Institut pour la Maîtrise des Risques
Sûreté de Fonctionnement - Management - Cindyniques

FICHE PROJET 12
INTÉRÊT GÉNÉRAL

1. TITRE DU PROJET : Méthodes statistiques de traitement et d'interprétation d'un retour d'expérience en langage naturel

2. OBJET ET CONTEXTE

Plusieurs projets en relation avec le retour d'expérience (REX) et le traitement automatique du langage (TAL) ont été récemment réalisés dans le cadre IMdR. Ces projets ont montré que :

- Que ce soit pour les REX orientés vers la technique et la sûreté de fonctionnement et pour ceux orientés vers les facteurs organisationnels et humains (FOH), il est nécessaire de travailler sur des données brutes du REX, de type texte, librement saisies par les opérateurs ou par les enregistrements automatiques de capteurs électroniques (ex. Health and Usage Monitoring System (HUMS)),
- Les données sont alors nombreuses, voire « massives » ; le REX sera à mener de plus en plus dans le contexte « big data », notamment à cause des systèmes HUMS; nous entendons par big data l'ensemble des méthodes de gestion, de traitement et d'interprétation de données « massives » ou fortement hétérogènes ou issues de capteurs de systèmes HUMS ou les trois à la fois,
- Ce trop-plein de données ne permet plus de traiter « manuellement » le retour d'expérience et nécessite l'utilisation de méthodes statistiques,
- Des études récentes (notamment (Tissot, 2017)) ont montré que la mise en œuvre de ces méthodes nécessite de bonnes connaissances informatiques/méthodologiques et du temps,
- La mise à disposition de tels outils accentue le risque d'effet boîte noire s'il n'est pas accompagné par un guidage à l'interprétation et une montée en compétences.

Les principales préoccupations des retours d'expérience technique, humain et organisationnel concernent principalement:

- Le traitement de données hétérogènes de retour d'expérience,
- La classification,
- La détection des signaux faibles,
- La détection de tendance et d'inflexions,
- La similarité,
- Les corrélations,
- La comparaison à des référentiels normatifs,
- Les langages contrôlés (domaine air-espace).

L'objectif premier de cette proposition est de mieux connaître les méthodes statistiques de traitement et d'interprétation d'un retour d'expérience en langage naturel et leur potentiel.

3. RESULTATS ATTENDUS

- Exemples de cas réels traités et interprétés concernant les retours d'expérience technique, humain et organisationnel.
- Guide de recommandations concernant l'utilisation des méthodes statistiques de traitement du TAL.
- Rapport de synthèse, résumé de synthèse, jeux de diapositives.

4. PROGRAMME DES TRAVAUX

Ce programme est donné à titre indicatif. Il sera explicité dans un cahier des charges qui intégrera les différents besoins des souscripteurs ainsi que les domaines d'applications souhaités.

Tâche 1 – Rappel des besoins des souscripteurs – Objectifs de l'expérimentation et choix de jeux de données tests.

Les jeux de données candidats peuvent être des données de retour d'expérience des souscripteurs ou des données publiquement accessibles (par exemple la base EPiCEA des accidents du travail de l'INRS et la base ARIA des accidents industriels et technologiques du BARPI).

L'IMdR se chargera des éventuels contacts préliminaires avec l'INRS ou le BARPI.

Tâche 2 – Recueil d'informations sur les méthodes de traitement candidates, état de l'art bibliographique.

Ces méthodes candidates peuvent être (Tanguy, 2017 ; Tissot, 2017) :

- Les méthodes de sémantique latente (LSA : Latent Semantic Analysis ; LSI : Latent Semantic Indexation),
- Le topic modeling,
- L'analyse distributionnelle (word embeddings, plongements lexicaux),
- D'éventuelles autres méthodes, comme par exemple le gradient boosting tree, peut-être adaptée à la détection de signaux faibles, ou d'autres méthodes plus « anciennes » comme celle de l'analyse de données.

Cet état bibliographique devra préciser : les concepts de la méthode, les hypothèses, les caractéristiques des données traitées, les outils logiciels disponibles, les applications existantes, les avantages et inconvénients relatés dans la documentation technique…

Tâche 3 – recueil d'informations sur les prétraitements de données

Faut-il un prétraitement ? Faut-il des ressources (lexique, terminologie, liste d'acronymes, …) ? Que peuvent apporter des prétraitements sémantiques ? Le prétraitement enrichit –il ou au contraire perturbe-t-il l'interprétation ?...

Tâche 4 – Traitement des jeux de données choisis par les souscripteurs.

Dans cette tâche on effectuera un prétraitement du (des) jeu(x) de données choisi(s), le traitement proprement dit. On spécifiera les difficultés rencontrées.

Tâche 5 – Interprétation des résultats obtenus – Conclusions quant à l'apport des outils TAL.

Cette interprétation sera effectuée par le prestataire avec l'appui d'un expert du retour d'expérience. Les représentations en sortie des outils TAL présentent-elles des difficultés d'interprétation pour les experts du REX ? Faut-il retourner aux données brutes initiales ? Peut-on intégrer la connaissance de l'expert ?

Tâche 6 – Accessibilité des méthodes aux utilisateurs.

Les utilisateurs visés sont : l'expert du retour d'expérience et l'opérationnel. On examinera en particulier : l'utilisation des outils logiciels, l'interprétation…

Tâche 7 – Synthèse.

5. RÉFÉRENCES

- Journée IMdR (16 mai 2017), Des méthodes aux applications du TAL dans le REX.
- Tanguy Ludovic (2017), Quel TAL pour le retour d'expérience? Réflexions sur les besoins et les solutions actuelles.
- Tissot Claire (16 mai 2017), Text mining sur des données d'accidentologie.
- Mason L., Baxter J., Bartlett P.L., Frean M. (1999), Boosting Algorithms as Gradient Descent, Advances in Neural Information Processing Systems, MIT Press, pp512-518.
- Projet IMdR (2013) – Détection de signaux faibles
- Projet IMdR (2013) – Retour d'expérience et TAL (Traitement Automatique des Langues)
- Projet IMdR (2017) – HUMS (Health and Usage Monitoring Systems)

6. DURÉE

9 à 12 mois

7. MONTANT DE LA SOUSCRIPTION

9000 € HT