Un cadre fédérateur de représentation des données et indices issus des forum de santé

Nathalie Bricon-Souf, Ghazar Chahbandarian, Lydia-Mai Ho-Dac, Mustapha Mojahid

Un cadre fédérateur de représentation des données et indices issus de forums de santé

Nathalie Bricon-Souf[1], Ghazar Chanbandarian[1], Mai Ho-dac[2], Mustapha Mojahid[1]

[1] Laboratoire IRIT, Université Toulouse3
{prénom.nom}@irit.fr
[2] Laboratoire CLLE, Université Toulouse2
{prénom.nom}@univ-tlse2.fr

Résumé : L'analyse des forums de santé est un enjeu important car ces espaces particuliers de discussions abordent, de façon novatrice, la connaissance médicale et le ressenti des patients. L'enjeu consiste également à proposer des outils facilitateurs pour contextualiser l'accès aux informations du forum en prenant en compte différents profils utilisateurs. Mais avant de tester quels indices caractérisant les discussions contenues dans les forums nous permettront de proposer ces outils avancés, nous avons besoin d'un cadre de référence pour représenter tant les informations directement contenues dans les forums que celles que nous pouvons déduire de leur analyse. Nous présentons ici le travail effectué pour élaborer ce cadre de référence s'appuyant sur les recommandations TEI (text encoding initiative). Les pages web des forums sont standardisées sous forme de documents XML et associées à des fichiers d'annotations contenant les indices émotionnels, discursifs, contextuels retenus. Enfin, nous abordons les perspectives qu'offrent ces travaux.

Mots-clés :Informatique médicale ; Forum de santé ; encodage, annotation.

1 Introduction

L'avènement du Web 2.0 permet depuis plusieurs années à tout un ensemble d'utilisateurs de s'informer et d'échanger à travers des outils mis à disposition sur Internet. L'utilisation des forums de santé est l'un des aspects grand public très généralisé et de récentes études montrent que 24% des *e*-patients utilisent Internet au moins une fois par jour pour y rechercher des informations à propos de leur santé [1]. Les forums de santé ont, dans ce cadre, une grande importance.

Il est intéressant de regarder attentivement les échanges effectués au travers des forums de santé car ils possèdent des spécificités qui ne se retrouvent absolument pas lors des échanges traditionnels mis en place entre les professionnels de santé et les patients. Asynchrones et souvent anonymes (ou tout au moins partiellement via l'utilisation de pseudos), les discussions de forum traitent par exemple de questions que les patients n'osent pas aborder, ou n'ont pas le temps de formuler ou d'envisager dans une consultation en face à face qui se déroule sur un temps très limité. Via les forums de santé, les patients anticipent leurs consultations, cherchent des explications, demandent du réconfort à leurs pairs, s'interrogent sur un statut d'aidant, cherchent des solutions alternatives ou des références avérées pour les soins, et ceci dans un contexte subjectif, peu normé, et pour lequel chacun peut choisir son implication (de l'internaute très actif au "*lurker*"qui se contente de lire les informations [2]). Les différents fils de discussions sont alors très divers, présentant des contenus informatifs ou

émotionnels variables mais recèlent des informations sur la prise en charge de la santé des individus qu'il convient d'explorer.

Nous avons identifié trois grands axes d'utilisations potentielles que seront :

- une aide à la navigation dans les forums pour le e-patient qui fournirait un aperçu global de la contenance d'un fil en caractérisant visuellement le type d'information contenue dans les différents messages (aspects émotifs, qualité de l'information, rôle dans l'interaction),
- une aide à la modération des forums de santé afin d'alléger par des systèmes intelligents le travail de modération à mener,
- une aide à l'extraction pour les professionnels de santé afin qu'ils puissent bénéficier des retombées de l'émergence de connaissances enfouies dans cette masse d'information que sont les fils de discussion en santé.

Le travail que nous présentons ici s'inscrit dans une collaboration de plusieurs équipes de recherche nationales s'inscrivant dans le cadre du projet interMSH "patient's mind", soutenu par la Maison des Sciences de l'Homme de Montpellier. Les travaux en cours ont ainsi proposé des analyses linguistiques sur les contenus textuels de messages de forums de santé et permettent d'extraire des informations variées telles que par exemple les traces d'émotion ou d'incertitude[3,4], de repérer les citations ou de retrouver les termes médicaux employés par les rédacteurs des messages mais également de proposer des outils de visualisation novateurs pour ces forums.

L'objectif du travail décrit dans ce papier est de proposer un cadre fédérateur de représentation des données de forums ainsi que des différentes couches d'annotations qui peuvent en être faites, afin de proposer des outils (notamment de visualisation et d'analyse) cohérents et généralisables quel que soit l'aspect sur lequel nous souhaitons nous focaliser dans le traitement des discussions. Nous souhaitons être ainsi à même de tester l'impact des combinaisons de traits et annotations mis à disposition par les différents travaux de recherche entrepris et disposer d'un outil exploratoire pour la visualisation enrichie des forums de santé.

2 Les données des forums

2.1 Un cadre de réflexion sur la nature des données forums

Les forums de discussion contiennent des fils de discussions : successions de prises de paroles par différents intervenants et peu ou prou orientées sur une idée principale mentionnée par le titre du fil de discussion, ils colligent les informations partagées par les différents intervenants sur ces thèmes. Afin d'affiner la clarté des informations proposées certains forums regroupent les fils en catégories structurantes, voire lorsque le forum est très généraliste tel *Doctissimo*, en une hiérarchie de catégories apportant une précision quant aux thèmes abordés par le fil.

Les messages constituant les fils de discussion peuvent s'envisager comme des *notes de communications* [5] et possèdent un certain nombre de caractéristiques inhérentes à ce type d'échange.

Ainsi le message

- est écrit par le **Producteur de la note**
- transmet un **Contenu**
- a une **Forme**
- est lié à une **Cible** par une **Ancre**
- s'intéresse à un **Objet de communication**
- est écrit dans un **Espace spatiotemporel de création**

- utilise une **Force illocutoire** (détermine dans quel sens il faut comprendre le texte)
- est écrit avec une **Intention de communication** (argumenter, constater…)
- est destiné à un type de **Lecteur cible**
- participe à une activité de **Lecteur**
- est lu dans le cadre d'une **Activité collaborative**
- est lu dans un **Espace spatio-temporel de lecture**

Les informations distinguées en première partie de l'énumération sont celles que nous allons pouvoir assez facilement extraire des données de forum. Le producteur de la note est parfois source d'informations multiples : pseudo, profil et signature. Le contenu est textuel mais s'enrichit d'une mise en forme matérielle (retraits, gras, ponctuation, ...) et d'artifices de communication (émoticons, répétitions de lettres ...). Cible, ancre et objet de communication se détectent via l'url et le topic des messages de forums. L'espace spatiotemporel de création est souvent réduit à une information temporelle (avec estampille de la date et de l'heure du post), certains sites peuvent néanmoins donner lieu à une récupération d'informations spatiales pour des modérateurs, liées à une localisation d'adresse IP.

Acquérir les informations mentionnées dans la seconde partie de l'énumération permettra de construire une exploitation intelligemment ciblée des forums de santé et constitue l'un des objectifs à long terme d'un tel projet.

Le fil de discussion s'apparente quant à lui à un dialogue avec des prises de paroles successives, menées par différents "acteurs"

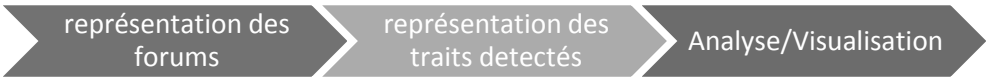
2.2 Structures accessibles des données forums

Les informations des forums sont le plus souvent accessibles au travers des pages web des sites qui permettent leur accès, et sont accessibles en HTML, ce qui présente tous les défauts de confusion entre forme et fond et ne permet pas facilement d'accéder à la structuration des données qui sont présentées.

3 Processus de traitement des données forums

Le processus mis en œuvre s'articule autour des trois étapes suivantes ;

- la représentation des forums : proposer un format permettant de représenter de façon homogène et cohérente le plus possible d'informations contenues dans les pages des forums.
- la représentation des traits détectés : proposer un format permettant d'encoder les différentes annotations disponibles actuellement ou à venir suite aux travaux d'analyse et de caractérisation des forums
- l'analyse ou visualisation : proposer l'exploitation des forums et de leurs caractérisations afin de produire un résultat pour l'utilisateur final. Dans le cadre de ce papier, seule la visualisation d'information a été envisagée.



3.1 Représentation des données forums

3.1.1 Format

Nous avons choisi un format XML pour être l'un des pivots de représentation des données de forums.

L'encodage des documents en respectant les recommandations TEI (Text Encoding Initiative) permet de représenter des corpus de texte de natures variées, il est souvent utilisé par les linguistes et facilite la normalisation et l'échange de documents textuels via un encodage des corpus en XML [6]. Nous nous sommes inspirés de la norme TEI P5 pour constituer le cadre de représentation des fils de discussion contenus dans les messages.

```
<sourceDesc>
    <biblStruct>
<!-- information sur la provenance du document -->
        <analytic>
            <title type="full">
                <title type="main">Nom du forum (name_for)</title>
                <title type="sub">Titre du fil de discussion (name_top)</title>
            </title>
            <date>Date de création du topic</date>
            <editor>Editeur du contenu des documents (nom du site web)</editor>
            <textLang>Langue des messages</textLang>
        </analytic>
    </biblStruct>
</sourceDesc>
    </fileDesc>
</teiHeader>
<!-- FIN de l'en-tête -->
```

Figure 1 : extrait du tei-header permettant l'encodage d'information liées au fil de discussion traité.

Les messages apparaissent alors à l'intérieur de ce corpus, des éléments de description XML suffisants pour répertorier les informations ont été prévus et permettent de représenter les informations de communication que nous avons soulignées en §1.

3.1.2 Validation

Cette représentation a été validée pour différents types de forums (*Doctissimo ; Cancer du sein ; Vivre sans thyroïde*) et a également été confrontée à un schéma de base de données relationnelle permettant de mobiliser les connaissances d'autres forums de santé mises au point dans le cadre du projet *patient's mind*.

Nous proposons ainsi un format générique, permettant la constitution d'un corpus de pages, facile à représenter et à échanger et capable de stocker les discussions des forums de santé via une structuration sémantique XML.

3.2 Représentation des traits détectés.

L'objectif est ensuite de permettre d'adosser à la représentation des informations liées aux forums de santé les informations issues des travaux de caractérisation de ceux-ci. La liste des caractéristiques que nous souhaitons aborder n'est ni stabilisée, ni finie et les caractéristiques sont de natures hétérogènes. Rappelons par exemple les indices suivants étudiés pour l'instant : émotion, incertitude, emphase via de la ponctuation, citation, présence de termes médicaux, taille des prises de parole, ... Nous souhaitons non seulement étudier en quoi chacune des caractérisations permet d'accroitre les connaissances issues des forums de santé mais également permettre de trouver quelles combinaisons pertinentes de caractérisations nous pouvons proposer pour extraire au mieux ces connaissances.

3.2.1 Format

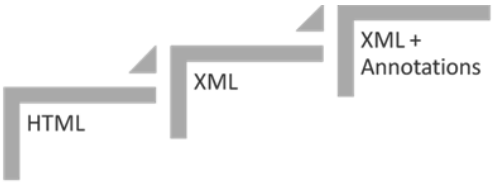
Nous avons choisi de représenter sous forme de fichier d'annotation les données d'analyse décrivant nos corpus. En proposant un fichier d'annotations spécifique à chacune des caractéristiques retenues, nous permettons la multiplicité des caractérisations retenues. En

choisissant un format d'annotation générique, nous permettons d'homogénéiser des caractérisations hétérogènes.

Nous avons ainsi décidé d'inscrire dans un fichier d'annotation de type *stand-off* [7], qui marque début et fin de chacune des annotations par une position absolue de l'annotation dans le corpus, chacun des travaux de caractérisation proposés. Ceci préserve le fichier d'origine et permet une superposition des différentes couches de descriptions des données.

3.2.2 Validation

Des annotations simples à mettre en œuvre car générées automatiquement via une comparaison entre dictionnaires et contenus textuels des forums ont servi de test pour ces propositions. Ainsi un fichier d'annotation des ponctuations, un fichier d'annotation des émoticons et un fichier d'annotation de termes médicaux restreints ont été mis en place. L'intégration avec des travaux d'annotations plus complets nécessitant des travaux linguistiques plus élaborés ou une annotation manuelle sera intégrée dans la suite du projet.



4 Résultats

Nous avons tout d'abord réalisé et testé un logiciel nous permettant de proposer une représentation XML unifiée pour les fils de discussions de différents forums de santé : doctissimo ; cancer du sein et vivre sans thyroïde ont ainsi été testés.

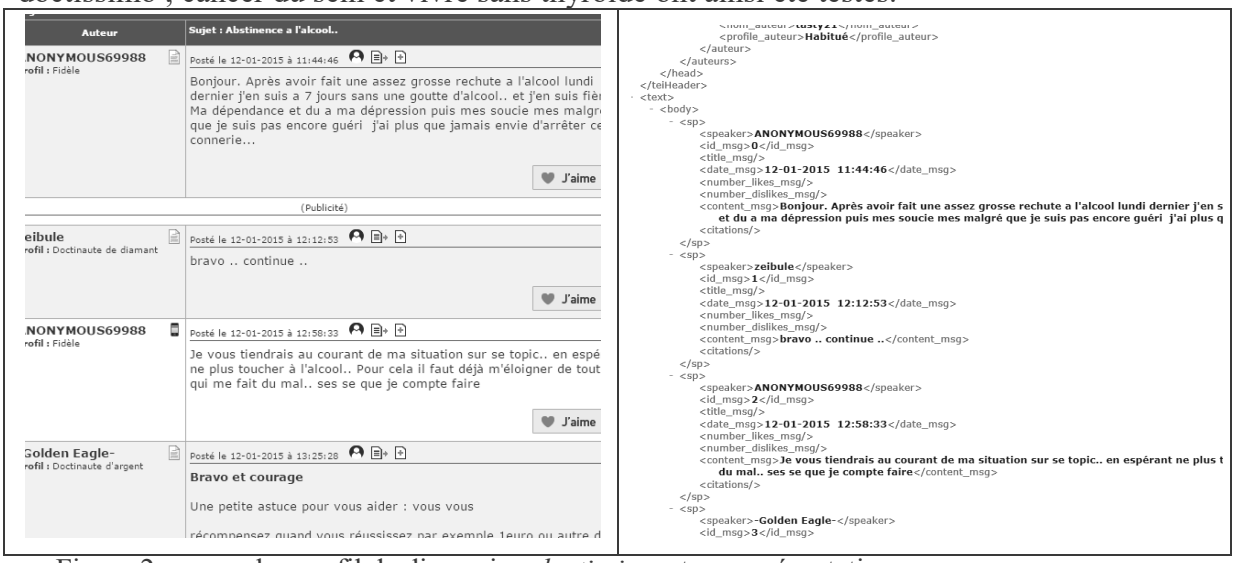


Figure 2 : exemple : un fil de discussion *doctissimo* et sa représentation

Nous avons ensuite mis en place l'annotation automatique de certains traits sur les corpus retenus.

```
<annotation mot="cure" type="Medical" positionEnd="135" positionStart="131" id_msg="0"/>
<annotation mot="cure" type="Medical" positionEnd="333" positionStart="329" id_msg="0"/>
<annotation mot="cure" type="Medical" positionEnd="634" positionStart="630" id_msg="0"/>
<annotation mot="cure" type="Medical" positionEnd="901" positionStart="897" id_msg="0"/>
<annotation mot="cure" type="Medical" positionEnd="1817" positionStart="1813" id_msg="0"/>
<annotation mot="cure" type="Medical" positionEnd="1988" positionStart="1984" id_msg="0"/>
```

Figure3 : exemple d'annotation d'un terme médical dans le fil de discussion.

Nous avons adapté un outil simple de visualisation de caractéristiques des forums afin de prendre en compte les formats de représentation XML et les fichiers d'annotation proposés. Nous pouvons ainsi proposer à l'utilisateur d'explorer les visualisations de traits "bruts", caractéristiques des messages contenus dans la description XML de ceux-ci, mais également de s'appuyer et de sélectionner les différents types d'annotations mis à disposition par les différents travaux de caractérisation des corpus mis en place.

5 Conclusion et Discussion

L'un des enjeux à long terme de nos travaux est de pouvoir associer à certains schémas de discussion une qualité informationnelle du fil étudié (exemple une discussion qui diverge du thème initial et se recentre entre 2 personnes devient peut être un aparté hors sujet ; une personne qui utilise dans sa question un registre émotif fort a sûrement plus besoin de réconfort de la part de pairs que d'informations factuelles, ...). Nous souhaitons pouvoir explorer les différents indices contenus ou déduits des messages échangés dans les forums afin de proposer des solutions pour mieux interpréter les caractéristiques de communications induites et aider les différents utilisateurs de forums via ces nouvelles connaissances. En particulier nous souhaitons explorer des indices de types Mise en Forme Matérielle [8] ou contexte afin de compléter les caractéristiques issues de l'analyse linguistique. En proposant un modèle de représentation stable des données des forums et des annotations d'indices ainsi que décrit dans ce papier, nous espérons faciliter grandement le travail exploratoire qui nous reste à mener dans ce cadre.

Références

[1] HON (Health On the Net) How Do General Public Search Online Health Information? Avril 2011 http://www.hon.ch/Global/pdf/Khresmoi/KRESMOI_internet_health_search_information_HON.pdf

[2] Chen, F. C. (2004, June). Passive forum behaviors (lurking): A community perspective. In *Proceedings of the 6th international conference on Learning sciences* (pp. 128-135). International Society of the Learning Sciences.

[3] Bringay, S., Kergosien, E., Pompidor, P., & Poncelet, P. (2014). Identifying the targets of the emotions expressed in health forums. In *Computational Linguistics and Intelligent Text Processing* (pp. 85-97). Springer Berlin Heidelberg.

[4] Thoumelin, P. C., & Grabar, N. (2014). La subjectivité dans le discours médical: sur les traces de l'incertitude et des émotions. In *EGC* (pp. 455-466).

[5] Bricon-Souf, N., Bringay, S., Hamek, S., Anceaux, F., Barry, C., & Charlet, J. (2007). Informal notes to support the asynchronous collaborative activities. *International journal of medical informatics*, *76*, S342-S348.

[6] TEI P5 Guidelines 2013 http://www.tei-c.org/Guidelines/P5/

[7] Pose, J., Lopez, P., & Romary, L. A Generic Formalism for Encoding Stand-off annotations in TEI. https://hal.inria.fr/docs/01/06/15/48/PDF/A_Generic_Formalism_for_Encoding_Standoff_annotations_in_TEI.pdf

[8] Kamel, M., Mojahid, M., & Rothenburger, B. (2012). " Quand rédiger c'est décrire"-Mise en forme matérielle des textes et construction d'ontologies à partir de textes. In *23es Journées Francophones d'Ingénierie des Connaissances-IC 2012* (pp. 133-148).

Remerciements

Nous remercions ici les partenaires du projet inter-MSH Patient's Mind ainsi que les étudiants de l'école d'ingénieur ISIS : Jonathan Brouttier, Lucas Leprêtre, Meggan Taussac et du master de linguistique : Marie Voisin, pour leur aide précieuse lors de ces travaux.