HAL open science

Breast cancer and quality of life: medical information extraction from health forums

Thomas Opitz, Jérôme Azé, Sandra Bringay, Cyrille Joutard, Christian Lavergne, Caroline Mollevi

To cite this version:

Thomas Opitz, Jérôme Azé, Sandra Bringay, Cyrille Joutard, Christian Lavergne, et al.. Breast cancer and quality of life: medical information extraction from health forums. 25th European Medical Informatics Conference (MIE), Aug 2014, Istanbul, Turkey. pp.1070-1074, 10.3233/978-1-61499-432-9-1070 . hal-01061891

HAL Id: hal-01061891

https://hal.science/hal-01061891v1

Submitted on 8 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Breast Cancer and Quality of Life: Medical Information Extraction from Health Forums

Thomas OPITZ[a,b,1], Jérome AZE[a], Sandra BRINGAY[a], Cyrille JOUTARD[b], Christian LAVERGNE[b] and Caroline MOLLEVI[c]
[a]*LIRMM, Université Montpellier, France*
[b]*I3M, Université Montpellier, France*
[c]*Biostatistics Unit, Institut de Cancérologie de Montpellier, France*

Abstract. Internet health forums are a rich textual resource with content generated through free exchanges among patients and, in certain cases, health professionals. We tackle the problem of retrieving clinically relevant information from such forums, with relevant topics being defined from clinical auto-questionnaires. Texts in forums are largely unstructured and noisy, calling for adapted preprocessing and query methods. We minimize the number of false negatives in queries by using a synonym tool to achieve query expansion of initial topic keywords. To avoid false positives, we propose a new measure based on a statistical comparison of frequent co-occurrences in a large reference corpus (Web) to keep only relevant expansions. Our work is motivated by a study of breast cancer patients' health-related quality of life (QoL). We consider topics defined from a breast-cancer specific QoL-questionnaire. We quantify and structure occurrences in posts of a specialized French forum and outline important future developments.

Keywords. breast cancer, quality of life study, query expansion, social media, text mining

Introduction

Health-related quality of life (QoL) is a multidimensional, subjective and dynamic concept with three main domains: physical, psychological and social functioning. Referring to the patient's perception of his treatment and illness, it constitutes a relevant alternative clinical endpoint. In 2012, the number of new cases of breast cancer in France amounts to 48,763 according to the French National Cancer Institute (INCa[2]). The overall relative survival rate 5 years after diagnosis is estimated at around 89% on average. Modern treatments are still exhausting to undergo such that many clinical research projects are now addressing the patients' QoL. We focus on stories told by patients in online health forums, where numerous topics deal with how to cope with symptoms of the illness and secondary treatment effects. Hancock et al. [1] showed that anonymous communication via computers facilitates the expression of affective states such as emotions, opinions and doubts compared to more traditional communication contexts. Recently, the impact of social media on health outcomes has been studied extensively. Merolli et al. [2] survey the positive effects of social media in

[1] Corresponding author

[2] http://www.e-cancer.fr

chronic disease management. Subirats et al. [3] affirm that social networks promote knowledge democratization and user-empowerment. In this study, we aim to capture topics of interest in patient stories related to breast cancer. Oncologists may have difficulty to obtain such useful patient-centred information directly from patients. It may 1) assist them to investigate how patients use forums in relation to their individual needs, 2) explain consistency or disparity between different patient health outcomes, 3) help collect patients' individual perceptions and preferences, 4) help improve the construction and validation of questionnaires for traditional clinical studies.

Methods

We seek for posts in specialized forums related to a list of topics of interest coming from questionnaires usually filled in by the patients. The studied French *CancerDuSein.org* forum with data from 2011 to 2013 holds 16,961 posts from 675 users in more than 1,050 threads, of which some contain more than 500 replies. Semi-automatic information extraction methods for these data remain a significant technological challenge. The text structure is not standardized (slang, spelling and grammatical errors, etc.) and simple text search for relating topics of interest and posts is not effective. Whereas a query such as *bouche sèche* (=*dry mouth*) returns many false negatives and misses related occurrences such as *langue sèche* (=*dry tongue*), too general queries like *bouche* return many false positives. Query expansion is therefore necessary, achieved in the following by using web resources to produce and validate topic expansions. Our method consists of 4 steps: 1) message collection, 2) manual topic identification and morphological expansion, 3) automatic synonym-based topic expansion, 4) global synthesis.

Step 1: Message collection. Classical pre-processing steps are applied to overcome difficulties owing to the specificity of the language. Named entities like user names or drugs[3] are identified. A part-of-speech tagger[4] is used to retrieve lemmas and to detect unknown words. We replace unknown words if a nearby word is found in a dictionary designed for the project (French words[5], named entities).

Step 2: Topic identification and morphological expansion. We define a list of topics of interest based on the breast-cancer specific module QLQ-BR23 of the QLQ-C30 questionnaire [6] proposed by the European Organisation for Research and Treatment of Cancer (EORTC), which evaluates function scales, symptom scales and the global health status. An expert manually defines the topics using questionnaire items and scales. Initially, we represent a topic through a topic set $T_I=\{TTI\}$ composed of n_I initial topic terms TTI, each composed of one or several lemmas without stop words. Basic morphological variations (nouns, verbs, adjectives) with the same word stem are taken into account and yield n_M additional topic terms TTM in the morphologically expanded topic set $T_{IM}=\{TTI\}\cup\{TTM\}$. Topic sets will be used to formulate the query for detecting topic occurrences in forum posts, see Step 4.

Step 3: Automatic synonymic topic expansion. Morphological variations do not cover all lexical variations found in forums. We extend T_{IM} by adding synonymic topic

---

3 http://medicament.comprendrechoisir.com/

4 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

5 http://www.aspell.net

terms TTS such that $T_{IMS}=T_{IM} \cup \{TTS\}$. Medical terms from the MeSH[6] are eligible for substitution by any related term that appears in the same MeSH hierarchy level. However, the MeSH is limited to technical vocabulary used primarily by health professionals. In addition, we use an online synonym tool[7] to enrich our list of proposed topic expansions. Only synonyms containing no stop words are considered for expansion. Since synonymy depends strongly on the context, we propose a web-based validation to rank synonymic expansions. The originality of our approach is to focus not on the number of occurrences as done by [7,8], but on the similarity of contexts built for the initial topic set T_{IM} and for each potential expansion TTS. The contexts are constructed from the most frequent co-occurrences within a large corpus of documents. For T_{IM}, we use the n_I+n_M topic terms of T_{IM} as queries submitted to an internet search engine[8]. We define a topic context as a weighted feature space based on the top N_{snip} search engine snippets for each query. Features taken into account are unigrams and bigrams of lemmatized nouns, verbs and adjectives. The weight of a feature f is calculated according to its frequency in the $(n_I+n_M) \times N_{snip}$ snippets, written $N(f|T_{IM})$, and the frequency in the French language[9], written N(f|Fr). A formula similar to the common Tf-Idf approach (cf. [9]) is applied to downweight weakly discriminant features:

$$weight(f \mid T_{IM}) = N(f \mid T_{IM}) / (N_{snip} \times (n_I + n_m) \times \log N(f \mid Fr))$$
$$weight(f \mid T_{IM}) = N(f \mid T_{IM}) / (N_{snip} \times (n_I + n_M) \times \log N(f \mid French)) \quad (1)$$

In our experimentations, we double the weight of bigrams, which are strongly discriminant features. Denote by E_i, i=1,…,k the weight of the top k features f_i ordered by decreasing weight after removal of features that are part of the proposed topic expansion TTS (*expected weights*). The context of each topic expansion is constructed analoguously: we submit the query <TTS> to the chosen internet search engine and replace T_{IM} by TTS and (n_I+n_M) by 1 in Eq. (1). Denote by O_i, i=1,…,k, the weight of the features f_i, i=1,...,k, in the expansion context (*observed weights*). To rank a topic expansion TTS with respect to its proximity to T_{IM}, various measures could be applied (cf. [8]). Here, we propose to use a similarity score S based on a variant of the chi-squared statistics

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

The score S takes values between 0 for disjoint contexts and 1 for identical contexts. The truncation at 0 avoids penalizing occurrences that are more frequent in the expansion context than in the topic context. Elevating to the power 2 penalizes the

---

6 French version of the Medical Subject Headings http://mesh.inserm.fr/mesh/

7 http://synonymo.fr

8 www.yahoo.fr

situation where observed weights are considerably smaller than expected weights. We include those topic expansions TTS into the topic set T_{IMS} whose score is sufficiently high, e.g. $S > S_0$ for some fixed threshold $0<S_0<1$.

Step 4: Global synthesis. We detect topic occurrences in lemmatized forum posts based on queries built from the expanded topic set T_{IMS}. Keywords constituting a topic term are combined by the AND operator "&".Topic terms are then joined by the OR operator "|". The query must be matched in a range of at most N_{lem} lemmas in a post. Retrieved occurrences can be represented in terms of the number of occurrences per topic, message IDs, phrases of occurrence, etc.

Results

The QLQ-BR23 questionnaire with 23 items is represented by 13 topics, which are aggregated to form 5 topic groups: *body image*, *sexuality*, *secondary effects* (exempt *hair loss*), *hair loss*, *breast symptoms*. For each topic, we calculate the score S for at most 2500 topic expansions. The number of snippets is fixed to N_{snip}=40 for each search engine query. We compare contexts based on the k=50 most relevant features of the topic context. Our experimentations suggest adding synonymic expansions TTS with score superior to 0.2 to the set T_{IMS}. The final query used to retrieve occurrences in forum posts is composed of 31 initial topic terms in T_I, 70 additional morphological topic expansions in T_{IM} and 596 automatically validated synonymic topic expansions.

When fixing the score threshold S_0 below 0.2 (e.g., to 0.1), we found a considerable increase in the number of false positives such that precision degrades, whereas we also find a significant number of additional true positives leading to an improvement in recall. We decided for a manual selection of relevant topic expansions with score between 0.1 and 0.2 by an expert. Figure 1 presents the number of retrieved posts. We attribute these occurrences as follows: occurrences of initial topic terms given by T_I always count for type "initial", occurrences of morphological variants count for "morphological" but not for "automated", and only the remaining occurrences are attributed to "automated". The manual validation of a stratified subsample of size 400 by two experts results in a precision of 0.89 with a validation agreement of 0.86 (Cohen's kappa). Automatic topic expansion has significantly increased the number of retrieved posts without loss in precision. However, only a small number of posts related to sexuality and body image were detected.

---

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

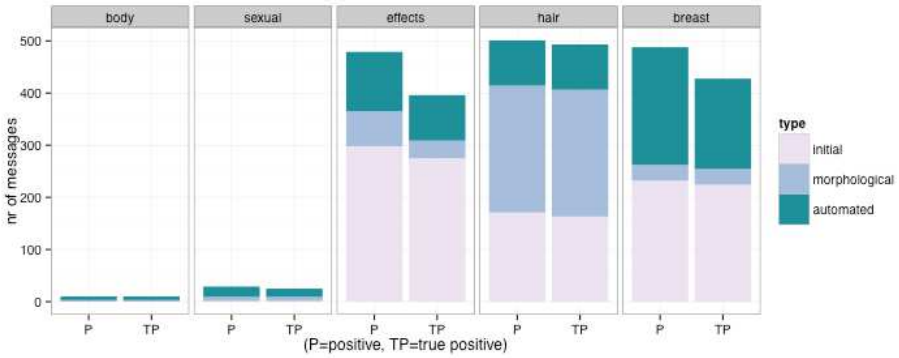

Figure 1. Number of retrieved forum posts by topic group and by type of topic term.

Discussion

A study of QoL through user content posted in health forums presents an interesting alternative to QoL analysis based on the collection of auto-questionnaires. The substantial number of occurrences retrieved for certain topics in our example is promising for a deeper analysis based on a classification of emotions (fear, joy…) associated to the occurrences. The small number of posts related to sexuality and body image is surprising. Lexical variations referring to such subjects may be more heterogeneous and hence more difficult to detect automatically. Moreover, the rare occurrence of these topics is an interesting result in itself since it hints at subjects that play a minor role in the actualities of patients' daily life. Moreover, self-censorship of forum users could be stronger on certain highly intimate subjects. For a proper statistical treatment of our results, we plan to develop classification methods for the retrieval of forum users' patient history from forum data. Finally, to reduce the number of false negatives returned by the topic query, it may be more efficient to construct topics not from existing questionnaires, but in a joint effort of clinical experts and text mining specialists to tailor topics for easier detection. Although forum content is usually publicly available, we stress that results of such approaches should be fully anonymized in order to respect the privacy of individual forum users.

Acknowledgement

The study was supported by a grant from the French Public Health Research Institute (www.iresp.net) within the 2009-2013 Cancer Plan.

References

1. .T. Hancock, C. Toma, and N. Ellison, The truth about lying in online dating profiles, in *Proceedings of the SIGCHI conference on Human factors in computing systems* (2007), 449–452.
2. M. Merolli, K. Gray, F. Martin-Sanchez, 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)** (2013), 957–969.

3. L. Subirats, L. Ceccaroni, R. Lopez-Blazquez, F. Miralles, A. García-Rudolph, J. M. Tormos, Circles of Health: Towards an advanced social network about disabilities of neurological origin, *Journal of Biomedical Informatic*s, **46(6)** (2013), 1006–1029.
4. A. Balahur, Sentiment analysis in social media texts, *4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (2013), 120–128.
5. N.K. Aaronson, S. Ahmedzai, B. Bergman, M. Bullinger, A. Cull, N.J. Duez, A. Filiberti, H. Flechtner, S.B. Fleishman, J.C. de Haes, QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology, *Journal of the National Cancer Institute*, **85** (1993), 365–376.
6. P.D. Turney, Mining the Web for synonyms: PMI-IR versus LSA on TOEFL, in *Proceedings of ECML* (2001), 491–502.
7. M. Roche, O.M. Garbasevschi, WeMiT: Web-Mining for Translation, in *Proceedings of PAIS/ECAI, Conference on Prestigious Applications of Intelligent Systems* (2012), Poster.
8. G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management, **24(5)** (1998), 513–523.