Analyse de commentaires libres par la technique des réseaux de segments

Hervé Rostaing, Hélène Ziegelbaum, Eric Boutin, Michel Rogeaux, Luc Quoniam

ANALYSE DE COMMENTAIRES LIBRES PAR LA TECHNIQUE DES RESEAUX DE SEGMENTS

Hervé Rostaing[1], Hélène Ziegelbaum[2], Eric Boutin[3], Michel Rogeaux[2], Luc Quoniam[1]

[1]CRRM, Université Aix-Marseille III, 13397 Marseille Cedex 20
fax : 04.91.28.87.12 e-mail : crrm@crrm.univ-mrs.fr

[2]TEPRAL, 68 route d'Oberhausgergen, 67 037 Strasbourg Cedex
fax : 03.88.27.40.53 e-mail : hziegelbaum@www.tepral.fr

[3]LEPONT, Université de Toulon et du Var, IUT TC BP 132 83957 La Garde
fax : 04.94.14.22.75 e-mail : boutin@rhodes.univ-tln.fr

Les techniques d'analyse du contenu d'un corpus de textes sont multiples. Le traitement exposé ici est basé sur une approche hybride intégrant l'analyse de réseaux, employée en sciences sociales ou en bibliométrie, et la technique de segmentation utilisée en analyse statistique textuelle.
L'application de cette méthode dans le cadre d'étude d'analyse sensorielle est présentée. Ces études ont pour objet l'interprétation d'un corpus de commentaires libres proposés par des consommateurs soumis à des tests de produits agro-alimentaires. Ces commentaires étant saisis sous formes de textes électroniques, la mise en œuvre d'outils informatiques spécifiques a permis l'analyse de réseaux des segments présents dans ces commentaires. La première phase du traitement de ces commentaires est leur postcodage : correction orthographique ; réduction du vocabulaire par lemmatisation et synonymie ; marquage des termes ou locutions selon leur appartenance à des classes (arôme/odeur, hédonique, perception, saveur, texture, aspect, intensité des sensations) ; découpage du texte en segments. La seconde phase passe par le dénombrement des segments et de leurs associations, construction d'un tableau exprimant ces données. La dernière phase du traitement est la représentation de ce tableau sous la forme d'un réseau. L'outil informatique qui génère ce réseau permet le renvoi vers les commentaires contenant les noeuds du réseau ainsi qu'une navigation hypertexte.

Introduction

Dans l'industrie agro-alimentaire, il est important de pouvoir étudier les impressions des consommateurs lorsqu'ils goûtent un produit. Analyser les sensations des consommateurs soumis à des tests permet de connaître l'impact que provoque un produit. Les résultats de ce type d'étude peuvent jouer un rôle très important lors de la conception d'un nouveau produit comme lors du lancement d'un produit sur le marché.

Une des méthodes d'analyse sensorielle est basée sur la réalisation d'une enquête portant sur un panel de consommateur (Rogeaux et col. 1996). Comme pour la plupart des enquêtes deux catégories de questions sont posées :

- des questions fermées concernant le sexe du consommateur, son âge, sa position sociale, son taux de consommation, son lieu de consommation, son intention d'achat...
- des questions ouvertes pour que le consommateur s'exprime librement sur les sensations qu'il ressent lors du test.

Les questions concernant les impressions sensorielles sont volontairement posées sous forme ouverte pour favoriser la spontanéité de la réponse, pour ne pas influencer ou transformer le jugement du consommateur en lui offrant un nombre limité de modalités par réponse.

Ces commentaires libres sont des sources d'information très riches mais aussi très complexe à exploiter de façon automatique. Différentes techniques d'analyse textuelle peuvent aider à l'interprétation d'un tel corpus mais toutes ne sont pas appliquées pour les mêmes objectifs (Lebart 1995). La technique exposée dans cette communication cherche à construire une représentation globale et synthétique du corpus ou de sous-ensemble spécifiques. Nous cherchons dans la mesure du possible à établir une cartographie synthétique de l'ensemble des données traitées. Ainsi, un découpage et une analyse du corpus par produit permet d'envisager de la comparaison de leurs caractéristiques. Dans le cas d'un découpage par tranche d'âge, catégorie sociale... pour un produit, l'analyse laisse apparaître les différences d'appréciation du produit selon chaque catégorie. Ainsi de suite...
La méthodologie mise en oeuvre est une approche d'analyse de textes pratiquée en bibliométrie. La bibliométrie est plus particulièrement axée sur l'exploitation de corpus de textes représentant de références bibliométriques (Rostaing 1996). Les principes de cette discipline sont de dégager à partir d'un grand volume de notices bibliographiques les tendances générales de leurs contenus et d'offrir une grille de lecture en déterminant les structures sous-jacentes à ces données.

Les commentaires libres et le postcodage

Lors de ces enquêtes consommateurs, l'acquisition des commentaires se réalise soit sous forme orale (commentaires enregistrés) soit sous forme papier (commentaires saisies de façon manuscrite sur des formulaires). Des opératrices les saisissent ensuite électroniquement par audition des bandes enregistrées ou par lecture et décryptage des formulaires. Le mode même de cette acquisition de données engendre deux problèmes majeurs :

- un grand nombre de termes erronés : les fautes de frappe systématiques ou occasionnelles, les fautes d'orthographes, les erreurs de lecture, les fautes de français.
- une très grande hétérogénéité du vocabulaire et des expressions employés. Il peut même apparaître des expressions propres au discours oral (bof, beurk...)

Une telle diversité de termes impose un traitement préalable de correction des erreurs et de postcodage des commentaires pour réduire le vocabulaire et augmenter la signification des traitements statistiques ultérieurs. Cette démarche correspond tout à fait au principe statistique de la bibliométrie qui au détriment d'une perte d'information offre un gain de signification.

Ce postcodage passe par plusieurs étapes :

- corrections des erreurs répertoriées
- élimination des mots-outils
- repérage des locutions et liaison des termes qui les composent
- lemmatisation
- regroupement synonymique
- gestion des ambiguïtés (polysémie et homographie)
- marquage des termes spécifiques à l'analyse sensorielle

Les cinq premières étapes sont totalement automatisées grâce à l'établissement de lexiques spécifiques au produit alimentaire étudié (lexique des erreurs, des mots-outils, des locutions, des lemmes, des synonymes). Ces lexiques sont systématiquement appliqués aux données

brutes grâce à un logiciel de reformatage du commerce (*Infotrans*[*]). Un tel logiciel ne sait pas traiter les aspects de catégorisation grammaticale et de syntaxe de phrase. Seuls des traitements de reconnaissance et de manipulation de formes graphiques sont réalisables.
L'automatisation complète de la sixième étape nécessiterait une analyse sémantique impossible à envisager avec un reformateur. Elle n'est donc que semi-automatisée. Un lexique des termes potentiellement ambigus a été établi. Ce lexique permet de les « marquer » de façon à pouvoir les retrouver facilement en fin de traitement. Il faut alors lire le contexte pour évaluer par quel autre terme il doit être remplacé (une table des termes ambigus et de leurs remplaçants potentiels a été rédigée pour aider le correcteur).
La dernière étape est là encore basée sur l'emploi de lexiques. Cette fois-ci, non pour réduire le vocabulaire mais uniquement pour « marquer » les mots ou locutions très appréciés pour l'analyse sensorielle. Ainsi, 6 catégories sont construites : les termes faisant appel à l'*arôme*, au caractère *hédonique*, à la *perception*, à la *saveur*, à la *texture* et à l'*aspect*. Tous les termes appartenant à ces classes étant marqués (voir exemple ci-dessous), il devient plus facile de les manipuler pour construire les tableaux croisant les termes des différentes catégories.

Exemple :
Avant postcodage
GOUT AGREABLE. ARRIERE GOUT ASSEZ AMER MAIS NE SUIT EN RIEN LA QUALITE DU PRODUIT. TRES RAFRAICHISSANT
Après postcodage
*@GOUT *AGREABLE. @ARRIERE_GOUT ASSEZ µAMER. QUALITE TRES_FAIBLE BIERE. TRES RAFRAICHISSANT*

Tous ces lexiques sont bien évidemment remis à jour après analyse de chaque nouveau corpus de commentaires libres. Chaque étude apportant son lot de nouvelles fautes, de nouvelles expressions, de nouveaux synonymes, il est indispensable de les prendre en compte pour les traitements futurs. Ce système de postcodage est donc conçu pour un contexte évolutif.

La segmentation et le comptage des associations de segments

Les données obtenues après postcodage offrent plusieurs voies de segmentations. La première est de tout simplement considérer toutes séquences de caractères encadrés d'un espace ou d'un point comme étant des formes graphiques à dénombrer. Un problème se pose alors lorsqu'il faut comptabiliser les associations de formes graphiques. Il faut rappeler que l'objectif de ces études d'analyse sensorielle est de cartographier au plus juste chaque produit testé. Pour cela, non seulement la liste des sensations évoquées par les consommateurs est importante, mais encore plus les associations de sensations. Or dans le cas où l'unité statistique textuelle est celle indiquée ci-dessus, deux cas de comptage d'association sont envisageables.

Cas A : associations des termes intra-phrase
Seuls les termes appartenant aux mêmes phrases se retrouvent associés. Pour l'exemple présenté plus haut, les associations seront *@GOUT ↔ *AGREABLE, @ARRIERE_GOUT ↔ ASSEZ, @ARRIERE_GOUT ↔ µAMER, ASSEZ ↔ µAMER, QUALITE ↔ TRES_FAIBLE, QUALITE ↔ BIERE*...

[*] *Infotrans* :logiciel développé par *Information & Communication, Alte Str.66, D-79249 Freiburg Merzhausen*

Dans ce cas, les associations précisant que le consommateur a trouvé le produit *agréable* avec un *arrière-goût amer* ou *très rafraîchissant* avec un *arrière-goût amer* sont négligés. Or ce sont justement ce type d'associations qui paraissent le plus intéressantes.

Cas B : associations des termes intra et inter phrase
Pour essayer de récupérer les associations précédentes, il est possible alors de considérer tous les couples de termes intra et inter phrases. Ce comptage fait bien ressortir les associations omises précédemment comme **AGREABLE ↔ µAMER*, **AGREABLE ↔ RAFRAICHISSANT*, mais il prend aussi en compte des associations comme **AGREABLE ↔ @ARRIERE_GOUT*, *@ARRIERE_GOUT ↔ TRES_FAIBLE*, voire **AGREABLE ↔ TRES* ou *µAMER ↔ TRES.* Ces dernières associations sont indésirables et ne peuvent être prises en compte lors de l'analyse de la cartographie des associations.

C'est pour cela qu'une troisième solution a été envisagée. Puisque les phrases dans les commentaires libres sont le plus souvent très concises et que la phase de postcodage a réduit leur composition aux idées essentielles, on peut considérer ces phrases comme des entités très homogènes, comme des concentrés d'information. L'unité statistique élémentaire peut alors être ramenée à l'échelle de la phrase. La segmentation pour le dénombrement de ces unités se fait donc grâce au point. Les associations des segments obtenus sont comptabilisées uniquement à l'intérieur d'un commentaire libre. Dans notre exemple, ce traitement donne les associations suivantes : *@GOUT *AGREABLE ↔ @ARRIERE_GOUT ASSEZ µAMER*, *@GOUT *AGREABLE ↔ QUALITE TRES_FAIBLE BIERE*, *@GOUT *AGREABLE ↔ TRES RAFRAICHISSANT*, *@ARRIERE_GOUT ASSEZ µAMER ↔ QUALITE TRES_FAIBLE BIERE*, *@ARRIERE_GOUT ASSEZ µAMER↔ TRES RAFRAICHISSANT*, *QUALITE TRES_FAIBLE BIERE ↔ TRES RAFRAICHISSANT.*

La cartographie des associations de segments

Le dénombrement des fréquences d'apparitions des segments (phrases postcodées) ainsi que le dénombrement des fréquences des co-présences des couples de segments sont des processus totalement automatisés grâce à exploitation du logiciel bibliométrique *Dataview* développé par le CRRM (Rostaing 1993). Parmi bien d'autres types de résultats, ce logiciel permet de réexprimer ces comptages sous la forme d'un tableau symétrique distribuant en ligne et en colonne l'ensemble des segments présents dans les corpus analysés. Une cellule d'un tel tableau comporte dans la diagonale, la fréquence d'apparition d'un segment, et hors de la diagonal, la fréquence des co-présences d'un couple de segments. Le tableau obtenu est alors exporté vers le logiciel *Matrisme* spécialisé dans la génération automatique de réseaux. Ce logiciel, mis au point grâce à la collaboration entre LEPONT et le CRRM (Boutin et col. 1995), produit une représentation infographique du contenu du tableau sous la forme d'un réseau (voir figure 1). Les segments du corpus sont symbolisés par les noeuds du réseau tandis que les arcs reliant les noeuds représentent la fréquence de co-apparition des couples de segments. Contrairement aux analyses d'inertie, la position des noeuds les uns par rapports aux autres ne dépend pas d'une métrique mesurant les distances entre noeuds. Ces positions sont fonction d'une mesure d'évaluation de l'esthétisme du graphe obtenu (optimiser l'espace occupé, réduire le nombre d'intersections, interdire les chevauchements de noeuds, limiter la longueur des arcs). Seules les nuances de couleur (ou d'épaisseur) des arcs donnent des indications sur les intensités d'association entre les noeuds (les segments).

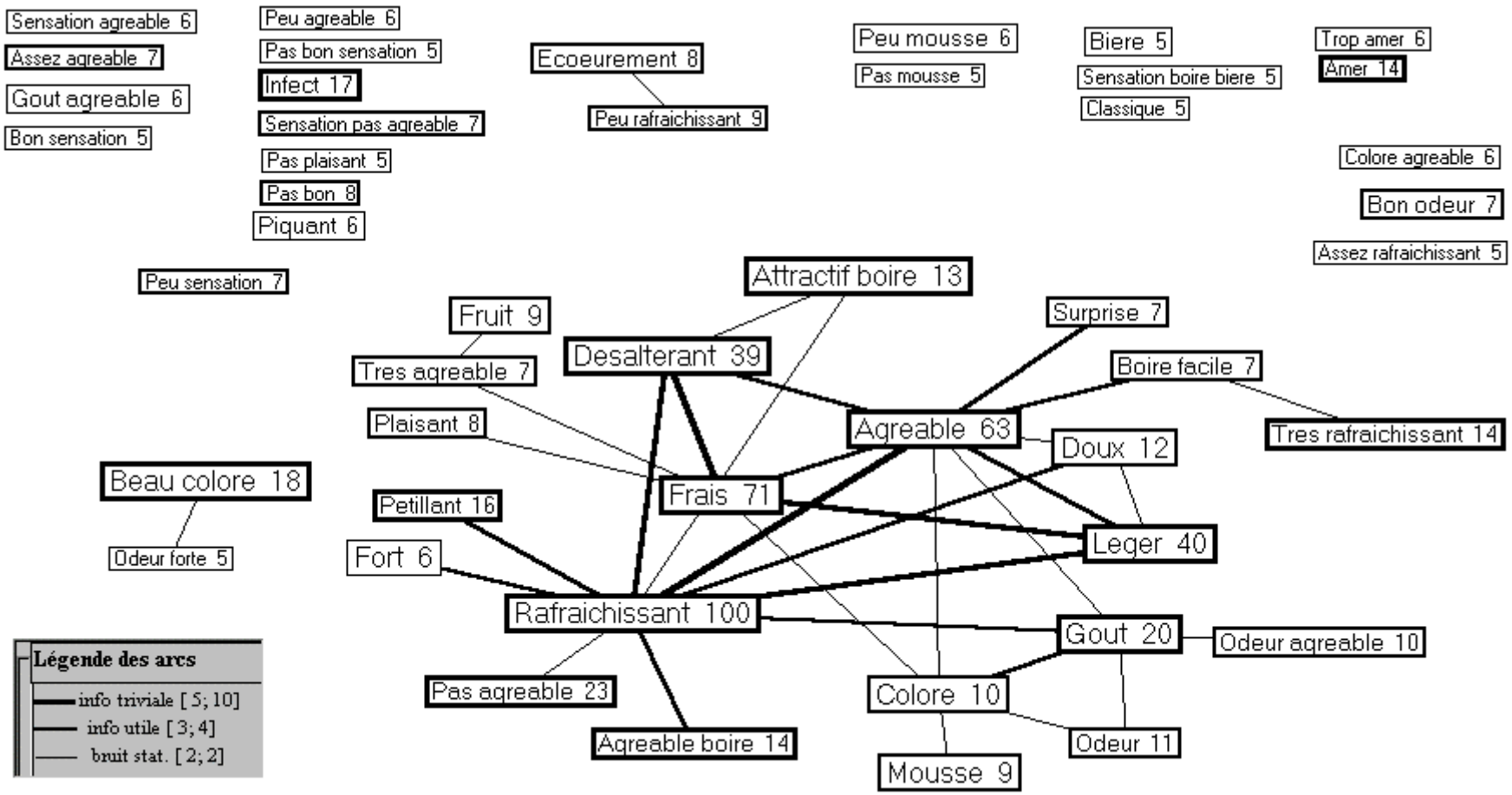

Figure 1 : Réseau de segments obtenus à partir de 1038 commentaires libres. Seuls les segments ayant une fréquence supérieure à 4 et seules les relations supérieures à 1 sont représentés sur ce réseau. La valeur présente à côté du segment correspond à sa fréquence. Les trois nuances graphiques des fréquences des co-présences de segments sont expliquées par la petite fenêtre intitulée *Légende des arcs.*

Conclusion

La technique de traitement automatique des commentaires libres de consommateurs qui a été mis au point est certainement encore perfectible, essentiellement dans sa phase de postcodage. L'approche de réduction du vocabulaire paraît indispensable si l'objectif est d'obtenir un gain de signification statistique suffisant. Les différentes phases de postcodage paraissent relativement bien au point pour permettre de répondre ne grande partie à cet objectif. Une analyse par catégorisation et une analyse sémantique pourrait nettement faciliter la phase de traitement des ambiguïtés mais ces approches seraient très coûteuses lors de leur mise au t-point. La technique choisie est peut-être frustre linguistiquement mais elle offre l'avantage d'être accessible à tous et rapide à mettre en place.

L'approche de segmentation des textes en phrase postcodée paraît bien appropriée au type de données collectées dans le cadre d'études d'analyse sensorielle (concision des phrases composées). Elle reste tout de même a être confortée lors d'étude ultérieure.

L'analyse des associations et des dépendances d'idées par la représentation cartographique sous forme de réseau est particulièrement bien adaptée à la phase d'interprétation. L'interprétation d'une étude ne peut s'envisager sans le soutien des professionnels du domaine étudié (dans notre cas des spécialistes en analyses sensorielles, des chercheurs, des commerciaux, des directeurs R&D...), il est préférable que les supports d'analyse ou de communication soit le plus accessible. La représentation réseau a cet avantage d'être compréhensible de tous sans aucun apprentissage spécifique, ce qui n'est pas le cas des méthodes basées sur un construit mathématique relativement complexe et difficilement explicable aux non-initiés (Boutin et col. 1996).

La possibilité de retourner aux commentaires originaux d'un segment par simple « clique » sur le noeud du réseau lui correspondant, puis de naviguer dans l'ensemble des commentaires par liens hypertextes en fonction des segments auxquels il est associé, offre un outil d'aide à l'interprétation et de validation incontestable. Cette fonctionnalité de génération automatique de fichier hypertexte à partir des commentaires originaux, structuré selon les associations exprimées dans le tableau analysé, est un atout supplémentaire. Pouvoir passer de la représentation synthétique du réseau aux données brutes qui ont permis de la construire est à nos yeux un instrument indispensable à la bonne réussite d'une telle analyse de contenu de textes.

REFERENCES

Rogeaux, M., Zieglebaum, H. (1996). Comment DANONE prend-il en compte les commentaires sensoriels des consommateurs après dégustation de boissons. *AGORAL 96 :* Lavoisier TEC&DOC, p139-147

Lebart, L. (1995). Analyse des données textuelles : quelques problèmes actuels et futurs. *JADT 1995 : Analisi Statistica dei Dati Testuali.* Università degli Studi di Roma : dalla Eurograf 2000

Rostaing, H. (1996). *La bibliométrie et ses techniques.* Toulouse : Sciences de la Sociétés

Rostaing, H. (1993). *La bibliométrie et la Veille Technologique : concepts, outils, applications.* Thèse

Boutin, E., Dumas, P., Quoniam, L., Rostaing, H., Dou, H. (1995). Génération automatique de réseaux bibliométriques. *Les systèmes d'informations élaborées 95.* Ile Rousse : SFBA.

Boutin E., Quoniam L., Rostaing H., Dumas P. (1996), Traitement de l'information : analyse des données classiques versus analyse réseau. Un cas d'application : la bibliométrie. *Inforcom'96.* Université Stendhal de Grenoble: Université Lille III, p. 571-587, 1996