E-marketing et textmining - Une application à l'analyse des opinions de consommateurs sur Internet

Claire Gauzente[1]

[1]LARGO – UFR Droit Économie Gestion – 49 000 Angers - France

Abstract

The paper aims at delineating the potential advantage of textmining for marketing applications. Even if qualitative inquiry has also been a significant avenue in marketing studies, the information flow engendered by the penetration of Internet reveals the need of quantitative approaches to qualitative material. A case study is proposed focused on the analysis of consumer opinions about a antivirus product.

Résumé

Le papier a pour objectif de délimiter les avantages du textmining pour les applications marketing. Si les investigations qualitatives constituent une voie de recherche réelle dans les études marketing, le flux informationnel engendré par la pénétration d'Internet met en exergue le besoin d'approches quantitatives pour des données qualitatives. Une étude de cas est proposée, centrée sur l'analyse des avis de consommateurs d'antivirus.

1. Introduction

Le développement des nouvelles technologies et, en particulier, la pénétration croissante d'Internet dans les foyers offrent aux marketers des opportunités quasiment sans limite d'exploitation de données clients. Le marketing des bases de données, aujourd'hui repris dans les démarches de CRM – customer relationship management ou GRC – gestion de la relation client, s'appuie sur des données quantitatives pléthoriques. Parallèlement à la multiplication des données structurées, les données non structurées connaissent également une croissance rapide (forums d'utilisateurs et de consommateurs, faq, listes de diffusion et de discussion, blogs).

De nouveaux besoins d'investigation d'une information qualitative quantitativement développée émergent. Le textmining propose une solution à ces besoins (Hearst, 1999) bien que sa définition suscite toujours des controverses (Kroeze, Matthee et Bothma, 2003). Toutefois, parmi les fonctions les plus couramment citées, on recense les suivantes (Crié, 2001 ; Nahm, online source, Miller, 2005 ; Fan et alii, 2005) : requêtes, résumé, visualisation & topographie, navigation, catégorisation, voire création de nouvelles connaissances.

Sur le plan des logiciels de textmining proposés, tout comme en analyse statistique textuelle, tous ne répondent pas aux mêmes objectifs (Jenny, 1997). L'utilité managériale des fonctionnalités possibles est variable. Il nous semble dès lors intéressant de mettre en perspective les apports du textmining à la prise de décision marketing en nous appuyant sur une illustration par les avis de consommateurs sur Internet et en exploitant les résultats fournis par un logiciel de textmining, Wordmapper.

2. L'apport du textmining au marketing

Deux principaux axes d'étude marketing peuvent bénéficier des techniques de textmining (cf. tableau 1).

Le premier concerne la compréhension du comportement du consommateur : ses critères de choix, ses processus de recherche d'information, ses modes d'évaluation des produits et services, l'usage qu'il fait des produits / services, les critères de satisfaction et d'insatisfaction, les recommandations (bouche à oreille positif ou négatif sur la toile), agrément suscité par les marques la communication des marques et enseignes.

Le second axe relève plus d'une visée stratégique d'anticipation et de détection d'évolutions en germe. Cette dimension à visée de veille est peut être plus délicate à mettre en œuvre.

Orientation	Problématiques marketing	Sources d'information textuelles
Comportement du consommateur	Conception produit	Forums d'opinion, espaces communautaires, emails ou commentaires sur sites commerciaux
	Communication positionnement	idem
	Suivi des usages, retours produits	idem
	Prix	Idem
	Place : différence des évaluations selon les communautés, étude des contagions	Idem + blogs
	Vente et Service après-vente	Emails [+ informations internes (BDD clients)]
Stratégie	Stratégie : • benchmarketing, • anticipation des tendances • évolutions des goûts et attentes des consommateurs	Sites de la concurrence + Blogs + forums et espaces ouverts.

Tableau 1. Problématiques marketing et sources d'information électroniques

Au total, ainsi que le souligne Crié (op. cit.) par rapport à l'évolution du marketing vers une prise en compte plus fine et personnalisée du client, le textmining est susceptible de permettre l'élaboration de réponses plus pertinentes et subtiles.

3. Une application au cas des évaluations consommateurs

L'application au cas des opinions de consommateurs revêt un intérêt pour le consommateur comme pour le responsable marketing car ces derniers permettent :

- aux consommateurs : de se renseigner (quels sont les produits disponibles) ; de jauger les produits (comparaisons entre produits) ; de juger les usages des produits ; d'établir leurs propres références (construction du prix de référence, des performances à attendre, des usages principaux et complémentaires du produit) ; de procéder à l'élaboration de leurs convictions propres en soupesant les opinions des autres ;
- aux entreprises fabricants / distributrices : de prendre conscience des critères sur lesquels leurs produits sont jugés ; de connaître la performance de leurs produits sur ces critères ; d'identifier les concurrents avec lesquels les consommateurs les mettent en comparaison ; de connaître les failles des produits ;

Globalement, une meilleure connaissance des opinions consommateurs permet de mieux cerner le processus de perception du risque par le consommateur au moment de son achat (Volle, 1995). Ainsi que le souligne Volle, le risque perçu constitue à la fois un critère de segmentation et un critère de ciblage des stratégies de réduction de risque pour certains segments. Les renseignements pris par la visite des magasins (réels ou en ligne), les descriptions produits, le bouche à oreille (qu'il soit ou non électronique) ont tous pour vertu de réduire les incertitudes liées à l'achat.

Les antivirus sont des produits dont la nécessité est volontiers reconnue par les détenteurs de PC. La presse informatique, et parfois les médias télévisuels et radiophoniques nationaux, relayent les nouvelles attaques virales et alimentent les craintes des utilisateurs. Dans de telles périodes, le besoin de protection est vivement ressenti. Parallèlement, le choix du consommateur 'lambda' n'est pas simple : offre variée, coexistence de produits commerciaux et de produits libres ; évaluation ex ante difficile. Le recours à une troisième partie est important. L'avis d'un tiers qui jouera le rôle plus ou moins prononcé de prescripteur devient alors essentiel (Stenger, 2005).

Hatchuel (1995) distingue trois formes de prescription : la prescription de fait, technique et de jugement. Seules la prescription technique et la prescription de jugement influencent de manière significative le processus d'achat. La prescription technique apporte à l'acheteur des notions qu'il ignore et qui lui permettront de mieux évaluer les solutions (produits) possibles tout en étendant le champ des possibles (autres usages, possibilités techniques alternatives, critères à prendre en compte). La prescription va au-delà car elle ne se contente pas de fournir des critères d'appréciation mais des appréciations elles-mêmes (ex : jugements sur un vin). Les consommateurs qui s'en remettent à de telles informations délèguent en grande partie leur décision d'achat au tiers prescripteur.

Le degré auquel les consommateurs se reposent sur la/les prescriptions peut varier en fonction de nombreux éléments mais un de ceux qui nous paraît important dans le domaine des opinions de consommateurs sur Internet nous paraît être la crédibilité de la source. En effet, les recherches sur la crédibilité des sources dans le domaine des médias (Newell et Goldsmith, 2001) soulignent l'importance de la confiance suscitée par la source autant que de son expertise perçue.

4. Méthodologie et mise en œuvre

4.1. Constitution du corpus

La collecte des données a pris en considération plusieurs critères que sont :

- l'identification d'un produit potentiellement impliquant et suffisamment complexe à évaluer pour nécessiter le recours à des sources d'informations complémentaires autres que le descriptif produit ;
- l'identification de sources différentes : commerciales et non commerciales ;
- l'identification d'un produit faisant l'objet de suffisamment de commentaires (inutile d'avoir 5 avis consommateurs) ;
- l'identification d'un corpus de taille modeste permettant d'évaluer rapidement l'apport du logiciel de textmining (permet-il de rendre compte du corpus ? de le comprendre différemment ?).

Ces critères pris en compte conduisent au corpus final suivant (tableau 2) :

Critères	Choix	Description corpus
Produit complexe	Kaspersky antivirus personal edition	31 avis JDN 30 avis FNAC 4391 mots (statistiques word)
Produit suscitant de nombreux commentaires		
Sources commerciales et non commerciales dont l'audience est importante	Journal du net Fnac	

Tableau 2. Constitution du corpus d'analyse

Le corpus a dû être préparé en vue de l'utilisation du logiciel de textmining Wordmapper (distribué par la société Grimmersoft). Voici les opérations effectuées :

- homogénéisation : le commentaire d'ouverture (en général un mot d'appréciation) a été intégré dans le corps du commentaire ;
- la note est rejetée en variable fermée ;
- la source est une variable fermée ;
- correction orthographique.

4.2. Analyse et arbitrages

Le logiciel réalise une première passe sur le texte importé (plusieurs formats peuvent être importés : pdf, html, text tags et méta tags, etc.) et offre une liste de mots candidats à l'analyse. Par défaut, ces mots sont sélectionnés ainsi : élimination des petits mots outils, fréquence minimum de 3, longueur minimum de 3 lettres. Ces paramétrages par défaut peuvent être modifiés et complétés. Nous les avons retenus tels quels et avons complété par une recherche des mots composés (trois mots composés ont été intégrés : anti-virus, anti hackers, mise-à-jour).

Sur cette base, Wordmapper a dressé une première liste de mots sur lesquels l'analyse sera effectuée, cette liste est appelée 'liste de mots signifiants'. Cent vingt mots étaient candidats à la liste que nous avons paramétrée manuellement (une sélection automatique est aussi possible, par tranches de 200 mots). Le choix un à un des termes signifiants permet différents arbitrages :

- lemmatisation ou non lemmatisation des termes examinés un à un ;
- regroupement par racine (exemple : détecte, détecter, détection) ;
- regroupement des synonymes (pas de cas sur ce corpus).

Une fois la liste constituée, le logiciel procède à l'élaboration du tableau lexical. La longueur des parties est ici de 10 mots (le guide de Wordmapper parle d'une fenêtre flottante, elle peut être paramétrée). La constitution des groupes de mots est réalisée par analyse ascendante hiérarchique. Le choix du nombre de groupe peut être manipulé par l'intermédiaire du nombre de mots maximal admis dans les groupes, par défaut ce nombre est établi à 1/8e du corpus, nous avons conservé ce paramétrage.

Une fois les groupes constitués, une matrice de cooccurrence des groupes est reprise pour une analyse MDS ce qui permet une représentation plus visuelle des groupes (cf. Annexe).

Les groupes (ou clusters dans la terminologie Wordmapper) peuvent être ouverts : une représentation MDS est à nouveau affichée. Plus loin, le réseau sémantique de chaque 'mot signifiant' peut être affiché.

Trois formes graphiques sont donc disponibles :

- Le graphique MDS des groupes de mots ;
- Le graphique MDS de la structure interne de chaque groupe ;
- Le réseau sémantique de chaque mot.

La navigation dans le corpus peut être réalisée par l'intermédiaire de la sélection d'un mot, l'affichage de son contexte de citation se réalise sous un navigateur web (IE est préférable). L'usage de couleurs différentes et d'intensité différente permettent de savoir à quel niveau de graphique on se situe et d'appréhender la fréquence du terme ou du groupe.

Les statistiques offertes comportent, notamment, la densité de chaque cluster qui représente le degré d'autonomie et de spécificité du cluster. Il est calculé comme suit :

$$\mathbf{d} = \frac{\Sigma \text{ cooccurrences internes au cluster}}{\Sigma \text{ cooccurrences entre les mots du cluster et tous les autres}}$$

Une ACP est également disponible, qui permet de projeter une variable fermée avec les mots du corpus.

Nous ne détaillerons pas toutes les fonctionnalités du logiciel, mais voici les analyses réalisées pour cette étude :

Analyses sous wordmapper	• graphe des clusters • graphe interne de chaque cluster • réseau sémantique du mot 'détecte' • navigation web / affichage des contextes (permet d'extraire des verbatims) • statistiques : densité des clusters ; acp (croisement mots/ variable note) ; cooccurrences entre clusters
Analyse complémentaire sous spss	Test de moyenne entre variables fermées (note d'appréciation globale / source)

Tableau 3. Analyses réalisées sur le corpus

5. Résultats obtenus et mise en perspective

Dans une perspective de textmining, notre objectif premier sera d'évaluer dans quelle mesure les sorties logicielles permettent de faire sens rapidement du corpus. Dans un second mouvement, nous chercherons à voir en quoi les résultats peuvent être interprétés par rapport à un cadre d'appréhension marketing du consommateur.

La visualisation 'à chaud' du graphe général ne permet pas véritablement de conclure. C'est par un processus itératif d'investigation et de navigation dans les données que les groupes de mots font sens progressivement. Toutefois, il convient de reconnaître que cette tâche est facilitée par le logiciel qui produit les graphes et les affichages de contexte à la demande.

Après avoir sorti le graphe général et les graphes de chaque cluster, nous avons ciblé des mots qui nous paraissaient a priori plus importants ou significatifs que d'autres (l'affichage détaillé de chaque réseau sémantique serait contre-productif).

Progressivement, la structure du corpus est apparue assez clairement en lien avec, d'une part, le cadre d'appréhension du risque perçu à l'achat et, d'autre part, le cadre d'analyse des formes de prescription tous deux cités plus haut.

Sept clusters sont identifiés, ils renvoient chacun à des aspects bien particuliers :

- le premier concerne la qualification du contexte d'utilisation du produit (l'utilise-t-on dans un cadre familial ? avec quel niveau d'expertise ? etc.) ;
- le second renvoie aux problèmes rencontrés : avant l'achat mais aussi après l'achat, le produit Kaspersky résout des problèmes mais en pose également !
- le troisième cluster concerne l'évaluation du produit : ses fonctionnalités, son efficacité, son rapport qualité-prix ;
- le quatrième concerne les dimensions achat et post achat ;
- le cinquième reflète la marque et sa réputation ;
- le sixième concerne les comparaisons faites par les utilisateurs entre Kaspersky et les produits concurrents ;
- le dernier est un cluster constitué de seulement trois mots, le réseau étoilé a donc été affiché, il concerne un aspect ciblé du produit, à savoir la détection de virus.

Les cooccurrences entre clusters ainsi que leur positionnement dans l'espace permettent de comprendre que les commentaires des utilisateurs s'organisent autour de deux pôles : un pôle d'évaluation « sur pièce » regroupe les commentaires concernent des appréciations après usage du produit ; un autre pôle réunit des jugements de valeur fondés principalement sur les croyances, la réputation du produit Kaspersky mais aussi des produits concurrents.

Un « balisage » du corpus initial (description de chaque commentaire au travers, ici, de la source et de la note attribuée) permet de constater que les commentaires 'sur pièce' ou les jugements de valeur peuvent se côtoyer au sein d'un même commentaire de consommateur.

Globalement, l'identification de ces deux pôles correspond assez bien à la distinction opérée par Hatchuel et reprise par Stenger sur les formes de prescriptions, la prescription technique renvoie au pôle 'évaluation sur pièce' alors que le pôle des jugements de valeurs est conforme à la prescription dite 'de jugement'.

L'examen de l'ACP met en évidence un point attendu à savoir que les termes mélioratifs sont associés à des notes (note sur dix) plutôt élevées alors que les termes dépréciatifs le sont avec des notes faibles. L'analyse sous SPSS montre que les commentaires positifs (notes sur dix / sources) sont plus fréquents sur la Fnac que sur le JDN. Il ne s'agit toutefois que d'une tendance, le pourcentage d'erreur s'élevant à 11%. Un tel résultat devrait conduire les consommateurs en recherche d'information à apprécier les commentaires issus de la Fnac avec plus de circonspection.

6. Conclusion

Le travail réalisé avait pour objectif d'illustrer l'utilisation d'un logiciel de textmining sur un corpus constitué dans une visée marketing. La problématique spécifique choisie était les

évaluations de consommateurs sur Internet. Par rapport au cadrage théorique rapide (risque perçu par le consommateur et formes de prescriptions), l'exploration permet à la fois de montrer que la compréhension du corpus peut être réalisée relativement rapidement (moyennant un cadre théorique préalable et une exploration progressive par les grands moyens offerts par le logiciel) et que la nature des évaluations de consommateurs diffère probablement selon les sources (commerciales ou non) dans lesquelles elles sont diffusées.

Les limites de cette étude reposent principalement sur la taille (volontairement) limitée du corpus. Afin de mettre en œuvre une véritable logique de textmining (exploration d'un grand ensemble textuel), d'une part, et, d'autre part, de confirmer les tendances pressenties ici (différence entre sources commerciales et non commerciales), il conviendrait de compléter le corpus par, par exemple, des avis de consommateurs collectés sur d'autres supports (amazon, zdnet, forum d'opinion comme toluna, ciao etc.). Le travail présenté a donc une valeur principalement exploratoire et illustrative.

Références

Beaudoin V., Fleury S.et Velkovska J. (2000). Études des échanges électronique sur internet et intranet : forums et courriers électroniques. *Journées d'analyse des données textuelles* : 17-24.

Bickart B. and Schindler R.M. (2001). Internet forums as influential sources of consumer information. *Journal of Interactive Marketing*, 15 (3) : 31-40.

Crié D. (2001). NTIC et extraction des connaissances. CLAREE Cahiers de la Recherche, Université de Lille.

Fan W., Wallace L., Rich S. and Zhang Z. (2005). Tapping into the power of textmining. *Communication of the ACM*.

Hatchuel A. (1995). Les marchés à prescripteurs. In Verin H. et Jacob A., *L'inscription sociale du marché*. L'Harmattan.

Jenny J. (1997). Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine – État des lieux et essai de classification, *Bulletin de Méthodologie Sociologique*, 54 : 64-122.

Journal of research for consumers. http://www.jrconsumers.com/.

Kroeze J.H., Matthee M.C. and Bothma T.J.D. (2003). Differentiating data- and text- mining terminology. *Proceedings of SAICSIT* (south african institute of computer scientists and information technologists) : 93-101.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Hearst M. (1999). Untangling Text Data Mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Miller T.W. (2005). *Data & text mining a business application approach*. Pearson.

Nahm U.Y. (online). A roadmap to textmining and web mining. http://www.cs.utexas.edu/users/pebronia/text-mining/overview/

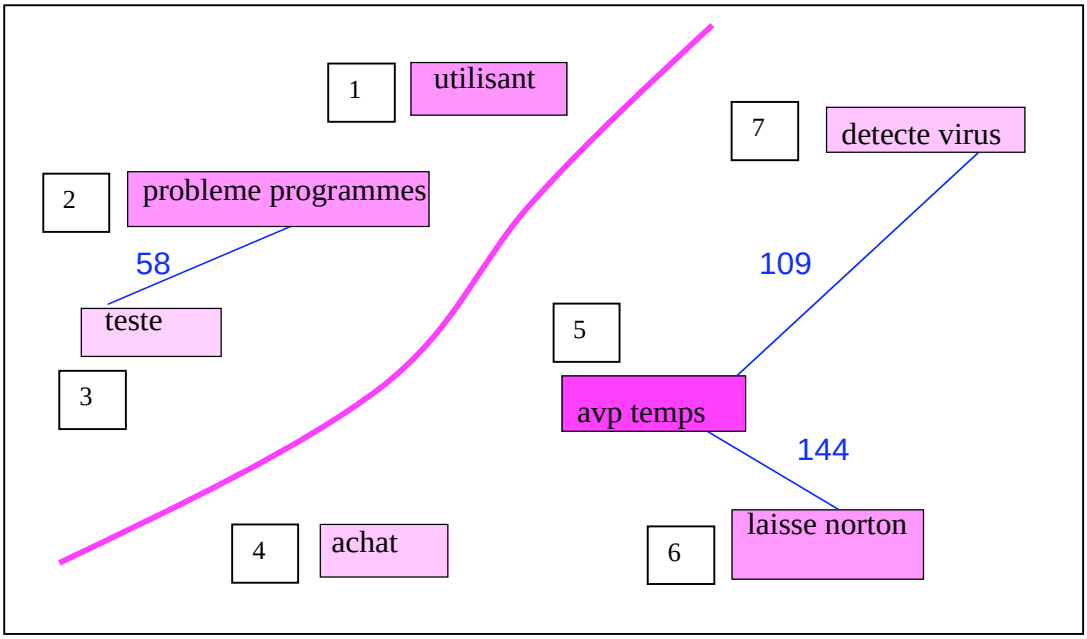
Newell S. and Goldsmith R.E. (2001). The development of a scale to measure perceived corporate credibility. *Journal of Business Research*, 52 : 235-247.

Stenger T. (2005). De la vente de vin par Internet à la modélisation des rapports de prescription dans la relation d'achat en ligne, *Actes de la Journée Nantaise E-Marketing, septembre*.

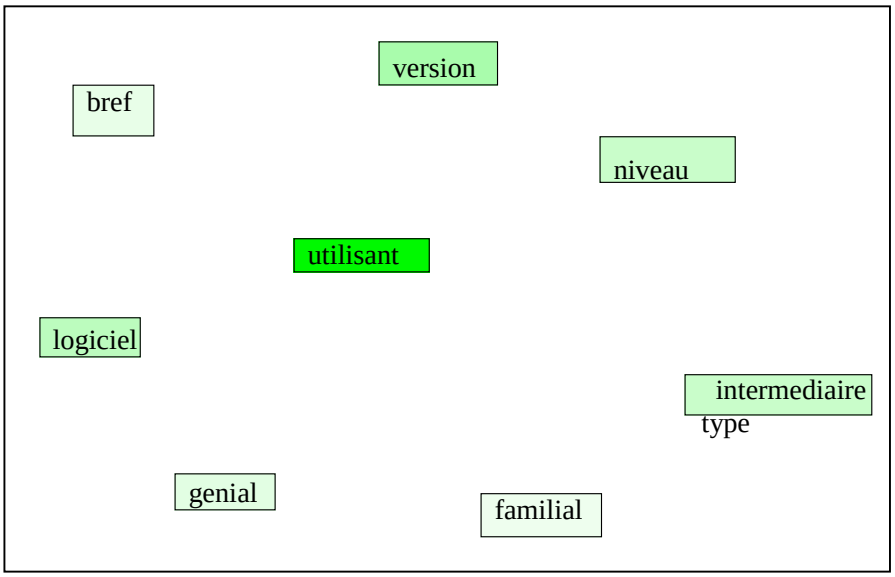
Ward J.C. and Ostrom A.L. (2003). The internet as information minefield – An analysis of the source and content of brand information yielded by net searches, *Journal of* Business *Research*, 56 907-914.

Wordmapper, *Guide d'utilisation*.

ANNEXES



Annexe 1. Graphique général du corpus



Annexe 2. Structure cluster 1 'Utilisation'

« niveau d'utilisation : intermédiaire
« redonnez un coup de jeune en profondeur à votre PC en utilisant ces deux logiciels ! vous ne le regretterez pas!
« une petite merveille, facile à installer, à utiliser, on ne peut plus clair
« la nouvelle version 5 est encore plus simple d'utilisation

Annexe 3. Verbatims cluster 1