Paroles de patients dans les forums de santé: une perspective originale sur la qualité de la vie

Thomas Opitz[1,2], Jérôme Azé[1], Sandra Bringay[1], Cyrille Joutard[2], Christian Lavergne[2], Caroline Mollevi[3]

[1] LIRMM, Université Montpellier, France
Thomas.Opitz@LIRMM.fr, Sandra.Bringay@LIRMM.fr, Jerome.Aze@LIRMM.fr

[2] I3M, Université Montpellier, France
Christian.Lavergne@univ-montp2.fr, Cyrille.joutard@univ-montp3.fr

[3] BIOSTATISTICS UNIT, Institut de Cancérologie de Montpellier, France
Caroline.Mollevi@icm.unicancer.fr

Résumé : Les forums de santé sur internet représentent une ressource textuelle riche, générée par les très nombreux échanges entre patients et, dans certains cas, professionnels de santé. Dans cet article, nous nous intéressons au problème de l'extraction d'informations médicales dans ces textes en nous focalisant sur des thèmes définis à partir d'auto-questionnaires utilisés dans les études cliniques. Il s'agit d'un problème difficile car les textes des forums sont peu structurés et comportent beaucoup de bruit comme des apartés ou des traits d'humour. Dans ce contexte, nous proposons une méthode originale pour l'expansion des requêtes basée sur la synonymie et limitant les faux positifs à l'aide d'une nouvelle mesure de similitude permettant d'écarter les expansions non pertinentes. Concrètement, à partir des thèmes abordés dans un questionnaire conçu pour le cancer du sein, nous quantifions et structurons les occurrences dans les messages d'un forum français spécialisé. Les extensions nécessaires pour un traitement statistique approfondi des informations retrouvées sont esquissées dans les perspectives de cet article.

Mots-clés : Cancer du sein ; Étude clinique ; Recherche d'information ; Médias sociaux ; Fouille de textes

1 Introduction

L'Organisation Mondiale de la Santé a défini en 1948 la qualité de vie (QdV) comme un concept multidimensionnel, subjectif et dynamique, portant sur les fonctions physiques, psychologiques et sociales. Ce concept fait référence à la perception que les patients ont de leurs maladies et de leurs traitements. La qualité de vie est un critère d'évaluation clinique alternatif pertinent pour évaluer les avantages et les inconvénients des traitements que ce soit pour le patient ou pour le système de santé. En 2012, 48,763 nouveaux cas de cancers du sein ont été recensés en France [1]. Si les traitements modernes permettent de sauver des vies, ils restent éprouvants. De nos jours, de nombreux projets de recherche clinique s'intéressent à la QdV des patients. Dans cet article, nous nous focalisons sur les histoires relatées par les patients dans les forums de santé en ligne, dans lesquels on retrouve de nombreux sujets de discussion liés aux symptômes de la maladie et aux effets indésirables des traitements. Hancock *et al.* (2007) ont montré que la communication anonyme via un ordinateur facilite l'expression d'états affectifs tels que les émotions, les opinions et les doutes. Typiquement, ces états affectifs sont réprimés plus fortement dans les contextes de communications plus traditionnels comme des interviews en face-à-face ou des enquêtes via des questionnaires.

1. Source : Institut National du Cancer, cf. www.e-cancer.fr

Récemment, l'impact des médias sociaux sur la santé a été largement étudié. Merolli *et al.* (2013) ont exploré les effets bénéfiques des médias sociaux dans le cas de maladies chroniques. Subirats *et al.* (2013) confirment que les réseaux sociaux favorisent la démocratisation et l'appropriation des connaissances par les internautes. Si la plupart des messages comportent des informations issues d'internautes non-professionnels et adressées à des non-professionnels, ces ressources sont une vraie mine d'or pour les professionnels de la santé. Dans notre étude, nous ciblons l'extraction automatique d'occurrences de thèmes d'intérêt dans les histoires relatées par des patients dans le contexte du cancer du sein. Ces informations sont utiles pour les oncologues qui les obtiennent souvent difficilement directement de leurs patients. Cette information centrée sur les patients peut les aider à : 1) comprendre comment les patients utilisent les forums par rapport à leurs besoins individuels ; 2) expliquer la cohérence ou la disparité existant entre différents résultats de convalescence ; 3) recueillir les perceptions et préférences individuelles des patients ; 4) évaluer la pertinence des items des questionnaires existants. Une étude similaire a été réalisée pour le médicament Januvia (Akay *et al.*, 2013) dans le but d'enquêter sur les avis des internautes et de détecter les utilisateurs influents.

Nous cherchons les messages du forum dont les thématiques sont associées aux éléments d'une liste de thèmes d'intérêt définie à partir de questionnaires remplis par les patients eux-mêmes durant leurs trajectoires de soins médicaux. Le nombre de messages dans les forums de santé est très important. Le forum *CancerDuSein.org* étudié ici comporte plus de 1,050 fils de discussions, dont certains comptent plus de 500 réponses. Le traitement informatique de ces données, basé sur des méthodes d'extraction d'informations semi-automatiques, reste un enjeu technologique considérable. En effet, la plupart des méthodes (semi-)automatiques appliquées au domaine de la santé l'ont été sur des données telles que les publications et les rapports hospitaliers. Leur adaptation directe à notre objet d'étude n'est pas possible car la structure des textes n'est pas standardisée (e.g. argot, fautes d'orthographe et de grammaire, sigles non standards). Le recours à une recherche textuelle simple afin d'associer les messages aux thèmes d'intérêts n'est pas efficace en raison de la richesse morphologique de la langue française. Par exemple, une requête comme *bouche sèche* ne permet pas de retrouver des occurrences comme *bouche desséchée* ou *langue sèche* (faux négatifs). Au contraire, des requêtes trop générales comme *bouche* renvoient de nombreux faux positifs (e.g. *bouche bée*). Par conséquent, l'expansion de requête permettant d'améliorer la description des thèmes par des termes sémantiquement proches est indispensable. Dans ce qui suit, nous proposons une méthode originale permettant de capter les messages d'intérêt dans les forums traitant du cancer et utilisant des ressources web afin de construire et valider des expansions de mots clés. Nous présentons les principaux résultats obtenus puis listons les perspectives importantes associées à ces travaux.

2 Méthodes

La méthode que nous proposons pour détecter des messages pertinents se structure en 4 étapes.

Étape 1 : collecte des messages. Le jeu de données comporte $16,961$ messages publiquement accessibles écrits par les 675 membres du forum *CancerDuSein.org* entre 2011 et 2013. Nous appliquons des pré-traitements classiques liés à la spécificité de la langue dans les forums

de santé. Un étiqueteur morpho-syntaxique [2] a été utilisé pour transposer les mots en lemmes et détecter des termes inconnus. Pour cela, nous calculons une variante de la distance *edit* (mesurant le nombre de lettres à modifier pour passer d'un mot à un autre) et remplaçons les termes inconnus par des termes proches trouvés dans un dictionnaire composé d'un dictionnaire général du français [3], des entités nommées (e.g. noms d'utilisateurs, entités géographiques) et d'une liste de médicaments [4] administrés aux patients atteints d'un cancer du sein.

Étape 2 : Identification des thèmes et enrichissement morphologique. Nous définissons une liste de thèmes d'intérêt sur la base du questionnaire EORTC QLQ-C30 spécifique au cancer et internationalement reconnu (Aaronson *et al.*, 1993). Ce questionnaire évalue des échelles de fonctions, de symptômes, de l'état global de santé et de la QdV. Il est souvent associé à des modules spécifiques comme le module EORTC QLQ-BR23 pour le cancer du sein. Afin de définir une liste de thèmes, nous utilisons les échelles fonctionnelles (image du corps, sexualité) et les échelles de symptômes (effets indésirables, symptômes au niveau du sein et du bras, perte de cheveux) de ce dernier questionnaire. Un expert définit les thèmes manuellement à partir des items du questionnaire. Au départ, nous représentons un thème par un ensemble thématique $T_{\mathrm{I}} = \{\mathrm{TTI}\}$ composé de n_{I} termes thématiques initiaux TTI, chacun associé à un ou plusieurs lemmes (sans les mots vides définis dans l'implémentation du `TreeTagger`). Certaines variations morphologiques fondamentales telles les noms/verbes/adjectifs sont prises en compte et amènent à n_{M} termes thématiques supplémentaires TTM dans l'ensemble thématique morphologiquement étendu $T_{\mathrm{IM}} = \{\mathrm{TTI}\} \cup \{\mathrm{TTM}\}$. Par exemple, l'item *Avez-vous eu la bouche sèche ?* est représenté par $T_{\mathrm{I}} = \{$*bouche*&*sec*$\}$. En utilisant les variantes morphologiques *buccal* pour *bouche* et *asséché*, *dessèchement*, ... pour *sec*, nous obtenons $T_{\mathrm{IM}} = \{(\textit{bouche}|\textit{buccal})\&(\textit{sec}|\textit{asséché}|\textit{dessèchement}|\ldots)\}$.

Étape 3 : Enrichissement synonymique automatique. Les variations morphologiques trouvées par l'expert ne couvrent pas toutes les variations lexicales dans les forums. Nous enrichissons T_{IM} en déterminant des termes thématiques synonymes TTS dans le but de proposer l'ensemble thématique $T_{\mathrm{IMS}} = T_{\mathrm{IM}} \cup \{\mathrm{TTS}\}$. Les termes médicaux présents au même niveau dans le thésaurus MeSH (*Medical Subject Headings*) [5] sont utilisés comme synonymes. Les ensembles thématiques résultants sont loin d'être exhaustifs car le vocabulaire technique du MeSH est utilisé principalement par les professionnels de santé et non par les patients. Pour cette raison, nous utilisons des outils de synonymie disponibles en ligne [6] pour enrichir notre liste d'expansions. Par ailleurs, la synonymie dépend fortement du contexte. À titre d'exemple, considérons le terme thématique <*bouche&sec*> dont <*langue&sec*> représente une expansion utile, contrairement à <*bouche&indifférent*>. Nous proposons donc une validation web pour mesurer la proximité d'une expansion synonymique à l'ensemble thématique concerné T_{IM}. L'originalité de notre approche consiste à ne pas considérer directement les nombres d'occurrences comme l'ont fait Roche *et al.* (2012); Turney (2001), mais la proximité de contextes construits pour T_{IM} et pour chaque expansion potentielle TTS. Le contexte de T_{IM} est construit à partir des co-occurrences les plus fréquentes dans un grand corpus. Dans ce but, nous utilisons

2. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger`
3. `http://www.aspell.net`
4. `http://medicament.comprendrechoisir.com`
5. `http://mesh.inserm.fr/mesh/`
6. `http://synonymo.fr` et `http://www.cnrtl.fr/synonymie`

les $n_{\mathrm{I}}+n_{\mathrm{M}}$ termes thématiques de T_{IM} comme requêtes soumises à un moteur de recherche web (`yahoo.fr`). Le contexte thématique est défini comme un vecteur de motifs pondérés, extraits des N_{snip} premiers snippets renvoyés pour chaque requête. Les motifs considérés sont les unigrammes et bigrammes des noms, verbes et adjectifs lemmatisés. La pseudo-fréquence (ou le *poids*) $f(m \mid T_{\mathrm{IM}})$ d'un motif m est calculé selon sa fréquence dans les $(n_{\mathrm{I}} + n_{\mathrm{M}}) \times N_{\mathrm{snip}}$ snippets et sa fréquence dans la langue française [7], notée $N(m \mid \mathrm{French})$. Une expression semblable à l'approche $\mathrm{Tf\text{-}Idf}$ couramment utilisée (Salton & Buckley, 1988) est appliquée afin de diminuer le poids des motifs peu discriminants :

$$f(m \mid T_{\mathrm{IM}}) = N(f \mid T_{\mathrm{IM}}) / \left(N_{\mathrm{snip}} \times (n_{\mathrm{I}} + n_{\mathrm{M}}) \times \log N(m \mid \mathrm{French})\right). \quad (1)$$

Pour les bigrammes, $N(m \mid \mathrm{French})$ est pris comme la fréquence minimale des deux mots constituant le motif. Nous doublons le poids $f(m \mid T_{\mathrm{IM}})$ des bigrammes qui représentent des motifs fortement discriminants. Notons A_i, $i = 1, \ldots, k$ les *poids attendus*, c'est-à-dire les poids des k meilleurs motifs f_i ordonnés par ordre décroissant des poids après l'éloignement des motifs compris dans l'expansion proposée TTS. La procédure pour définir le contexte de chaque expansion TTS est la même : on soumet la requête $< \mathrm{TTS} >$ au moteur de recherche et on remplace T_{IM} par TTS et $(n_{\mathrm{I}} + n_{\mathrm{M}})$ par 1 dans l'équation (1). On note O_i, $i = 1, \ldots, k$ les *poids observés* résultant des motifs f_i pour le contexte de l'expansion. Plusieurs mesures ont été expérimentées pour la proximité entre une expansion TTS et un ensemble thématique T_{IM}, comme la similarité cosinus ou l'information mutuelle (Roche *et al.*, 2012). Ici, nous avons retenu un score de proximité S basé sur une variante de la statistique khi-deux :

$$S = 1 - \left(\sum_{i=1}^{k} A_i\right)^{-1} \times \sum_{i=1}^{k} \max(A_i - O_i, 0)^2 / A_i. \quad (2)$$

Les valeurs du score S se situent entre 0 pour des contextes disjoints et 1 pour des contextes identiques. La troncature à 0 évite de pénaliser des occurrences plus fréquentes dans le contexte de l'expansion que dans la thématique. Enfin, nous ajoutons une expansion TTS à T_{IMS} si son score est suffisamment élevé, par exemple $S > S_0$ pour un certain seuil fixé $0 < S_0 < 1$. Par exemple, l'expansion **<*langue&sec*>** proposée pour le thème dont le terme initial est **<*bouche&sec*>** obtient un score $S = 0.40$. En utilisant le seuil expérimental de $S_0 = 0.2$ choisi lors de nos expériences, nous aujoutons cette expansion à T_{IMS}.

Étape 4 : Requêtes dans la base des messages et synthèse globale. Nous détectons les occurrences des thèmes dans les messages en soumettant les requêtes construites à partir de l'ensemble thématique enrichi T_{IMS}. Les mots clés constituant un terme thématique sont composés par l'opérateur "ET". Ensuite, les termes thématiques sont réunis par l'opérateur "OU". La requête doit apparaître dans un intervalle de N_{lem} lemmes au maximum, par exemple $N_{\mathrm{lem}} = 12$ dans notre application. Selon l'objectif de l'étude, les occurrences renvoyées seront représentées en termes du nombre d'occurrences pour les thèmes, de la phrase d'occurrence, ...

3 Résultats

Le questionnaire QLQ-BR23 avec 23 items est représenté par 13 thèmes agrégés dans 5 groupes thématiques selon les dimensions du questionnaire : *image du corps*, *sexualité*, *effets*

7. `http://eduscol.education.fr/cid50486/liste-de-frequence-lexicale.html`

indésirables (excepté *perte de cheveux*), *perte de cheveux*, *symptômes au niveau du sein*. Pour chaque thème, nous calculons le score S pour un nombre maximal de 2500 expansions proposées. Le nombre de snippets est fixé à $N_{\text{snip}} = 40$ pour chaque requête au moteur de recherche. Nous comparons les contextes sur la base des $k = 50$ motifs les plus pertinents du contexte thématique. Afin d'écarter des expansions non pertinentes et polysémiques, une expansion proposée est retenue lorsque son score S dépasse le seuil $S_0 = 0.2$ déterminé de façon expérimentale. La requête finale T_{IMS} utilisée pour trouver les occurrences dans les messages du forum est composée de 31 termes initiaux dans T_{I}, de 70 expansions morphologiques supplémentaires dans T_{IM} et de 596 expansions synonymiques validées automatiquement. En fixant un seuil S_0 en-dessous de 0.2 (par exemple, $S_0 = 0.1$), nous avons constaté une forte dégradation en termes de précision. Toutefois, on détecte un nombre considérable de vrais positifs supplémentaires améliorant le rappel, d'où nous mandatons un expert pour une sélection manuelle des expansions thématiques pertinentes ayant un score S entre 0.1 et 0.2. La figure 2 présente le nombre de mes-

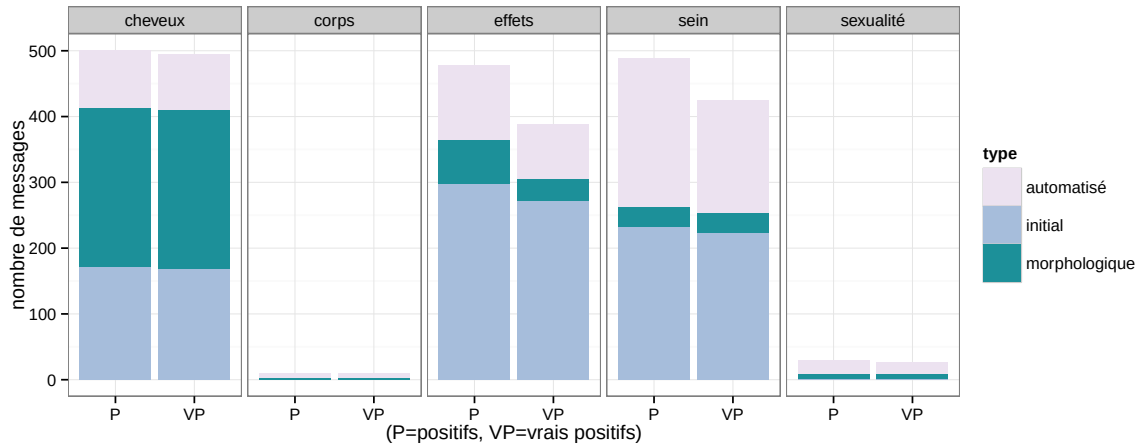

FIGURE 1 – Les occurrences trouvées selon leur groupe thématique et le type du terme détecté.

sages détectés selon la requête générée. Les occurrences d'un terme sont comptabilisées dans la catégorie *initial* si issues de T_{I}. Les occurrences des variantes morphologiques sont comptabilisées dans la catégorie *morphologique*. Seules les occurrences restantes sont attribuées à la catégorie *automatique*. Nous vérifions la cohérence des occurrences avec le thème cherché à l'aide d'une validation manuelle par deux experts informaticiens nous permettant d'évaluer la précision sur un sous-échantillon stratifié de taille 400. L'accord mesuré par un Kappa de Cohen est 0.86 (accord favorable). En général, les expansions automatiques ont augmenté le nombre d'occurrences de façon significative sans provoquer une perte de précision. Le nombre faible de messages associés à la sexualité et à l'image du corps est surprenant. Les variations lexicales caractérisant ces thèmes semblent être plus hétérogènes et donc moins accessibles à la détection automatique. L'occurrence rare d'un thème peut représenter un résultat intéressant en lui-même qui indique des thèmes jouant un rôle secondaire dans l'actualité de la vie quotidienne des patients.

4 Conclusion

Une étude de la QdV basée sur les discussions dans les forums de santé représente une alternative intéressante à l'analyse de la QdV basée sur l'évaluation des auto-questionnaires. Le nombre considérable des occurrences que nous avons détectées pour certains thèmes est un résultat prometteur pour une analyse plus profonde basée sur la classification des émotions (e.g. peur, joie) associées à ces occurrences. Ces résultats pourront être contrastés avec des résultats plus classiques obtenus dans le contexte des études cliniques. Nous espérons enrichir la description des messages en détectant par exemple les messages produits par les patients avant ou après une chirurgie. L'ensemble de ces métadonnées seront utilisées pour réaliser un traitement statistique plus poussé permettant de suivre l'évolution des préoccupations des patients. En outre, nous projetons d'explorer certaines extensions algorithmiques, à savoir une automatisation des expansions morphologiques et l'utilisation de relations comme l'antonymie dans les expansions thématiques automatiques. D'autres méta-données du forum comme ses sous-catégories et la structure des fils de discussion pourront être exploitées. Enfin, afin de diminuer le nombre de faux négatifs renvoyés par une requête, il peut être plus efficace de construire les thèmes différemment dans un effort collaboratif des experts cliniciens et des spécialistes de la fouille de textes afin d'adapter leur définition pour une détection plus facile dans les médias sociaux. Notre approche est une première étape vers une analyse plus globale des sentiments des patients.

Références

AARONSON N. K., AHMEDZAI S., BERGMAN B., BULLINGER M., CULL A., DUEZ N. J., FILIBERTI A., FLECHTNER H., FLEISHMAN S. B., DE HAES J. C. *et al.* (1993). QLQ-C30 : a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, **85**(5), 365–376.

AKAY A., DRAGOMIR A. & ERLANDSSON B.-E. (2013). A novel data-mining platform leveraging social media to monitor outcomes of Januvia. In *35th Annual International Conference on Engineering in Medicine and Biology Society (EMBC)*, p. 7484–7487 : IEEE.

HANCOCK J. T., TOMA C. & ELLISON N. (2007). The truth about lying in online dating profiles. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 449–452 : ACM.

MEROLLI M., GRAY K. & MARTIN-SANCHEZ F. (2013). Health outcomes and related effects of using social media in chronic disease management : A literature review and analysis of affordances. *Journal of Biomedical Informatics*, **46**(6), 957–969.

ROCHE M., GARBASEVSCHI O. M. *et al.* (2012). WeMiT : Web-mining for translation. In *Conference on Prestigious Applications of Intelligent Systems*, p. 993–994.

SALTON G. & BUCKLEY C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, **24**(5), 513–523.

SUBIRATS L., CECCARONI L., LOPEZ-BLAZQUEZ R., MIRALLES F., GARCÍA-RUDOLPH A. & TORMOS J. M. (2013). Circles of health : Towards an advanced social network about disabilities of neurological origin. *Journal of Biomedical Informatics*, **46**(6), 1006–1029.

TURNEY P. (2001). Mining the web for synonyms : PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML)*.