Evaluation d'outils de Text Mining dans un contexte industriel

Yasmina Quatrain, Anne Peradotto, Sylvaine Nugier

Evaluation d'outils de Text Mining dans un contexte industriel

Yasmina Quatrain, Anne Peradotto et Sylvaine Nugier

EDF Recherche et Développement, 1 avenue du Général de Gaulle 92140 Clamart

{sylvaine.nugier, anne.peradotto, yasmina.quatrain}@edf.fr

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. 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.

ABSTRACT. Electricité 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 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.

MOTS-CLÉS : Text Mining, évaluation, outils, corpus, grille de test.

KEYWORDS : Text Mining, evaluation, tools, corpora, test grid.

Actes de CIFT 2004, pages 103 à 115

Introduction

Le volume croissant de données textuelles provenant de l'Internet et des contacts clients apportent une quantité d'informations non exploitable manuellement et actuellement peu utilisée à Electricité 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 ceux-ci, notamment dans le contexte actuel d'ouverture du marché de l'électricité à la concurrence.

Le groupe de recherche Statistique et Outils d'Aide à la Décision (SOAD) du département Innovation Commerciale et Analyse des Marchés et de leur Environnement (ICAME) a développé un pôle de compétence dans le domaine du Text Mining au sein d'EDF Recherche et Développement. Les efforts de l'équipe se portent sur :

- le problème **du formatage, de la modélisation de données commerciales** en entrée des logiciels de Text Mining et de l'enrichissement de celles-ci à l'aide d'outils de traitement du langage naturel. La réflexion porte en particulier sur la mise au point d'architectures et de formats pivots facilitant le pré-traitement des données.
- **les outils de Text Mining** afin de connaître les logiciels actuellement sur le marché et ceux dont les résultats sont prometteurs en matière de traitement de données clients.
- **les méthodes d'analyse** de Text Mining sous-jacentes les plus efficaces pour le traitement des données commerciales spécifiques d'EDF.

Nous abordons tout d'abord le Text Mining en tant que discipline répondant aux besoins opérationnels de notre entreprise. Nous présentons ensuite notre expérience autour de la création d'une grille de test qui nous semble incontournable avant une étape d'évaluation de logiciels. De plus, nous décrivons notre procédure d'évaluation ainsi que le choix des logiciels et des corpus. Enfin, nous présentons nos résultats avant de conclure et faire part des perspectives possibles. Nous insistons sur le fait que cette démarche repose sur nos besoins spécifiques et dans un contexte précis.

1 Le Text Mining en tant que discipline opérationnelle en entreprise

Le Text Mining, ou Text Data Mining, ou encore Knowledge Discovery in Texts (KDT) peut être défini comme du Data Mining sur données textuelles. C'est cette définition qui est utilisée à EDF. Le Text Mining comprend l'ensemble des techniques issues du traitement automatique du langage naturel, qui permet de transformer les données textuelles en données « codées », et de la fouille de données

permettant de trouver des informations cachées dans de larges bases de données textuelles.

Le volume croissant de données textuelles provenant de l'Internet, des Intranets, des Forums de discussion, des contacts client (par mails, retranscription de messages téléphoniques, lettres de réclamation, enquêtes…) propose une quantité d'informations potentiellement pertinentes pour toute entreprise mais qui ne sont pas exploitables manuellement. Les enjeux économiques attachés à l'utilisation de ces informations sont très importants et l'utilisation du Text Mining est devenue incontournable à EDF.

Le traitement de gros volumes de données textuelles s'est fait ressentir suite à plusieurs demandes internes à l'entreprise concernant des données commerciales (réponses à des questions ouvertes provenant d'enquêtes de satisfaction et commentaires saisis lors de contacts avec la clientèle). L'objectif était tout d'abord d'exploiter des informations capitalisées dans les bases de données, en fouiller le contenu. Le second consistait à en extraire des éléments permettant une meilleure connaissance de nos clients.

L'équipe a, en particulier, été confrontée au choix d'un logiciel traitant de données textuelles 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.

2 Elaboration d'une grille de test

3.1 Démarche

La mise au point de cette grille de test a ainsi deux principaux objectifs :

- Permettre l'évaluation comparée de logiciels analysant des données textuelles ;

- Etre 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 comporte un volet évaluation décliné en domaines de traitement du langage naturel (par exemple EVALDA-ARCADE. II pour l'évaluation de l'alignement de corpus multilingues ou EVALDA-CESTA pour les systèmes de traduction automatique) dans lequel ne figure pas le Text Mining.

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...).

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 souhaitant trouver des méthodes élaborées d'analyse de données ou de modélisation, l'autre plus sensible aux moyens mis à disposition pour l'analyse du texte. De plus, nous ne souhaitions pas lister exhaustivement les fonctionnalités disponibles dans un ensemble d'outils sans nous pencher plus particulièrement sur les méthodes applicables au texte. Il convient d'utiliser un langage compréhensible par les deux communautés et 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.

3.4 Contenu

Les critères retenus se déclinent en 10 thèmes.

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). 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.

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.

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'insère-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 sur la transparence des méthodes statistiques utilisées - Appréciation générale sur l'ergonomie, la facilité d'apprentissage, la convivialité.

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 :

Passage par outil externe d'extraction des données

Formats de documents supportés en entrée (Liste) - Accès direct aux SGBD - Appréciation sur le degré de pré-formatage utilisateur.

Nombre de documents maximums autorisés - Nombre d'unités linguistiques maximums analysées – Nombre de variables illustratives maximums autorisées - Nombre de caractères maximums des champs textes - Taille maximum des noms de variables illustratives.

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 construit.

Items évalués :

Pré-normalisation des documents (orthographe, harmonisation, majuscules...) - Langues supportées (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 – Possibilité de regrouper manuellement des termes - 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) - Réutilisation du tableau lexical entier pour analyses utilisateurs - Importation possible d'un tableau lexical entier (interne ou externe).

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.

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 grandes méthodes 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

Modélisation et classement : Liste des méthodes - Méthodes de validation – Scoring.

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 pré-existants, 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

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

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.

Les perspectives

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

3 Procédure d'évaluation des outils de Text Mining

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[1] 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.

4.1 Sélection des corpus

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 Lincoln sur internet.

La nature des données (langage naturel ou retranscription) est différente suivant nos corpus : les corpus « QO » et « forums » sont l'expression directe des clients contrairement à celui des « commentaires » où les motifs d'appels des clients ont été retranscrits de façon plus ou moins abrégée 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 400 interventions et « QO » environ 2 000 réponses à une question ouverte sur la satisfaction.

---

[1] Stage de fin d'étude en collaboration avec la société Lincoln, 92774 Boulogne-Billancourt Cedex – France.

4.2 Choix des outils

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 en particulier par les sociolinguistes depuis plusieurs années à la R&D d'EDF pour l'analyse de textes (enquêtes). 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 origines ou philosophies des produits se distinguent en : outils linguistiques intégrant des traitements statistiques ou outils statistiques s'enrichissant de modules linguistiques.

- 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.

Nous présentons, dans le chapitre suivant, le résultat détaillé de notre évaluation. La figure 1 synthétise ces résultats.

4 Résultats de l'évaluation

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 de 0 à 3 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).

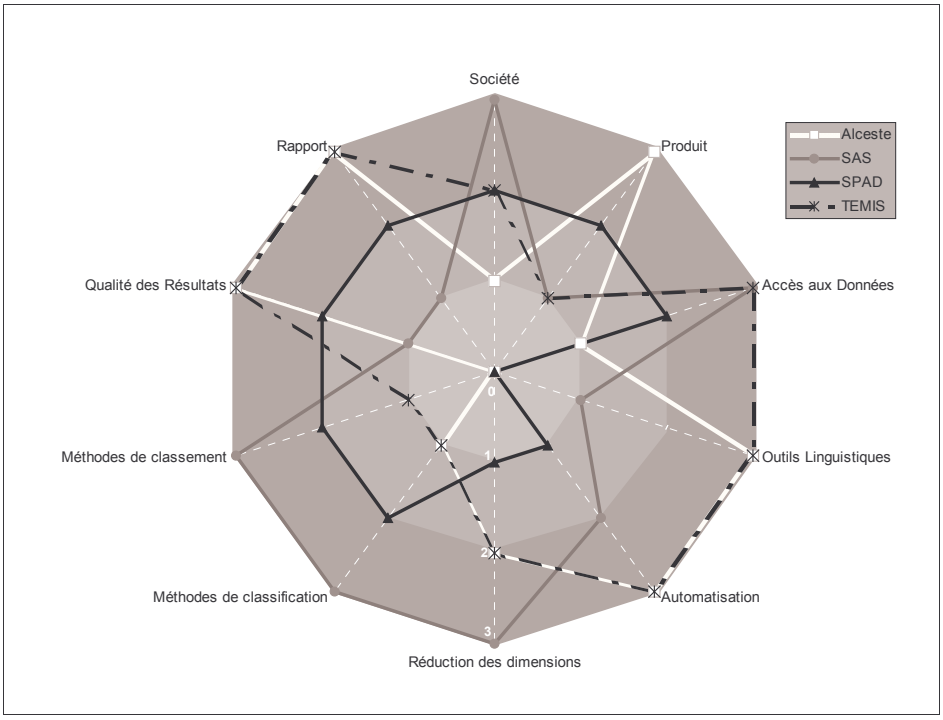

Figure 1. positionnement des quatre outils de Text Mining

5.2 Synthèse des résultats

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).

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 ne présente pas de caractérisation des classes obtenues. ni de classement par pertinence des documents d'une classe. L'interface de visualisation est peu ergonomique. Cet outil de Text Mining s'adresse plutôt 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.

5 Conclusion et perspectives

Notre objectif était de proposer des solutions de traitement de gros volumes de données textuelles réparties dans les bases de données clientèle d'EDF, les enquêtes de satisfaction, les échanges par mails ou d'autres canaux. A l'issue de notre évaluation des quatre outils de Text Mining, nous avons pu juger **du rôle important**

d'un protocole d'évaluation. Cette expérience conforte l'idée d'utiliser une grille de test détaillée, 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 (linguiste, statisticien, data miner, …).

La démarche d'évaluation, qui avait comme premier objectif de préconiser un logiciel pour l'entreprise, a évolué vers une évaluation comparative des logiciels. En effet, il est apparu que, la diversité des corpus à traiter ainsi que les objectifs visés par ce traitement sont tels que les **outils testés se révèlent davantage complémentaires que concurrents**.

Par rapport aux besoins spécifiques et aux corpus d'EDF, nos choix se tournés vers Alceste comme 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, quant à elle, a été jugée 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...). Les messages provenant du forum de discussion Lincoln étant trop spécifiques par rapport à nos besoins, ils ne sont donc pas directement intervenus dans nos préconisations.

L'évaluation réalisée reste incomplète car elle n'étudie pas la pertinence et l'efficacité de certaines méthodes relativement nouvelles de Data Mining telles que les Support Vector Machines ou l'algorithme EM, par exemple, appliquées au texte. On trouve sur ce sujet de nombreuses références anglo-saxonnes, mais très peu sur le traitement des textes français. Il serait intéressant dans un avenir proche de se pencher sur la pertinence de l'utilisation de ces « nouvelles méthodes » en fonction des types de données que nous manipulons et des objectifs que nous visons. De plus, 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) et d'autres types de corpus (corpus multilingues) et d'autre part en choisissant un panel d'utilisateurs de niveaux de connaissances variables.

6 Bibliographie

Aït Mokhtar S., Hagège C. et Sàndor A.(2003). *Problèmes d'intersubjectivité dans l'évaluation des analyseurs syntaxiques*. 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). *Evaluation automatique pour résumé automatique. 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 : *JADT 2000 (5èmes Journées Internationales d'Analyse Statistique des Données Textuelles).*

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

JF Marcotorchino, '*Les technologies avancées de l'analyse de l'information* : *Text Mining, Data mining et fusion Data Mining – Text Mining'*, REE N°7/8, juillet-septembre 2001.

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.

R.Feldman, I. Dagan , *'KDT - Knowledge Discovery in texts'*, Proceeding of the Conf. On Knowledge Discovery (KDD) 1995.

U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy '*Advances in Knowledge Discovery and Data Mining*', AAAI Press, MIT Press, 1996.