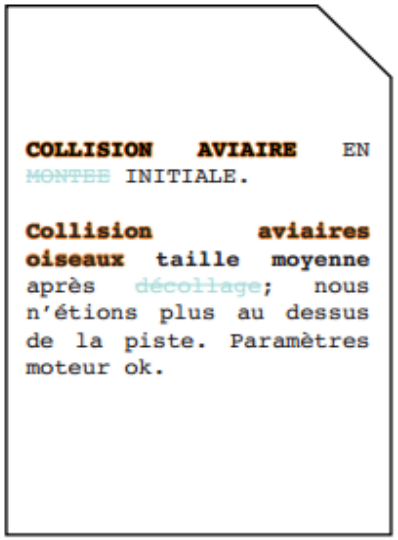
~~Turbulences~~ modérés en **courte** finale non prévues. Effet de sol. **Atterrissage** long (environ 800 m) Sans rapport avec fermeture W35/36 !!


~~COLLISION~~ ~~AVIAIRE~~ A L'**ATTERRISSAGE**
Runway : 26 Altitude : Environ 300 ft

En **courte**, ~~collision aviaire~~ (~~chouette~~ me semble-t-il) sans effet apparent autre qu'un bruit de ~~choc~~, confirme sans dégât par la maintenance.


COLLISION AVIAIRE, PAS DE DOMMAGE

Collision aviaire de nuit à 4000 ft en ~~descente~~ , phare ~~atterrissage~~ sur ON, sur la partie droite du radome avec un **oiseau** de **taille moyenne** . Trace de sang et plumes au niveau de l impact sans aucun dégat.


COLLISION AVIAIRE EN ~~MONTEE~~ INITIALE.

Collision aviaires oiseaux taille moyenne après ~~décollage~~; nous n'étions plus au dessus de la piste. Paramètres moteur ok.

Effacement de dimensions de similarité > Tests en corpus

- Moyenne de Recouvrement (MR) au rang 30
- Taux de perturbation (TP) au rang 30

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

Effacement de dimensions de similarité > Dimensions sous-jacentes

- Effacer les termes associés aux 4 classes

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

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

Sans Filtre

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-

Avec Filtre

Effacement de dimensions de similarité > Conclusions

- Passer à l'échelle (toutes les classes)
  - Problème de l'interdépendance et la forte disproportion des classes.
- Non pertinence de la valeur de Information mutuelle
  - Pouvoir « calmer » les termes plutôt que les effacer
- Intégrer à l'outil d'analyse de similarité
  - Informer l'utilisateur des classes les plus représentés dans les résultats
  - Donner la possibilité à l'utilisateur de paramétrer le calcul

Plan

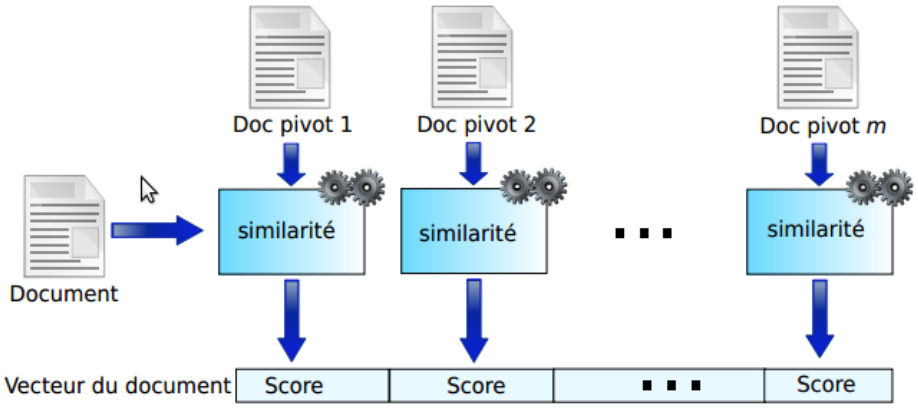
- 1. Contexte
- 2. Similarité et temps
- 3. Calculs de similarité
- 4. Dimensions de similarité
- 5. Perspectives

  - Similarité de second ordre
  - Calculs transparentes et paramétrables
  - Evaluations?

Perspectives
> Sim² > Principe

- Similarité de 2e ordre
  - Pour chaque document, calculer une similarité "classique" avec *m* documents pivots.
  - Indexer les documents avec des vecteurs à *m* dimensions dans un espace "sémantique"

  - (Claveau, 2012) : "Vectorisation" → Pivots choisis dans la collection
  - (Gabrilovich & Markovitch, 2007) :"ESA" → Pivots articles de Wikipédia (+ concepts Wikipédia)



Perspectives
> Sim²> Avantages

- Meilleurs performances sur différentes tâches
  - RI, *Amaryllis*
  - Fouille de Texte , *DeFT 2011*
  - Segmentation thématique

- Tout terrain
  - Domaines spécialisés
  - Sans ressource externe

- Dimensions "explictes"
  - Access aux documents pivots

- Supporte tout calcul de similarité classique

Perspectives
> Sim²> Sim. Inter-langue

- Projet TALN2013

- Calcul de similarité Anglais/Français pour des rapports d'incidents
  - Bases mélangés (Air France, DGAC)

- Documents pivots en paires
  - Bouts de rapports d'accidents canadiens (bilingues)

- Identification de la langue source des documents à indexer
- Similarité avec le bout du pivot correspondant à la langue du document

- Evaluation sur des rapports d'accident canadiens
  - Indexer séparément les versions EN et FR d'un même rapport
  - Voir si, pour un doc-source dans la langue A, son homologue dans la langue B est au 1e rang.

Perspectives
> Calculs transparents et paramétrables

- Mémoire de N. Ribeiro

- Voir les raisons du rapprochement
  - Soulignage des mots en commun
    - Représenter importance aussi
    - Pour la Sim^2?
    - En intégrant des termes complexes dans le calcul?

- Influencer le calcul
  - Poursuivre avec l'effacement de dimensions
  - Proposer d'exclure des mots/termes jugés non-pertinents

Perspectives
> Evaluation ?

- Comment évaluer l'apport de la similarité au processus de traitement de REX???
  - Pas de gold
  - Evaluations par la tâche? Quelle tâche???
  - Applications industrialisés, utilisées
    - Analyse des logs?
    - Reconstruction des sessions?
    - Feedback direct (Bon/Pas bon, Champs de commentaire)
  - Accès aux experts
    - Observations de l'utilisation des outils?
    - Entretiens?
    - Magicien d'oz?
    - Comparaisons jugement humain/auto (limites…)?

Merci de votre attention!

Références

Claveau, V., (2012). Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF, in: *Actes De La Conférence Conjointe JEP-TALN-RECITAL 2012*, Volume 2: TALN. ATALA/AFCP, Grenoble, France, pp. 85–98.

Johnson, C.W., (2003). Failure in Safety-Critical Systems: A Handbook of Accident and Incident Reporting. *University of Glasgow Press*, Glasgow.

Macrae, C., (2007). Analyzing Near-Miss Events: Risk Management in Incident Reporting and Investigation Systems. *Centre for Analysis of Risk and Regulation*.

Moreau, F., Sébillot, P., (2005). Contributions des techniques du traitement automatique des langues à la recherche d'information (No. 1690). IRSIA.

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

Tirilly, P., (2010). Traitement automatique des langues pour l'indexation d'images. Thèse de doctorat,
Université de Rennes 1 / IRISA, Rennes, France

Tulechki, N., Tanguy, L., (2012). Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques, in: *Actes De La Conférence Conjointe JEP-TALN-RECITAL 2012, Volume 2*: TALN. ATALA/AFCP, Grenoble, France, pp. 439–446.