Évaluation d'outils de Text Mining : démarche et résultats

Yasmina Quatrain[1], Sylvaine Nugier[1], Anne Peradotto[1], Damien Garrouste[2]

[1]EDF R&D – 92141 Clamart Cedex – France
[2]Lincoln – 92774 Boulogne-Billancourt Cedex – France)
yasmina.quatrain@edf.fr, anne.peradotto@edf.fr

Abstract

Électricité de France, the French electricity provider is moving into new markets, and needs to know very well its clients and tracks their satisfaction. Thus, different kinds of text mining tools were tested in order to analyze a huge quantity of heterogeneous textual documents, including mails, open-ended customer survey questions, discussion forums, and comments contained in large databases concerning customer contacts. In this context, it seemed unavoidable to build a test grid, in order to facilitate the comparison between different tools. This grid was created according to a test on data mining software, achieved by a consulting firm. However, we tried to give equal importance to the statistics and the linguistics aspects. It is divided in ten points, ranging from the company which developed the tested software to the methods implemented and data extraction. This paper describes the test grid carrying out, the software selection, the way the evaluation of four tools (Alceste, SAS Text Miner, TEMIS Insight Discoverer and SPAD/CRM) was achieved and the results. In conclusion, this work underlines that text mining tools are of two types: statistics software including linguistics modules and linguistics software using statistical methods to analyze text. Furthermore, the evaluation stressed on the obligation of using detailed test grids and creating evaluation protocols, in order to define the functionalities which are essential to be tested in a software, according to the user's aims and profile and the kind of corpora he or she would like to analyze.

Résumé

Dans le contexte de l'ouverture du marché de l'électricité, EDF désire analyser les gros volumes de données textuelles qui lui permettront de mieux connaître ses clients. Dans cette optique, plusieurs outils de text mining destinés à l'analyse de cette information hétérogène de taille importante ont fait l'objet d'une évaluation à l'aide de trois corpus de nature différente. La constitution d'une grille de test facilitant la comparaison des logiciels est apparue indispensable. Inspirée d'une expertise sur les outils de data mining, elle a été réalisée en évitant de privilégier la communauté statistique au détriment de la linguistique. Cette grille se compose de dix thèmes variant de la société éditrice aux champs d'application en passant par l'accès aux données et l'analyse du tableau lexical. Outre le déroulement de l'évaluation et ses résultats sur quatre outils du marché (Alceste, SAS Text Miner, TEMIS Insight Discoverer et SPAD/CRM), cet article retrace la démarche de constitution de la grille de test, le choix des outils évalués et les critères retenus. Il conclut sur l'existence de deux types d'outils en text mining : ceux statistiques enrichis d'un module de traitement du texte et ceux originellement linguistiques. De plus, cette expérience conforte l'utilisation d'un protocole détaillé permettant de déterminer et évaluer les fonctionnalités incontournables en fonction des objectifs et du profil de l'utilisateur du logiciel et de la nature du corpus à analyser.

Mots-clés : évaluation, protocole, text mining, logiciels.

1. Introduction

Le volume croissant de données textuelles provenant de l'Internet et des contacts clients (par mails, retranscription de message téléphonique, lettres de réclamation, enquête…) apportent une quantité d'informations non exploitable manuellement et actuellement peu exploitée à

Électricité De France (EDF). Pourtant ces données sont indispensables à une bonne connaissance des clients et à l'amélioration de la gestion de la relation avec celui-ci, notamment dans le contexte actuel d'ouverture à la concurrence.

Le processus d'extraction de cette information à partir d'un large volume de données textuelles non structurées est appelé text mining. Il débute par une étape primordiale de préparation des données allant de la structuration à l'enrichissement en passant par le filtrage. Les méthodes employées ensuite pour extraire l'information pertinente sont principalement des tâches de classification et de classement.

C'est dans ce contexte que nous avons été confrontés au choix d'un logiciel traitant de données textuelles (text mining), pour différents types d'études. Par outils de text mining, nous entendons ici des outils permettant l'analyse de données non-structurées (textuelles dans notre cas) associées à des données structurées telles que les données relatives à la consommation d'un client ou à son type de logement. Afin d'aider à l'évaluation et à la comparaison de ces outils, la constitution d'une grille de test est apparue nécessaire.

Cette grille avait ainsi deux principaux objectifs :
– Permettre l'évaluation comparée de logiciels analysant des données textuelles ;
– Être une aide à la décision pour l'achat d'un tel logiciel adapté au besoin des utilisateurs.

La démarche s'est voulue pragmatique en recherchant tout d'abord les expériences similaires dans le domaine du data mining et du traitement du langage naturel.

Une évaluation de logiciels de data mining (CXP, 2001) a permis d'avoir un bon aperçu des volets devant figurer dans notre grille. Il ne traitait pas la partie relative aux fonctionnalités de traitement du texte, inexistante dans ce domaine. Nous avons trouvé assez peu de références sur la réalisation de protocoles d'évaluation ou l'évaluation d'outils de cette nature (Brugidou *et al.*, 2000). Les articles traitant du sujet concernent des logiciels très spécifiques tels que l'évaluation pour le résumé automatique (Barthel *et al.*, 2002) ou celle d'analyseurs syntaxiques (Aït Mokhtar *et al.*, 2003). Le projet européen TECHNOLANGUE, co-financé par le Ministère français délégué à la recherche et aux nouvelles technologies, le Ministère français délégué à l'Industrie et le Ministère de la Culture et de la Communication, comporte un volet évaluation, décliné en domaines de traitement du langage naturel, dans lesquels ne figure pas le text mining (dont EVALDA-ARCADE II pour l'évaluation de l'alignement de corpus multilingues ou EVALDA-CESTA pour les systèmes de traduction automatique).

La grille devant permettre l'évaluation de nos outils a été construite à l'aide des références précédemment citées, et a évolué tout au long des tests sur les trois types de corpus retenus (questions ouvertes d'enquête, champs commentaires d'une base de données sur les contacts Clients, forums de discussion).

Nous avons finalement retenu un ensemble de critères organisés en trois axes :

– le commercial (prix, prestations, documentation…) ;
– le technique (architecture, limites volumétriques…) ;
– le fonctionnel (traitements possibles, convivialité, exploitabilité des résultats...).

Cet article contient la démarche, la grille de test appelée PITT (grille Pour l'évaluation de logiciels statistiques Incluant le Traitement du Texte) et les résultats sur les quatre outils évalués. Un objectif à plus long terme est la création d'un véritable protocole de test d'outils de text mining.

2. Élaboration de la grille de test

2.1. *La démarche*

Au cours de la construction de la grille de test, nous avons cherché à éviter de privilégier la communauté statistique au détriment de celle du traitement automatique des langues, l'une plus sensible aux moyens mis à disposition pour l'analyse du texte, l'autre souhaitant trouver des méthodes élaborées d'analyse de données ou de modélisation. De plus, nous ne souhaitions pas lister plus ou moins exhaustivement les fonctionnalités disponibles dans un ensemble d'outils sans se pencher plus particulièrement sur les méthodes applicables au texte.

Pour le premier point il convient d'utiliser un langage compréhensible par les deux communautés. Pour le second, et à l'inverse des outils de data mining qui offrent des gammes de plus en plus larges de méthodes, il semble important de recentrer les études à partir des propriétés essentielles (capacité à classifier, capacité à classer) que l'on attend d'un outil de text mining et dont les résultats sont interprétables et pertinents.

2.2. *Le choix des logiciels évalués*

Les logiciels testés ne forment pas une liste exhaustive de l'offre du marché, cependant le choix s'est porté volontairement sur des logiciels différenciés offrant ainsi une large couverture technique (dans les méthodologies proposées) et une large couverture opérationnelle (dans les applications possibles).

Le choix des logiciels s'est fait en deux temps : SAS étant l'outil statistique de référence à EDF, le test du nouveau module SAS/Text Miner est apparu évident. De même pour Image/Alceste, outil utilisé depuis plusieurs années à la R&D d'EDF pour l'analyse de textes. Les deux autres outils ont alors été choisis pour compléter la gamme avec comme objectif de balayer au maximum l'ensemble des fonctionnalités offertes par ces outils dits de text mining. Nous avons ainsi sélectionné la suite TEMIS Insight Discoverer présentée comme une véritable solution de text mining, et un logiciel de statistique permettant l'analyse du texte, SPAD/CRM de la société DECISIA.

Ces quatre logiciels donnent donc un très bon aperçu des outils existants dans le paysage du text mining pour les raisons suivantes :

– Les orientations « commerciales » sont différentes.

Alceste est un logiciel dédié à l'analyse de données textuelles utilisé pour traitement du discours sans fonctionnalités inférentielles. SAS Text Miner est une solution de text mining intégrée dans une suite logicielle de data mining. SPAD/CRM se positionne également sur le marché du data mining. TEMIS Insight Discoverer est un outil exclusivement text mining.

– Les méthodologies et fonctionnalités adoptées sont larges.

On observe des différences très importantes dans la partie du traitement purement textuel (présence ou non d'un outil linguistique plus ou moins performant), dans la construction et la réduction des tableaux lexicaux (analyses factorielles ou autres) et enfin dans les méthodes d'analyse proposées.

– Les degrés de maturité sont différents.

Alceste et SPAD/CRM (anciennement appelé SPAD-T) sont des logiciels éprouvés dans l'analyse de données textuelles, la solution SAS Text Miner est par contre très récente (la version que nous testons de SAS est la première version commercialisée), de même que la solution TEMIS Insight Discoverer.

2.3. *Le déroulement du test*

Il est important que les outils de text mining soient tous testés selon le même principe afin de garantir la comparabilité des résultats et sa relative pérennité dans le temps. Les tests ont été effectués à partir d'une machine unique et par la même personne, dans le cadre d'un stage de licence professionnelle en data mining.

Pour chacun des corpus, les fonctionnalités de chaque logiciel ont été éprouvées en renseignant au fur et à mesure les différents points de la grille de test.

Les corpus choisis pour les tests représentent un échantillon non représentatif de l'ensemble des données textuelles pouvant être analysées par les méthodes de text mining. Cependant ils ont été sélectionnés pour leurs caractères différenciés et parce qu'ils correspondent aux différents types de corpus que nous avons été amenés à analyser. Ces derniers sont les suivants :
– Corpus « QO » : réponses à une question ouverte dans une enquête de satisfaction EDF ;
– Corpus « commentaires » : champs commentaire extrait d'une base de données EDF ; ce champ est renseigné, en cas de nécessité, par l'agent à la suite de contacts téléphoniques avec les clients ;
– Corpus « forums » : forums de discussions sur internet.

La nature des données (langage naturel ou retranscription) est différente suivant nos corpus : Les corpus « QO » et « forums » sont du langage naturel contrairement à celui des « commentaires » où les motifs d'appels des clients ont été retranscrits en abréviations par un opérateur.

La volumétrie est également très variable d'un corpus à l'autre : Le corpus « commentaires » est constitué de 100 000 motifs d'appel, « forums » contient 2 000 interventions et « QO » 400 réponses à une question ouverte sur la satisfaction.

2.4. *Les critères retenus*

Les critères retenus se déclinent en 10 thèmes, chaque thème contenant des sous-thèmes non détaillés dans ce chapitre.

1. La Société

Cette rubrique a pour but d'indiquer l'origine de l'outil. Cela permet de déterminer si ce dernier suit une école en particulier (statistique à la française ou autre). Par le biais du pays d'origine, cette rubrique renseigne aussi sur la langue pour laquelle le produit a été conçu. Les autres entrées permettent de déterminer la santé financière de la société à l'origine du logiciel, la pérennité de ce dernier et son potentiel d'évolution au sens financier.

Items évalués :
Origine – Nom de la Société – Effectif France/Monde – CA France/Monde - Gamme de produits – Existence Club utilisateurs – Site Web.

2. Le Produit – Aspects Financiers

Cette partie concerne le coût de la licence et permet ainsi de définir si un déploiement de large envergure nécessite des moyens financiers importants. Les aspects formation et conseil (prestation) permettent d'évaluer la durée de mise en place d'une première étude.

Items évalués :
Achat/location – Coût Licence – Coût d'assistance / maintenance évolutive – Formation – Existence de partenaires en conseil et développement sur le produit.

3. Le Produit – Aspects Techniques

Trois sous-parties composent ce thème : l'architecture, la prise en main et les généralités.

Les items relatifs à l'architecture doivent permettre de répondre à des questions du type : Quelle architecture faut-il mettre en place ? Le produit est-il adapté à l'environnement de travail de l'utilisateur ? S'insert-t-il dans l'environnement technique en vigueur dans l'entreprise ? L'achat du logiciel nécessite-t-il l'achat de machines supplémentaires plus récentes ? …

Les items relatifs à la prise en main nous renseignent sur la facilité avec laquelle l'utilisateur s'approprie l'outil. Elles font également l'inventaire des supports à sa disposition.

Des remarques générales sur le logiciel, apportant des détails sur l'appréhension par l'utilisateur de sa philosophie, sa fiabilité et la manière dont les erreurs sont gérées sont réunies sous la dernière rubrique.

Items évalués :
Mode Client/Serveur – Systèmes d'exploitation supportés – Ressources mémoire nécessaires – Espace disque nécessaire – Logiciels nécessaires – Historique des versions
Niveau de l'utilisateur – Existence d'aide en ligne, manuel d'utilisation – Documentation Méthodologique
Niveaux de fiabilité (fréquence des bugs) – Présence d'un système de gestion des erreurs (traces) – Présence d'une organisation en mode projet (organisation des travaux) – Degré d'ouverture et de personnalisation de l'outil – Existence de sauvegarde des paramétrages et résultats – Appréciation générale sur le temps de calcul – Appréciation sur la transparence des méthodes statistiques utilisées – Appréciation générale sur l'ergonomie, la facilité d'apprentissage, la convivialité.

4. L'accès aux données

Cette partie permet de déterminer d'une part la facilité d'intégration des données textuelles et extra-textuelles au logiciel (accès, formatage, langage de programmation spécifique, etc.) et d'autre part ses limites en terme de volumétrie des données pouvant être analysées.

Items évalués :
Possibilité d'accéder à distance aux documents – Duplication obligatoire des documents – Passage par outil d'extraction des données
Formats de documents supportés en entrée (Liste) – Accès direct aux SGBD – Appréciation sur le degré de pré-formatage utilisateur
Nombre d'observations maximums autorisées – Nombre de colonnes maximums autorisées – Nombre de caractères maximums des champs textes – Taille maximum des noms de variables.

5. Les pré-traitements et l'identification des unités textuelles

Les fonctionnalités propres au pré-traitement du texte avant le lancement des analyses sont réunies dans cette section. Cette dernière indique la marge de manœuvre et les facilités offertes à l'utilisateur par le biais de l'interface mise à sa disposition, afin de réaliser les tâches suivantes : nettoyer son texte et le mettre aux normes du produit, l'étiqueter, l'enrichir et enfin visualiser, éditer, importer ou exporter le corpus traité ou le tableau lexical (voir définition au critère 6) construit.

Items évalués :
Pré normalisation des documents (orthographe, harmonisation, majuscules...) – Langues sup-

portées (liste) – Reconnaissance automatique – Mélange de langues – Reconnaissance du rôle grammatical dans la phrase – Reconnaissance des groupes nominaux – Lemmatisation, stemmatisation, autres – Reconnaissances d'entités (noms propres, nombres, adresses,...)
Reconnaissances co-occurrences – Regroupement de termes possible (utilisateur) – Création liste de synonymes – Existence liste de synonymes – Mise à jour liste de synonymes – Création liste de mots vides – Existence liste de mots vides – Utilisation possible d'une liste de départ – Outil de visualisation du corpus sélectionné – Outil d'édition du corpus sélectionné – Statistiques lexicométriques sur le corpus (nombre de termes, d'hapax) – Exportation du tableau lexical entier pour analyses utilisateurs – Importation possible d'un tableau lexical entier (interne ou externe).

6. La transformation et la réduction des tableaux lexicaux

On appelle « tableau lexical » le croisement des unités lexicales (n-grammes, mots lemmatisés, …) et des documents.

Cette rubrique concerne les transformations et réductions possibles du tableau lexical dans l'outil. La transformation de ce tableau aborde notamment les différents types de pondérations des fréquences sur les unités lexicales. La réduction peut être réalisée par filtrages sur ces dernières (sur les fréquences, pondérées ou non, sur un nombre d'unités…), ou par des méthodes plus complexes de type analyses factorielles.

La possibilité de paramétrage du logiciel au niveau de ces transformations est également évoquée, ainsi que la visualisation et la modification manuelle du tableau réduit avant de nouvelles analyses.

Items évalués :
Méthode(s) de transformation des fréquences (richesse, pertinence) – Méthode(s) de réduction des tableaux lexicaux (richesse, pertinence) – Paramétrage possible – Visualisation et d'édition du tableau lexical réduit.

7. L'analyse du tableau lexical

Cette rubrique répertorie les méthodes d'analyse du tableau lexical réduit en les divisant en deux catégories : la classification et le classement, qui sont les deux grands objectifs du text mining. La caractérisation des classes à l'aide de variables non textuelles est également abordée sous cette rubrique.

Items évalués :
Croisement avec des variables non textuelles
Classification : Méthode(s) de classification implémentée (Liste) – Possibilité d'inclure des variables externes actives – Détection automatique du nombre de classes
Autres méthodes
Classement : Liste des méthodes – Méthodes de validation – Scoring.

8. La gestion et la présentation des résultats

Cette section permet d'indiquer la nature de la restitution des résultats par l'outil. Une génération automatique de rapports, l'intégration aisée de graphiques dans des documents préexistants, ou encore l'aide à l'interprétation sont autant d'éléments jouant en la faveur de l'outil et permettant à l'utilisateur de réduire ses délais de traitement.

Items évalués :
Retour au texte initial – Représentations graphiques des résultats statistiques

Éditions d'aides à l'interprétation – Éditions de rapports – Appréciation générale sur la gestion et la présentation.

9. Les champs d'applications

Les champs d'application ont pour objectif général d'aider un futur utilisateur à choisir le logiciel le plus adapté à ses besoins et à ses connaissances. Ils représentent une synthèse des rubriques précédentes, à laquelle s'ajoute l'expérience acquise lors du test des logiciels sur les différents corpus.

Item évalué :
Type d'utilisation conseillée.

10. Les perspectives

Les perspectives servent à estimer la viabilité du logiciel et sa marge d'évolution.

Des outils peuvent être issus de travaux de recherche, du secteur industriel, être plus récents que d'autres, tous ces éléments doivent être renseignés afin de permettre une meilleure appréhension de l'avenir du logiciel.

Items évalués :
Pérennité de l'éditeur – Maturité de l'outil (marge de progression) – Défauts majeurs – Mises à jours prévues.

3. Les résultats et l'interprétation

3.1. *Le test des quatre outils*

Le tableau suivant synthétise les résultats du test des outils Alceste, SPAD/CRM, Text Miner et TEMIS Insight Discoverer[1] :

	ALCESTE	SPAD/CRM	SAS/Text Miner	TEMIS/Insight Discoverer
La Société				
Origine	France	France	USA	France
Pérénité	*	**	***	**
Le produit				
Coût du produit	Très accessible	Moyen	très cher	très cher
Architecture	PC	PC	CL/Serveur	CL/Serveur
Prise en main	Très rapide	Très rapide	Difficile (SEM)	Moyenne
Accès aux données				
formats	*	* *	* * *	***
limites (volumétrie)	*	* *	* * *	***
Pré-traitement				
outils linguistiques	Très complet	Interface pour effectuer manuellement des regroupements et filtrages	Présence de nombreuses fonctionnalités, mais résultats erronés	Très complet
Intervention manuelle	Possible mais inutile	Obligatoire, mais assez conviviale	Obligatoire en l'état actuel du produit mais pas conviviale	Possible mais inutile
Classification				
Diversité des méthodes de réduction du tableau lexical	*	* *	* * *	*

[1] Le détail de l'évaluation est disponible sur demande.

	ALCESTE	SPAD/CRM	SAS/Text Miner	TEMIS/Insight Discoverer
Diversité des méthodes d'analyse des données	* *	* * *	*	*
Classement				
Méthodes de modélisation	0	* *	* * *	*
Résultats				
Lecture des résultats	* * *	* *	*	***
Rapport	* * *	* *	*	***
Champs d'application préconisés	Outil linguistique efficace, Détection rapide des thèmes d'un corpus et bonne caractérisation des classes thématiques	Outil linguistique inexistant, Outil très efficace pour l'analyse des données	Outil linguistique très défaillant, Outil efficace pour effectuer du Data Mining	Outil linguistique et statistique efficace mais absence de caractérisation des classes thématiques

La notation de certaines rubriques varie de médiocre (*) à très satisfaisant (***).

Les évaluations effectuées dégagent un certain nombre d'oppositions et de rapprochements entre les différents outils tant au niveau de l'utilisation, de la méthodologie, des fonctionnalités linguistiques et des fonctionnalités statistiques. Une synthèse a été effectuée afin de pouvoir déterminer les positionnements respectifs de chaque outil. Ainsi, une note a été attribuée à chacun d'entre eux sur 10 macro-critères résumant ceux abordés en détail dans notre grille de test :

– La société (pérennité de l'éditeur, le pays d'origine…)
– le produit (architecture, coût du produit, prise en main)
– l'accès aux données (volumétrie, pré-formatage)
– les outils linguistiques (présence et la qualité de l'outil linguistique)
– l'automatisation (nécessité d'une intervention manuelle et qualité de l'outil fourni)
– la réduction des dimensions (qualité et diversité des méthodes pour transformer le tableau lexical)
– les méthodes de classification (diversité des méthodes pour classer les documents ou dégager les thèmes abordés)
– les méthodes de classement (diversité des méthodes pour réaliser des modèles de classement automatique)
– la lecture des résultats (présentation et lisibilité des résultats ou aides à l'interprétation)
– et enfin le rapport (présence et qualité du rapport d'étude produit automatiquement par le logiciel)

3.2. *L'interprétation*

Les quatre produits se positionnent très différemment.

Pour Alceste, la partie pré-traitement des données est automatique et efficace (enrichissement important des données grâce aux outils linguistiques). Les classes thématiques obtenues sont homogènes et leur caractérisation avec des variables exogènes, si elles existent, est performante et utile. Le rapport d'analyse, généré automatiquement, est également un atout. Les deux principaux problèmes résident dans la volumétrie limitée et dans l'absence de méthodes de modélisation (aucune méthode de classement par exemple). Alceste est ainsi un outil très efficace pour détecter rapidement les thèmes d'un corpus, mais ne peut se positionner comme un outil de text mining.

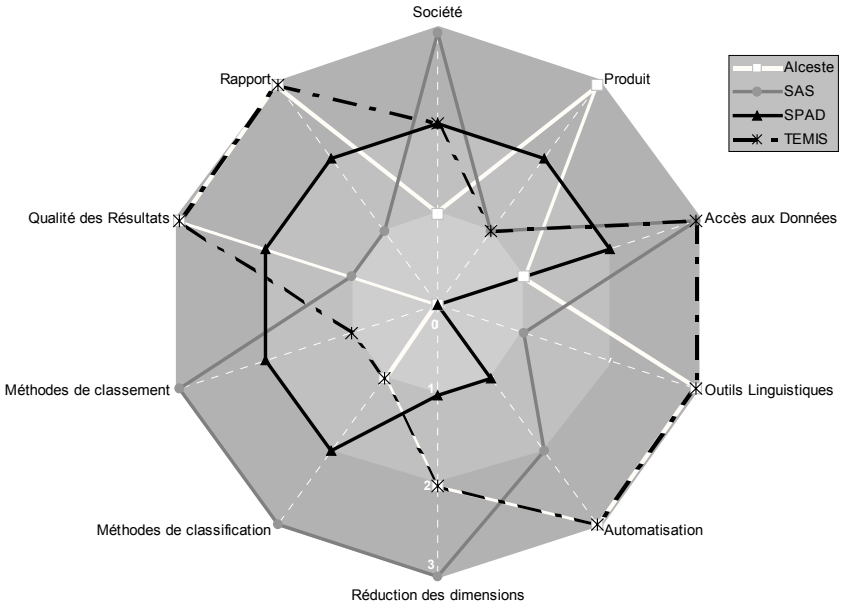

Figure 1. Positionnement des quatre outils de Text Mining

À l'opposé, SAS Text Miner possède un grand nombre de méthodes de modélisation et la volumétrie n'est limitée que par les caractéristiques de la machine. Par contre, pour la partie linguistique, il faut attendre une prochaine version pour véritablement tester ses possibilités (résultats erronés, apparemment non testé pour le français). L'aide à l'interprétation des résultats est également très décevante (pas de caractérisation des classes obtenues, pas de classement par pertinence des documents d'une classe, interface de visualisation peu ergonomique…). Cet outil de text mining s'adresse clairement à des « data miners » confirmés, connaissant déjà le produit de data mining SAS Enterprise Miner dans lequel est inclus le module de text mining.

La suite logicielle TEMIS Insight Discoverer se positionne comme un outil de text mining, permettant une analyse efficace du texte et supportant une grosse volumétrie. Il ne possède cependant pas de fonctionnalités statistiques permettant par exemple de caractériser les classes obtenues par la classification du texte avec des variables illustratives non textuelles. Afin d'utiliser toutes les potentialités du produit, il est pour le moment nécessaire de connaître un langage de programmation permettant de manipuler les données.

Le positionnement de SPAD/CRM se situe entre Alceste et Text Miner. Ce produit dispose d'outils d'analyse de données permettant une exploration fine du corpus et d'outils de modélisation permettant de créer des modèles de classement, sur de grands volumes de données. Par contre, il ne possède aucun outil linguistique, et même si l'interface de filtrage des « mots » est assez conviviale, la préparation des données avant analyse est une étape longue et fastidieuse pour les gros corpus.

4. Conclusion

L'objectif premier de cette étude, à savoir la préconisation d'un logiciel pour l'entreprise, a évolué vers une évaluation comparative des logiciels, dans le sens où la diversité des corpus à traiter ainsi que les objectifs sont tels que les outils sont davantage complémentaires que concurrents.

En ce qui concerne les besoins spécifiques et les corpus d'EDF, Alceste demeure l'outil privilégié de dépouillement et d'exploration des réponses aux questions ouvertes d'enquêtes de satisfaction, d'une volumétrie réduite et présentant des variables explicatives illustratives. La suite logicielle TEMIS Insight Discoverer est quant à elle bien adaptée au traitement des champs commentaires de notre base de données des contacts clients dont la volumétrie dépasse les 100 000 documents et dans lesquels on trouve un langage très spécifique (vocabulaire technique, utilisation d'abréviations...).

Cette évaluation conforte l'idée d'utiliser un protocole de test détaillé, afin d'éprouver un ensemble de fonctionnalités incontournables à la fois en fonction des objectifs visés par l'utilisation d'un outil de text mining, des corpus à analyser mais également du profil de l'utilisateur (langage naturel, statisticien, data miner). De plus, à l'issue de cette dernière, il semble que les outils dont la vocation initiale était l'analyse du texte restent les plus performants.

La grille de test présentée a pour vocation d'évoluer vers un véritable protocole de test d'outils de text mining, d'une part en testant d'autres types d'outils (outils dédiés à la veille par exemple) ou d'autres types de corpus (corpus multilingues) et d'autre part en choisissant un panel d'utilisateurs de niveaux de connaissances variables.

Références

Aït Mokhtar S., Hagège C. et Sàndor A. (2003). Problèmes d'intersubjectivité dans l'évaluation des analyseurs syntaxiques. In *Actes de TALN 2003*. www.sciences.univ-nantes.fr/irin/taln2003/articles/eval1.pdf/

Barthel M.P., Khouas L., Sanford E. et Couillault A. (2002). Évaluation automatique pour résumé automatique. In *Journées d'étude de l'ATALA sur résumés de texte automatiques : solutions et perspectives.* http://www.atala.org/je/021214/Barthel.pdf

Brugidou M., Escoffier C., Folch H., Lahlou S., Le Roux D., Morin-Andréani P. et Piat G. (2000). Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles. In *Actes des JADT 2000.*

CXP. PackExperts 2001 « Business Intelligence : outils de Data Mining », société CXP International. 19-21 rue du rocher, 75008 PARIS.

Nugier S. Garrouste D., Peradotto A. et Quatrain Y. (2003). Grille de test de logiciels de text mining. Note Interne à EDF. Disponible à la demande.