TALN 2010, Montréal, 19-23 juillet 2010

L'apport des concepts métiers pour la classification des questions ouvertes d'enquête.

Ludivine Kuznik[1 3] Anne-Laure Guénet[1] Anne Peradotto[2] Chloé Clavel[2]

(1) Electricité de France, Direction Commerce, 92050 Paris La Défense, France
(2) Electricité de France R&D, 92141 Clamart, France
(3) Société Lincoln , 92570 Boulogne-Billancourt, France
ludivine-externe.kuznik@edf.fr, anne-laure.guenet@edf.fr, anne.peradotto@edf.fr, chloe.clavel@edf.fr

Résumé EDF utilise les techniques de *Text Mining* pour optimiser sa relation client, en analysant des réponses aux questions ouvertes d'enquête de satisfaction, et des retranscriptions de conversations issues des centres d'appels. Dans cet article, nous présentons les différentes contraintes applicatives liées à l'utilisation d'outils de text mining pour l'analyse de données clients. Après une analyse des différents outils présents sur le marché, nous avons identifié la technologie Skill Cartridge™ fournie par la société TEMIS comme la plus adaptée à nos besoins. Cette technologie nous permet une modélisation sémantique de concepts liés au motif d'insatisfaction. L'apport de cette modélisation est illustrée pour une tâche de classification de réponses d'enquêtes de satisfaction chargée d'évaluer la fidélité des clients EDF. La modélisation sémantique a permis une nette amélioration des scores de classification (F-mesure = 75,5%) notamment pour les catégories correspondant à la satisfaction et au mécontentement.

Abstract The French power supply company EDF uses text mining tools to improve customer insight by analysing satisfaction inquiries or transcriptions of call-centre conversations. In this paper, we present the various application needs for text mining tools. After an analysis of the various existing industrial tools, we identify the Skill Cartridge tool provided by TEMIS company as the more relevant to our needs. This tool offers the capability to model expressions linked to reason for satisfaction/dissatisfaction. The contribution of this modelling is illustrated here for the classification of satisfaction inquiries dedicated to the evaluation of customer loyalty. The semantic models provide a marked improvement of classification scores (F-mesure = 75.5%) for the satisfaction/dissatisfaction categories in particular.

Mots-clés : outils de *text mining*, modélisation de concepts métier, classification supervisée
Keywords: text mining tools, business concept modelling, supervised classification

LUDIVINE KUZNIK ANNE-LAURE GUENET ANNE PERADOTTO CHLOE CLAVEL

1 Introduction

Une bonne connaissance de ses clients et de leurs besoins est essentielle pour EDF, notamment afin d'adapter ses produits et services. Depuis 2003, EDF analyse de manière automatique les différentes sources de données textuelles dont elle dispose, à l'aide de techniques de fouille de contenu et de modélisation. Ces techniques ont pour vocation de faciliter le traitement et l'analyse de ces données par les *marketeurs* du Groupe. Par exemple, une description des thématiques abordées dans les centres d'appel a été obtenue par l'analyse des champs libres remplis par le conseiller en ligne, consignés dans nos Systèmes d'Information. Dans la même optique, EDF analyse des retranscriptions automatiques de conversations issues de ses centres d'appels. Un autre moyen de connaître les besoins de nos Clients est l'analyse d'enquêtes de satisfaction. Les réponses aux questions ouvertes offrent un point de vue complémentaire à celles des questions fermées, et sont susceptibles de refléter davantage l'état d'esprit du client et ses attentes car elles proviennent de son expression libre.

L'outil utilisé à EDF pour le traitement de ces différentes données textuelles permet d'effectuer de la classification, supervisée et non supervisée, de documents. Les méthodes tout d'abord employées ne prenaient pas en compte la sémantique véhiculée par les mots : après une lemmatisation puis un filtrage des mots basé sur la catégorie grammaticale et la fréquence (TF-IDF), des rapprochements entre documents étaient fait à l'aide de méthodes statistiques classiques, chaque document étant considéré comme un 'sac de mots' (perte de l'ordre d'apparition des mots). L'information fournie par l'analyse thématique de ces documents a été jugée pertinente par les *marketeurs* d'EDF. Cependant, certains aspects se sont avérés trop détaillés par rapport à leurs besoins, tandis que d'autres ne l'étaient pas suffisamment.

Deux problèmes applicatifs ont notamment été soulevés. Le premier concerne la **prise en compte de la négation et plus généralement de la sémantique**. Pour la négation par exemple, dans les deux phrases suivantes « J'ai un problème » / « Je n'ai jamais eu de problème », notre analyse statistique repose uniquement sur le nom « problème » et l'antinomie des deux phrases est perdue, ce qui n'est pas adapté pour l'analyse de la satisfaction et du mécontentement. Plus généralement pour la **sémantique,** les concepts métier sont parfois mal discriminés. Par exemple, pour la thématique métier FACTURATION, les réponses identifiées seront celles contenant les mots « facture », « mensualisation », « prélèvement », « estimation », « duplicata », « relance » ou « consommation estimée »…. Par contre, les documents contenant des phrases du type « la facture est chère » devront être intégrés à la thématique métier PRIX. Un autre exemple : les deux phrases suivantes « oui, tout va bien » et « non, cela ne va pas du tout » deviennent, après lemmatisation et filtrage grammatical, « aller ». Le sens et la polarité sont perdus. Le second problème applicatif concerne le **ciblage de l'étude par rapport aux besoins opérationnels**. Suivant les cas, il conviendra de détailler certains aspects spécifiques, pour aider à la création d'un nouveau service par exemple, alors que dans d'autres études, c'est une analyse fine de la satisfaction et du mécontentement qui devra primer.

Afin d'améliorer nos résultats, et afin que ces derniers soient plus pertinents du point de vue du besoin opérationnel, une part de sémantique a été introduite dans nos analyses. Ainsi, des mots et des expressions ont été détectés comme étant des concepts métier, ou regroupés au sein de concepts liés à la satisfaction ou au mécontentement de nos clients, de manière plus précise.

Depuis quelques années, de nombreux travaux se focalisent sur la détection des sentiments et opinions. De même, des éditeurs commencent à proposer une détection automatique de ces expressions. Nous suivons

ces travaux de près, notamment en participant au projet collaboratif DoXa[1]. Cependant, la spécificité de nos corpus nous impose de pouvoir ajuster ces modélisations génériques.

L'objet de cet article est dans un premier temps de décrire le protocole suivi pour choisir un outil de text mining permettant l'ajout de sémantique métier et répondant à nos exigences de production. Cet outil, LUXID®de l'éditeur TEMIS®, a été acquis par EDF fin 2009. La seconde partie de l'article illustrera sur un cas d'usage et des données client, l'apport de l'enrichissement du texte par des concepts sur les tâches prédictives (classement des verbatims dans des catégories prédéfinies).

2 Recherche d'un outil de text mining adapté à nos besoins : le choix de Luxid® (TEMIS)

Après avoir recensé les outils de Text Mining présents sur le marché, à partir des études comparatives mises à jour de (Chaumier, Dejean, 2003), (Deveaux et al., 2005) et (Filippone, 2006), nous leur avons appliqué les critères définis dans le protocole d'évaluation des outils de text mining (Quatrain et al. , 2004). Outre les fonctionnalités présentes, nous avons également pris en compte des critères liés à la finalité industrielle de son utilisation à EDF, soit un outil commercialisé par une société pérenne, proposant un support de qualité et une offre de prestation. De fait, ont été exclus de l'étude les outils « open-source ». L'outil devra également permettre de traiter une volumétrie importante, disposer de ressources en français, et enfin offrir des traitements linguistiques et statistiques de qualité. Les critères considérés pour les traitements linguistiques sont ceux cités dans l'introduction : ils doivent permettre le ciblage de thématiques métier et une **modélisation sémantique fine** (création de concepts) pour analyser des opinions, par exemple.

Trois outils ont ainsi été sélectionnés et évalués en fonction de ces critères : PASW Modeler Text Analytics de SPSS ; le couplage de l'outil linguistique XEROX/XIP[2] et d'un outil statistique ; la suite logicielle LUXID de l'éditeur TEMIS[3]. C'est ce dernier qui a été acquis par EDF fin 2009. Cette suite logicielle propose des fonctionnalités d'extraction d'information, d'analyse du contenu et une interface de navigation dans les résultats. En ce qui concerne la modélisation sémantique des concepts métier, la technologie Skill Cartridge™ est particulièrement puissante. Ensembles de règles et de composants linguistiques définissant l'information à extraire, les cartouches de connaissance ont la particularité d'obéir au principe de lectures successives. Après une première lecture à la charge de l'analyseur morphosyntaxique de l'outil (Xelda), la cartouche va permettre de réaliser une analyse sémantique sur plusieurs niveaux : reconnaissance des entités, du lexique, puis des relations de premier niveau, de deuxième niveau, etc. A chaque lecture, le texte étiqueté est remplacé par le concept correspondant. Lors de la lecture suivante, le serveur d'extraction ne voit plus le texte mais seulement les concepts. Une règle peut ainsi contenir des entrées lexicales, des expressions régulières, des étiquettes grammaticales mais également des concepts de niveau inférieur. Par ailleurs, TEMIS® est une société pérenne proposant un accompagnement au niveau des formations et des prestations. Même si l'interface de navigation et l'utilisation des différents modules n'est pas toujours ergonomique et intuitive, c'est cette solution qui nous est apparue comme la plus adaptée : les résultats de la classification non supervisée des corpus de test ont été jugés comme étant les plus pertinents par trois experts métier.

[1] Traitement automatiques des opinions et sentiments (2009-2011), https://www.projet-doxa.fr/-Presentation-du-projet-.html

[2] XIP est commercialisé par Celi France

[3] http://www.temis.com/index.php?id=201&selt=1

3 Apport de la sémantique pour catégoriser un document

L'objectif de cette partie est d'illustrer l'apport de la sémantique pour une tâche de catégorisation. Le corpus est constitué de 3.653 réponses à une question ouverte posée aux clients sur leur intention de rester fidèle à EDF. Il est constitué de documents très courts, de une à quelques phrases en moyenne. Les données ont été annotées manuellement par un expert en huit thématiques présentées dans le tableau 2.

Tableau 1- Catégories utilisées pour la classification et exemple d'expressions caractéristiques de la catégorie.

	Description	**Exemples**
Satisfaction globale	Clients satisfaits de manière générale : ils n'estiment pas avoir de raison valable ou suffisante pour changer de fournisseur.	"Parce que ca fonctionne. Pourquoi changer ?" , "J'ai toutes les raisons de rester fidèle."
Satisfaction tarifs	Ces clients mettent en avant les tarifs préférentiels dont ils bénéficient, ainsi que la régulation des tarifs à laquelle est soumise EDF.	"EDF nous propose des tarifs jaunes et pas les autres fournisseurs."
Confiance EDF	Ces clients évoquent leur confiance en EDF.	"On connaît EDF alors que les autres on ne les connaît pas."
Attachement service public	Ces clients sont favorables à la défense du service public, des monopoles d'Etat et/ou à la sauvegarde des entreprises françaises .	"Il faut un seul et unique fournisseur " "Fidèle aux services publics de EDF "
Fidélité passive	Ces clients n'ont pas envie de chercher un nouveau fournisseur d'électricité.	"Je n'ai pas envie de me casser les pieds. "
Indécision	Ces clients sont indécis, ils n'ont pas pris le temps d'étudier d'autres offres et n'ont pas été démarchés par les concurrents.	"Je ne sais pas ce que font les autres" , "Je n'en sais rien"
En attente propositions.	Clients ouverts à l'éventualité de changer de fournisseur si la concurrence propose mieux, notamment en termes de prix.	"Si je trouve moins cher ailleurs je change "
Insatisfaction	Clients insatisfaits, qu'il s'agisse des tarifs, du manque de suivi ou de proximité, de la qualité de l'accueil, etc.	"Je ne suis pas satisfait des coûts"

3.1 Méthode

Le corpus est analysé via une *cartouche de connaissance* appelée « Fidelité », construite par nos soins afin de récupérer les concepts métier caractéristiques des différentes catégories (première colonne du tableau 2). Pour cela, l'analyse des résultats d'une classification non supervisée a permis de détecter les thèmes mal repérées, comme par exemple séparer les marques de satisfaction et d'insatisfaction, ou encore modéliser la notion de confiance.

Cette cartouche a été construite en trois étapes : les concepts outils, les concepts lexicaux et les règles. Les concepts outils permettent de faciliter le travail du développeur en regroupant des éléments syntaxiques et sémantiques que l'on utilisera ensuite dans d'autres règles. On retrouvera ici des regroupements sur les notions de fréquence (« maintes reprises, déjà 3 fois, continuellement, encore , toujours, tout le temps, …) , sur les adverbes intensifieurs (« absolument, beaucoup, de plus en plus, trop, …) ou encore la modélisation de la négation. Les fichiers lexicaux regroupent également du vocabulaire comme des noms qualifiant la qualité d'un service (« amabilité, disponibilité, courtoisie, sérieux, … »), ou encore des adjectifs (« conciliant, professionnel, … »), mais aussi du vocabulaire propre à EDF regroupant par exemple ses concurrents, ses services, ses offres, … Enfin, des règles ont été créées pour améliorer l'affectation des verbatims aux catégories définies par les experts. Par exemple, pour illustrer la catégorie insatisfaction, on retrouve des règles comme :

(1) *~negation-patt / ~satisfait-lex / (~negation-patt~contexte)?* , basée sur l'enchaînement des trois concepts de niveau inférieur que sont la modélisation de la négation, une liste d'adjectifs concernant la satisfaction des personnes, et éventuellement l'expression « du tout ». Cette règle permettra d'extraire par exemple « pas satisfait du tout», « vraiment pas acceptable », …

(2) *(manque|~negation-patt|(il/#NEG/y/avoir/~negation-patt))/(#PREP_DE)?/ (conseil|contact|~services-lex)*, basée sur un des éléments entre parenthèses, à savoir le lemme « manque », un des termes du concept gérant la négation (« pas vraiment, pas du tout, n'a pas », par exemple) ou encore une expression de type « il n'y a pas » suivi éventuellement de la préposition de (étiquettes grammaticales issues de l'analyseur morphosyntaxique de TEMIS, Xelda), puis d'un des termes entre parenthèses (les lemmes conseil ou contact ou les termes regroupés précédemment dans le concept services-lex). On y retrouvera des verbatims comme « manque de qualité de service », « il n'y a vraiment pas eu de contact », …

Cette cartouche de détection de concepts a été couplée à une analyse morphosyntaxique (cartouche Analytics™ de TEMIS®) pour récupérer également les termes ne faisant pas partie des expressions modélisées. C'est le lemme qui sera conservé, en filtrant sur la catégorie grammaticale pour garder uniquement les noms, groupes nominaux, adjectifs et verbes.

Une fois le traitement linguistique effectué, une catégorisation est réalisée par une méthode basée sur les réseaux bayésiens[4] fournie dans LUXID. 80% du corpus est utilisé pour l'apprentissage et 20% pour le test. L'évaluation de l'apport de la sémantique sur le résultat de la catégorisation est obtenu en comparant ces résultats à celui d'une analyse morphosyntaxique seule (cartouche Analytics™).

3.2 Résultats

La qualité des résultats est évaluée selon trois mesures : le rappel, la précision et la qualité (moyenne lissée des deux autres). Afin d'éviter les biais engendrés par le tirage aléatoire, le protocole a été itéré 20 fois. Le tableau ci-dessous fournit une moyenne des vingt mesures obtenues.

Tableau 1- résultats de la catégorisation avec ou sans l'utilisation de la cartouche "Fidélité".

	Précision	**Rappel**	**F-mesure**
Sans la cartouche « Fidélité »	76,3%	69,6%	72,8%
Avec la cartouche « Fidélité »	79,0%	72,3%	75,5%

Les performances sont légèrement meilleures avec la cartouche « Fidélité » avec un **gain de 2,7 points pour la F-mesure** qui intervient de manière quasi-égale sur la précision et le rappel. Il est intéressant de noter que l'amélioration des performances varie en fonction de la classe considérée. La F-mesure par catégorie oscille entre **44,1% et 86,1%** lorsque la catégorisation repose sur une analyse morphosyntaxique seule et entre **60,4% et 85,4%** lorsque l'analyse morphosyntaxique est utilisée conjointement à la cartouche « Fidélité »**.** Les résultats avec cette cartouche sont particulièrement meilleurs pour les catégories *F8-Insatisfaction, F2-Satisfaction tarifs* et *F1-Satisfaction globale*, avec une F-mesure qui passe respectivement de 44,1% à 61,6% , de 52,8% à 63,3% , et de 76,1% à 79,8% . Ces trois catégories sont les plus sujettes au problème de la **négation** présenté en introduction car elles permettent de discriminer le mécontentement de la satisfaction. La modélisation par une cartouche de connaissance est donc

[4] Méthode d'apprentissage vectoriel assignant aux documents un score d'appartenance à une catégorie. Elle permet également de n'affecter aucune catégorie à un document si les scores sont trop faibles.

particulièrement pertinente pour ces catégories et permet une nette amélioration des résultats. Cela permet notamment de réaffecter correctement à la catégorie Insatisfaction des verbatims comme « Pas contente », « je n'ai pas eu de suivi sur mes demandes», « manque de communication », ou à la catégorie Satisfaction « On n'est pas insatisfait d'EDF », « je n'ai rien à leur reprocher ».

Tableau 2-Résultats de la catégorisation avec ou sans l'utilisation de la cartouche "Fidélité".

Catégories *(1ère colonne : avec Analytics,* ***2ème colonne : avec notre cartouche Fidélité)***	Précision		Rappel		F-mesure	
1.Satisfaction globale	80,6	**85,5**	72,1	**74,8**	76.1	**79,8**
2.Satisfaction tarifs	86,5	**67,2**	38,0	**59,8**	52.8	**63.3**
3. Confiance EDF	82,3	**85,6**	53,4	**53,8**	64.8	**66.2**
4. Attachement service public	90,6	**92,9**	82,0	**79,1**	86.1	**85.5**
5. Fidelité passive	79,2	**88.0**	74,6	**70,4**	76.9	**78.2**
6. Indécision	55,4	**64,2**	58,6	**57,1**	57.0	**60.4**
7 En attente propositions	73,2	**72,4**	88,7	**90,5**	80.3	**80.5**
8. Insatisfaction	66,7	**78,7**	32,9	**50,6**	44.1	**61.6**

4 Conclusion

Les besoins opérationnels concernant l'analyse de données textuelles pour une meilleure connaissance de la clientèle EDF sont spécifiques et passent par la **modélisation sémantique de concepts métiers** concernant les motifs de satisfaction ou de mécontentement des clients. Nous avons illustré l'apport de cette modélisation - réalisée grâce à la technologie Skill Cartridge™ de l'éditeur TEMIS - pour la classification des réponses des clients à une enquête chargée d'analyser la fidélité des clients. Cette modélisation a permis d'obtenir une F-mesure supérieure à 75% et de faire passer les performances les plus basses (obtenues pour la catégorie liée à l'insatisfaction) de 44.1% à 60,4%.

La modélisation sémantique des concepts est actuellement réalisée à EDF par des experts linguistes. La suite envisagée pour cette étude concerne la recherche de méthodes statistiques capables d'aider le linguiste à la modélisation de ces concepts.

Références

CHAUMIER J., DEJEAN M. (2003). Recherche et analyse d'information textuelle. Tendances des outils linguistiques. *Documentaliste-Sciences de l'information* vol. 40, 14-24.

DEVAUX V., VIDAL S., FABRY C. (2005). Benchmarking outils de veille. Projet mené au sein du service veille de l'INstitut de l'Information Scientifique et Technique, http://outils.veille.inist.fr/index.html

FILIPPONE D. (2006). Les outils de Text Mining. Article publié sur *Le Journal du Net*, http://www.journaldunet.com/solutions/0606/060630-panorama-text-mining/1.shtml

QUATRAIN Y., PERADOTTO A., NUGIER S. (2004). Evaluation d'outils de Text Mining dans un contexte industriel. Actes de *CIFT 2004*, 103-115.