Bandaragoda Tharindu (Orcid ID: 0000-0001-5047-3496)

Text mining for personalised knowledge extraction from online support groups

Tharindu Rukshan Bandaragoda[1], Daswin De Silva[1], Damminda Alahakoon[1], Weranja Ranasinghe[1,2], Damien Bolton[2]

*Email: *T.Bandaragoda@latrobe.edu.au*, T: +61 3 9479 4700*

[1]Research Centre for Data Analytics and Cognition, La Trobe University, Victoria, Australia

[2]Austin Hospital, Heidelberg, Victoria, Australia

The traditional approach to healthcare is being revolutionised by the rapid adoption of patient-centred healthcare models. The successful transformation of patients from passive recipients to active participants is largely attributed to increased access to healthcare information. Online support groups present a platform to seek and exchange information in an inclusive environment. As the volume of text on online support groups continues to grow exponentially, it is imperative to improve the quality of retrieved information in terms of relevance, reliability and usefulness. We present a text mining approach that generates a knowledge extraction layer to address this void in personalised information retrieval from online support groups. The knowledge extraction layer encapsulates an ensemble of text mining techniques with a domain ontology to interpose an investigable and extensible structure on hitherto unstructured text. This structure is not limited to personalised information retrieval for patients as it also imparts aggregates for crowdsourcing analytics by healthcare researchers. The proposed approach was successfully trialled on an active online support group consisting of 800,000 posts by 72,066 participants. Demonstrations for both patient and researcher use-cases accentuate the value of the proposed approach to unlock a broad spectrum of personalised and aggregate knowledge concealed within crowdsourced content.

Introduction

A public health article published in 2004 (Godlee et al., 2004) proposed the World Health Organisation adopt a focal position towards attaining the goal of "Health Information for All" by 2015. This is yet to be achieved, as substantiated by commonplace knowledge as well as research reviews (Finch et al., 2013; Proaño et al., 2016). Among most strategies proposed for achieving universal access, information cycles and communities of practice are given equal prominence as clinical initiatives, national policies and interdisciplinary networks (Godlee et al., 2004). More recently, the importance of a federated health information architecture for addressing diverse information needs and enabling data-based decision-making has been postulated (Kumar et al., 2017). The dispersion of healthcare information across researchers, publishers, reviewers, healthcare professionals, policymakers as well as patients, carers and other health information consumers form the primary information cycle with high potential to increase relevance and reliability of information as well as build skills, understanding and ownership. A strong emphasis is expounded upon sustainability and inclusivity of such information cycles. In direct contrast, the prevalence and popularity of online support groups (OSGs) is a compelling indication that autonomous, self-standing communities are responding to their own information needs by partaking in online conversations to seek, provide and exchange healthcare information. In the absence of institutional support for these communities and their information cycles, it is pertinent to focus on research endeavours that contribute towards an enhanced information retrieval experience for health information consumers.

From a clinical perspective, the rise of patient-centred care (PCC) models postulate the increasingly important role of healthcare information for patient empowerment. PCC is defined as care that is "respectful of and responsive to individual patient preferences, needs, and values, and ensuring that patient values guide all clinical decisions" (Institute of Medicine, 2001). PCC models transcend traditional boundaries that separate patients and their families from clinical contexts. As highlighted by Epstein et al., (2010) and Bechtel & Ness, (2010), PCC aims to achieve a state of shared information, shared deliberation and shared mind. Information sharing should occur across all stakeholders, clinicians, groups of patients, carers and family members.

Most OSGs are structured as topics (or threads) with each topic starting with a question, comment or an experience. Users respond to such threads and thereby create a conversation. Common topics and corresponding conversations are grouped together and indexed as free-text. Subsequent visitors (end-users) begin information perusal with a simple search query on this free-text index. A single search query is inadequate to capture all information expectations of an end-user. A dilemma further complicated by the voluminous results generated by such a query. Simplicity of the search query leads to irrelevant results (Lee et al., 2014). Voluminous results that lack domain-specific structure lead to unreliable and/or ambiguous information. An OSG that retrieves irrelevant and unreliable information, such as contradictory demographics or disease settings, is an inefficacious and harmful resource to end-users. Therefore, it is imperative to improve relevance, reliability and usefulness of information retrieved from OSGs. A wealth of implicit information encapsulated within the OSG's unstructured text data can be extracted to fulfil this purpose. In this paper, we propose an approach that generates this knowledge extraction layer, which then interposes an investigable and extensible structure on hitherto unstructured OSG text.

Contributions:

- Extend and evaluate existing text mining techniques to retrieve age, gender and narrative from OSG text
- Design and develop a layer for extraction of knowledge from OSGs by integrating above techniques with domain ontologies
- Evaluate relevance, reliability and demonstrate usefulness of the knowledge extraction layer using an active OSG containing 800,000 posts by 72,066 participants.

Rest of the paper is organised as follows. The following section discusses related work, followed by a section introducing the proposed text mining approach. The next section reports on an experiment conducted on a current and active OSG with 800,000 posts and a demonstration of information retrieval for health information consumers and OSG analytics for healthcare researchers. The next section evaluates the three knowledge extraction modules developed in this work. Conclusion and a discussion on future work sees to the end of the paper.

Related work

Tanis, (2008) presented six general motives for information seekers on OSGs; to seek information, seek emotional support, inclusion, support others, pass time and convenience. The study indicates active participation is a key predictor of better coping with the medical condition. Improved relevance and reliability of retrieved information further supports the role of a OSG in managing a medical condition. Oh, (2012) examined factors that motivate information providers to share knowledge and information on OSGs. Altruism, enjoyment and efficacy were key motivators for those responding to information seekers. Improved relevance and reliability has been noted in this study as further encouragement for continuous participation of information providers. A further study examined the sustained use of OSGs for the attainment of long-term health benefits (Zhang, 2015). Emotional needs such as individuality, small group interactions, respectful treatment and decision-making awareness aided sustained use of the OSG. As proposed in this paper, personalisation of information retrieval leads to individuality and thereby continuous use of OSGs for the achievement of such long-term goals.

Gupta et al., (2014) presented a method for the identification of medical entity types in OSGs. They demonstrate performance improvements in the extraction of drug and treatment types and symptoms and condition information from two OSGs based on the expansion of medical dictionaries. Lexico-syntactic patterns are used to extract new entities not found in seed dictionaries, abbreviated terms and spelling errors. This method was tested on two OSGs: Asthma OSG and ENT OSG, consisting 39,137 and 215,123 sentences, respectively. In evaluation, it outperformed MetaMap, OBA and a CRF classifier (Gupta et al., 2014).

Tang et al., (2013) developed a hybrid clinical temporal information extraction system to derive temporal expressions and temporal relations from hospital discharge summaries. The three components of the system, event extraction, temporal expression extraction, and temporal relation extraction used both rule-based and machine learning (classification) approaches for extraction. Classification was based on a dataset containing 310 clinic notes; 190 used as the training set and remainder used for testing.

Cho et al., (2013) proposed the use of OSG text for Comparative Effectiveness Research (CER). CER is defined as the generation and synthesis of evidence that compares the benefits and harms of different prevention and treatment methods (Cho et al., 2013). They proposed demographic information extraction from OSG text as an alternative to the traditional approaches of observational studies and randomized clinical trials. The authors report the proposed approach is capable of extracting more than 30% of demographic (age and gender) information. Experiments were conducted on the topics of breast cancer (40,996 posts) and heart disease (98,644 posts) in MedHelp OSG.

Ku et al., (2014) proposed a text mining framework to determine self-disclosing health information with unusual messages on OSGs. Their motivation was to inform public health agencies to re-allocate resources and deliver services through social media. The framework comprised three components: data acquisition, feature generation and text representation; and classification. Two major HIV/AIDS OSGs (5,500 posts) in Taiwan were used to demonstrate its feasibility.

In summary, information seekers, information providers and long-term end-users alike advocate relevance and reliability of retrieved information for effective and sustained OSG use. Moreover, scale is an essential element in crowdsourcing. As noted above, hitherto research endeavours on information retrieval from OSGs have only focused on small-scale OSGs. In this context, it is important to note that our approach has been trialled on a voluminous OSG consisting 800,000 posts with demonstrable accuracy and value.

The proposed method

It is appropriate to start this section with an illustrative comparison (Figure 1) of the existing method for information retrieval from OSGs and the proposed knowledge extraction based structure.

The existing method requires end-users to perform a full text search within the posts of the OSG or browse through a selected topic. OSGs that provide full-text search, index terms found in each post (except stop-words) so that they are searchable. Full-text search is demonstrated below using an example query *"I'm a 40 year old woman taking Nexium for heartburn"*.

1. Tokenize the query into words *(I, m, a, 40, year, old, woman, taking, Nexium, for, heartburn)*.
2. Remove stop-words and stem-words to their root form *(40, year, old, woman, take, Nexium, heartburn)*.
3. Search the OSG database for the posts that have any of these words.
4. Determine a relevance score for each identified post by aggregating tf-idf scores of each matching word (tf-idf is a statistical measure of how important a word is to a document in a collection of documents).
5. Present the top *n* results with the highest relevance score.

The top *n* results consist of posts that contain several matching words in the end-user query. However, it does not recognise that *heartburn* is a symptom, *Nexium* is a medication and the end-user is interested in posts that mention both of these words. In addition, it does not recognise *40* is an age mention and *woman* is a gender mention. This highlights the two key issues of the existing full-text search.

- **Low relevance**: demographics of the end-user (age and gender mentions in the query) are disregarded. Relevant posts by authors of same gender and age group are overlooked.
- **Low reliability**: the importance of certain terms (such as symptom or medication mentions) are disregarded. These terms introduce context to a search and thereby crucial for reliability.

The existing method based on full-text search leads to low relevance and low reliability mainly because it's limited to matching terms in the query to the indexed post without consideration for the context introduced by these terms.

As illustrated in Figure 1(b), the proposed method overcomes these issues by introducing a knowledge extraction layer that structures OSG posts and queries based on extracted knowledge. The same query is handled by the proposed method as follows:

1. Structure the query by extracting context based on text mining techniques (*{symptom- heartburn, medication- Nexium, age- 40, gender- female})*
2. Match extracted context to the structured knowledge in OSG posts using different criteria (e.g., symptoms, medication, gender, and age group)
3. Distinguish between posts narrated as advice and experience from retrieved results

This approach returns relevant and reliable results compared to the full-text search.

- **High relevance:** normalised information (e.g., woman -> female and 40 -> 35-45 age group) increases the retrieval of posts relevant to the end-user.
- **High reliability:** strict matching criteria is followed on contextual information to filter out unreliable results (e.g. must match symptoms and medications)

Figure 2 presents the expanded illustration of the proposed method which shows the proposed knowledge extraction layer in more detail.

The proposed knowledge extraction layer is extensible, thus can be enhanced with additional knowledge extraction modules (e.g., identify ethnicity, family history, etc.). However, for this work we have limited this layer to extract age, gender, narrative type and medical information mentioned in OSG posts. Note that narrative type aims to distinguish between posts containing experience and advice.

The rest of this section describes the knowledge extraction modules: narrative type, age, gender, and medical concepts (symptoms/medications).

Determining the narrative type

As OSGs are used to seek, provide and exchange information (Zhang, 2015), individuals tend to share similar experiences and also provide advice based on their own experience. In past work, Liu et al., (2015) and Park et al., (2010) extract sentences that express patient experience using a classifier trained on different linguistic features. However, such methods are still in its early stages of development as experience extraction is a relatively new area of research.

In this work, we assume that each post can be broadly categorised as expressing experience or providing advice. We further subdivide expressions of experience as expression of own experience or expression of a third party experience (often family member, close relative or a friend). Thereby, posts are grouped into three categories (i) experience: first-person, (ii) experience: second-person, and (iii) advice

For this categorisation, we look at nouns and pronouns that mention humans to decide the narrative type. The rationale for this approach is that different types of nouns and pronouns are employed in different narrative types. For example, the first column of Table 1 shows three sample posts taken from a OSG. First post is a patient expressing his experience, therefore majority of the human mentions are first-person pronouns. In the second post a child is expressing about his mother, so there is human mention noun 'mother' and third person pronoun 'she'. The last post is a piece of advice given by the author to someone else, so there are significant number of second-person pronoun 'you' present in that post.

However, posts sometimes contain several nouns and pronouns, for example, patients expressing his experience might mention their family history using words such as 'mother' and third-person pronouns like 'her'. Moreover, third-person pronouns are often used to mentions other people as well (e.g., I went to see GP and he prescribed).

In order to resolve these issues, we first employed pronoun resolution tools to identify their antecedents. We first attempted JavaRAP tool (Qiu, Kan, & Chua, 2004) which implements algorithm presented by Lappin & Leass, (1994). However, this approach relies on parsing text to identify its structural elements and grammatical roles. We found this approach is less effective for OSG posts as they rarely conform to rules of formal grammar. Therefore, we selected CogNIAC (Baldwin, 1997), a rule based pronoun resolution algorithm that uses a simple set of rules to resolve pronouns. For our task we did a partial implementation of CogNIAC, to resolve male and female pronouns.

Column 2 of Table 1 presents posts with resolved pronouns. It should be noted that this process has made the primary human mentions more evident by resolving the pronouns to its relevant noun. Column 3 of Table 1 shows the prominent human mention noun of sample posts, which are used to determine the narrative type (column 4).

Age extraction

Age is an important feature for patient profiling because medical information such as symptoms are interpreted differently for different age groups. Patients often mention their age within the text of the OSG post. However, such mentions are highly diverse and often expressed using shorten forms of English words. Some sample age mentions are shown in column 1 of Table 2. Also, they need to be differentiated from other number mentions such as duration, dose, weight etc.

Previous literature on age extraction from OSGs are limited mainly due to its challenging nature. Yang et al., (2014) focused on standard age mentions such as ‘I am 35 years old’ and identify them as age mentions. Kim et al., (2013) employed regular expressions such as ‘age’+number to extract age. These approaches achieve higher precision but as shown in Table 2, age mentions are diverse thus these approaches would overlook significant amount of age mentions resulting in a low recall.

Zhu et at., (2012) look for clue words such as ‘years’, ‘old’, ‘aged’ etc. appear within two-word distance of each two-digit number mentions and such mentions are extracted as the age of the patient. In addition, they look for clue words such as ‘teenager’, ‘toddler’, ‘child’ to approximate the age of the patient. This approach would result in high recall however, can lead to significant number of misclassifications (e.g., ‘about 10 years ago’ can be misclassified as an age mention).

A common issue of these approaches is that they do not resolve whether the age mentions is about the patient. For example, patient might mention age of relatives when talking about their family history. Also, there can be age mentions in past incidents of the patient (e.g., ‘I have this issue since 11 years old’). Moreover, there can be multiple age mentions in a OSG post. Thereby, it is necessary to develop an extraction method to resolve age from multiple age mentions.

Machine learning has been proposed for text classification tasks in clinical and medical literature (Ford et al., 2016). In order to overcome above mentioned issues we developed a text classifier that identifies relevant age mentions from OSG posts. This technique is illustrated in Figure 3 and elaborated below.

Potential Pattern Identification

In the first step we break the text into sentences using a *sentence splitter* and then employed *regular expressions* to capture text chunks with one or two digit numbers in the middle. Five words before the number and five words after the number are included in this extracted text chunk. Note that, numbers co-exist with several letters are also considered to accommodate age mentions such as ‘20s’ and ‘34yrs’

For each post p , a set of potential text chunks $\{t\}^p$ were identified

Feature Extraction

In this step feature vectors (f) are extracted from the text chunks identified from the previous step $t_1^p \rightarrow f_1^p$. As shown in Figure 4, we divided the extracted chunks into three segments (L, M, and R) and different features were extracted separately from each segment.

We engineered 29 features that are able to differentiate age related text chunks and others collected from step one. Table 3 presents a sample of the extracted features.

Most of these features are constructed to capture age mentions while others such as dose and time mentions are to differentiate other key classes such as dosage (e.g., *atenolol 50 mg*) and duration (e.g., *10 years ago*).

Classifier

Once the features were extracted from text chunks, we employed a classifier to identify the chunks that are most likely to be age mentions. To train this classifier we labelled 2,212 text chunks identified from the first step of this process. We labelled them ‘age’ and ‘other’ based on whether it is an actual age mentions or not. In the labelled dataset there are 1,186 ‘age’ and 1,026 ‘other’ labels.

We employed WEKA (Hall et al., 2009) data mining toolkit to try two classifiers Naïve Bayes and Random Forest to select what works best for this task. With 5-fold cross validation Naïve Bayes and Random Forest result f-measures 0.80 and 0.85 respectively. Therefore, Random Forest was selected as the classifier for this task.

Classifier determined a confidence value (c_t) and based on that feature vectors that have a higher confidence of being an age mention are identified. For this task we employed the confidence threshold $\tau = 0.7$.

$$\{t_1, t_2, \ldots, t_n \square c_t > \tau\}$$

Aggregate and resolve age

This step aims to resolve the age of each individual author. An author A often have multiple posts $\{p\}^A$ in different discussion threads. Hence, for this task we considered all high confidence age mentions captured from those posts $\{p\}^A \rightarrow \{t\}^A$ to resolve the authors age.

First, the numerical value of each age mention (a_t) is determined. Note that age mentions are parsed to get the age value normalised to year (e.g., 6 months → 0.5 year). Also, two other parameters are derived from each age mention as follows: (i) tense of age mention (present, other), and (ii) subject type of the age mention (first-person, other).

Tense of the age mention is derived based on tense of the verb in the chunk (if present). Subject type of the age mention is determined based the subject in the chunk (if present, otherwise it is unknown). The rationale of these two parameters is that if the chunk is in present tense then it is more likely to be the authors current age. Also, if the subject if first-person then it is more likely to be about the author. Therefore 0.25 boost is added if the age mention is present tense and 0.25 boost is added if it has a first-person subject.

The aggregated confidence value $C(a)$ for all the age values mentioned in $\{t\}^A$ is determined as follows:

$C(a) = \sum^{\{t\}^A} I(a_t = a)(c_t + r)$, where $r \in \{0, 0.25, 0.5\}$ and I is the indicator function.

r takes value 0.5 when t is both present tense and first person, it takes value 0.25 when t is either present tense or first-person, and it takes value 0 if t is neither present tense nor first-person.

Age value with the highest aggregated confidence is assigned as the age of the author.

$$a_{resolved} = \arg\max_a(C(a))$$

Gender extraction

Gender is another important demographic information that is important for personalised retrieval. Some symptoms can relate to different diagnoses depending on whether the patient is a male or female. Gender is sometimes explicitly mentioned (e.g., I'm a female), and in some cases gender can be inferred based on gender mentions (e.g., my mother) or gender specific medical term mentions (e.g., pregnant). However, similar to age extraction, this task is challenging due to unstructured and diverse nature of gender mentions in OSG posts.

(Cheng et al., 2011) focused on language style of the text to predict the gender of author. The idea is that males and females often follow different language styles for written communication. However, the accuracy of this approach will be low as OSGs attract information seekers and providers with very diverse language styles from across the world. Another approach is to predict the gender by looking at the first name of the author (Herdağdelen & Baroni, 2011). It keeps two lists of male and female first names and resolve the gender based on that. This approach does not work in our scenario as many people do not use their real name in OSGs, they often use a nickname or a part of their name.

Zhu et al., (2012), look for gender clues: (i) gender specifying words such as 'men', 'woman', etc. and (ii) gender specific medical terms such as 'hot flashes', 'prostate cancer'. It is not mentioned that how they resolve the gender, if terms related to both genders appear in posts of the patient. Also, another key issue is to resolve the whether the gender clues are about that patient.

We extended the above approach to resolve gender more accurately for each author. First we filtered out the posts which are marked as *advice* by our narrative type identification module. This step is taken mainly because gender clues in *advice* can be misleading as they often relate to other patients.

Selected posts were first mined for gender specific medical terms. We collected a list of gender specific medical terms by looking at men's and women's sections of OSGs. Gender extraction from gender specifying terms in text is relatively less straightforward than using medical terms. This is because of certain ambiguities that can exists in a post. For example, 'he' may be referring to the doctor not the patient. Therefore, we first need to resolve the pronouns to identify the gender specifying terms that are referring to the patient. For this task we employed the pronoun resolved sentences generated to identify the narrative type. We look for different gender specifying terms based on the narrative type.

If the narrative type is *experience: first-person* then the post is written by the patient (using first-person pronouns). Hence, we first look for direct gender mentions. Examples for such mentions are: ‘a male’, ‘a mother’, ‘a widower’. Those mentions have to exist with in close proximity of first-person pronoun i.e., ‘I’ in a sentence to verify that it is about the patient. If direct gender mentions cannot be found, then we look for indirect gender mentions. Examples for such mentions are: ‘my husband’, ‘my fiancée’. We take the opposite gender of such mentions (e.g., ‘my husband’ infers that patient is a female).

If the narrative type is *experience: second-person*, then a narrator is relating the experience of someone they knew (often about a family member or close relative). In such cases, the patient’s relationship to the narrator can be used to resolve the gender most of the time (e.g., male: husband, son, uncle etc., and female: daughter, wife, mother, etc.). If the post is about ‘husband’ then the patient is a male. Note that this is quite the opposite of the previous case where the post is a *first-person experience*. If the patient’s relationship to the narrator is gender-neutral (e.g., partner, friend) then the pronoun resolution process is examined to check if gender specific pronouns were resolved to that noun (he -> friend) and gender is assigned accordingly. Table 4 shows three examples for this gender resolution process.

Similar to the age resolution process, gender is resolved for each profile by aggregating the gender resolutions in the posts. For this aggregation task we only considered the posts with the type *experience: first-person* as they are about the author. In the aggregation process each gender specific medical term adds a weight of 1.0 to the relevant gender. Gender resolved using the gender specific terms gets a weight of 2.0 as it is more accurate. Once aggregated, gender of the profile is resolved based on the highest weighted gender. Profile gender is then re-assigned to all the posts of that author with the type *experience: first-person*. Posts with the type *experience: second-person* were resolved separately for each post as they are about experience of someone else.

Medical concept extraction

Medical concept extraction from text is a further formidable task. Natural language processing (NLP) tools are used to extract terms from text that can be mapped to the medical concepts found in medical thesauruses such as the UMLS Metathesaurus (Bodenreider, 2004). There are several state-of-the-art tools that extracts medical concepts from text such as: MedLEE (Friedman et al., 1994), cTAKES (Savova et al., 2010), and MetaMap (Aronson et al., 2010).

Gupta et al., (2014), shows that better precision and recall can be achieved by developing a tool to cater special characteristics of patient-authored text. Consumer health vocabularies extracted from community generated corpora (Vydiswaran et al., 2014) are employed to identify the relevant medical terms and a different stack to NLP processes is followed to extract key phrases from text. However, as noted by the authors, the existing system is limited to certain subcategories of the OSG (Asthma and ENT). Therefore, for our task we adhere to the well-established medical concept extraction tools.

We employed cTAKES (Savova et al., 2010) tool for our medical concept extraction task. It identifies noun phrases in the text and then conduct a dictionary-look up in SNOMED CT[1] and RxNORM (Nelson et al., 2011) medical concept databases. Each identified term is then mapped into five sematic types: disorder/disease, sign/symptoms, procedures, anatomical sites, and medications. In the proposed method, we subject each OSG

[1] http://www.ihtsdo.org/snomed-ct

post to this process and the identified medical concept are extracted. As pointed-out by Gupta et al., (2014), some ambiguities exists in the identified concepts. For example, ‘today’ is mapped to a drug named Today and ‘web’ is mapped to the disorder ‘congenital webbing‘. However, both these words are found frequently in OSG posts mostly referring to their usual meanings. To overcome this issue, we constructed a list of terms that often mapped incorrectly and filtered out the identified concepts based on such terms.

Demonstration

This section demonstrates underlying workings of the proposed method as well as its relevance and reliability in addressing information needs of patients (end-users) and researchers. Patients aim to find cases that are more relevant to them both medically and demographically. Researchers aim to extract aggregated insights on medical conditions based on different dimensions such as age, gender and time. We employ two use cases, one each for a patient and a researcher, to demonstrate how their information needs can be addressed using knowledge extraction layer generated by the proposed method.

The rest of this section is organised as follows. First we introduce the test dataset collected from an active OSG and then present the implementation of the proposed knowledge extraction layer. Next we delineate implementation of personalised search with a patient use case. Finally, we demonstrate OSG analytics capabilities using a researcher use case.

OSG dataset collection and implementation

We collected approximately 800,000 posts by 72,066 authors from the reputed OSG patients.info[2]. Each post belongs to a particular OSG thread, categorised based on a set of high level topics (e.g., ‘Brain and nerves’ and ‘Women’s health’) which is the only available structure for posts in the OSG.

Narrative type of each post is identified individually for each post. However, age and gender is resolved for each author. We used the associated *author-id* of each post to aggregate all posts of a particular author. These aggregated posts are then utilised to resolve age and gender of each author. As explained in the previous section, resolved gender and age is assigned to the relevant posts of that author considering the identified narrative type of each post.

Age, gender and narrative type extraction techniques are linear in time complexity ($O(n)$) and thus efficient and scalable to handle large number of posts. In addition, each patient profile can be independently processed by individual threads in a multi-threading environment. Figure 5 shows experimental execution time results obtained against number of posts. The experiments are carried out in a 16 core 1.4Mhz server. It shows that the time complexity is linear with number of posts. A total of 1,000,000 posts (an average of 153 words per post) are extracted in 6060 seconds in a single threaded environment, and only 3000 seconds with five parallel threads. Note that the slight increase of execution time with ten parallel threads is due to I/O resource bottlenecks.

As discussed in a previous section, the medical concept extraction is carried-out using cTAKES (Savova et al., 2010) which runs on the highly scalable Apache UIMA framework (Ferrucciet al., 2009).

[2] http://patient.info/forums

Each post is structured with the extracted knowledge (narrative, demographic and medical) and published date. We employed open source search platform Elasticsearch[3] to store this dataset since it is handles both full-text and structured search. Elasticsearch is a distributable full-text search engine, designed to be scalable to handle very large datasets (Kononenko et al., 2014). Indexing and search performance metrics for Elasticsearch benchmarks[4] are available online.

Table 5 provides a distribution of the extracted information from the above dataset. The proposed demographic extraction modules managed to resolve age and gender of 47% of the posts in the collected dataset compared to the reported 30% success by Cho et al., (2013). Approximately 80% of the posts discuss patient experience and among them majority (78%) are first-person experience.

Among the gender resolved posts, 83% is authored by females where only 17% is from males. We have further investigated this in research literature. (Kummervold et al., 2017) report that women are more participative in OSGs than men. Li, Lin, & Wang, (2015) state that women tend to self-disclose more information than men. Therefore, we assume there is less male participation in OSGs and also even the participating males tend to expose less information to identify their gender.

Personalised search to extract relevant information

As discussed in previous sections, the proposed method provides personalised (relevant and reliable) information in response to patient search queries based on their medical and demographic information.

We first employed the medical information extracted from the patients query to identify the posts that are experiences and contain same medical information. Such posts are then ranked based on the demographic similarity of the author to the demographics of the patient. For this ranking we use a custom relevance measure based on age and gender.

Let age and gender of a patient P be P_a and P_g respectively. Using the same notation, let age and gender of an experience E be E_a and E_g. The relevance measure r_E is defined as follows:

$$r_E = \quad W_a \times \exp(-|P_a - E_a|/\sigma_a^2) \quad + \quad W_g \times I(P_g = E_g)$$

Where W_a and W_g are the weights for age and gender respectively. I is the indicator function which is 1 if P_g equals E_g and 0 otherwise. Weight of age is associated with a Gaussian decay function which is 1 if P_a equals P_g. σ_a is used to control the granularity of age matching where smaller values make the decay function to decrease rapidly with the age difference and vice versa. W_g and W_a are set to 0 if the patient does not provide their demographic details. The OSG posts are ranked based on this relevance measure and presented to the patient.

[3] https://www.elastic.co/products/elasticsearch
[4] https://elasticsearch-benchmarks.elastic.co/

Patient use case

In this patient use case we demonstrate how personalised retrieval can provide relevant and reliable information to the patient compared to the existing full-text search. We use the same query: "*I'm a 40 year old woman taking Nexium for heartburn*". Contextual information is initially extracted from the query and used to identify medical information (symptom: heartburn, medication: Nexium) and demographic information (age: 40, gender: female). Most relevant experiences are retrieved from the database using the abovementioned method.

For comparison of results we employed two approaches of full-text search: (i) default search in the OSG, and (ii) search key terms with the Boolean aggregation 'AND' (retrieves the posts that contain all the search terms).

Excerpts from the top five results from our method and the two approaches of the full-text search are provided in Table 6. The key terms that are relevant to the query is highlighted in each excerpt. Note that, some experiences retrieved by the proposed method does not have age or gender mentioned in that post. This is because age and gender is resolved for each author using all posts by that author, so age or gender of that author is inferred from other posts and not the retrieved post.

Above results show that the proposed method retrieves more relevant posts for the given query. Instead of taking key terms of the query as-is, the proposed method identifies the patient is a female and her age is 40. Therefore, it retrieves similar experiences from females who are aged close to 40. Also, the posts do not necessarily need to have age and gender mentions in the post itself as they were resolved for each author.

In comparison, the second query (full-text search in the OSG) attempts a direct string matching with the search terms and retrieves partially matched results that contain any (unknown) combination of search terms. It is apparent that the last match of this query is irrelevant, because it is only matching 'woman' and '40' but does not have the symptom 'heartburn' or the medication 'Nexium'.

The third query has a very low recall with only one retrieved post, as it is rare to have all four terms in a single post. Also, it is clearly noticeable that the match term '40' is not an age mention.

OSG analytics for researchers

Medical research is often conducted using a small samples of patients due to the associated cost (both time and money) of such research. On the other hand, such information is accumulated in OSGs, crowd sourced by real patients. These untapped resources are inaccessible to researchers due to inherent noise, unstructured nature and diversity of information representation. Researchers have to attempt the formidable task of executing full-text queries and manually extract information from the resulting posts.

As previously illustrated in Figure 2, the proposed method builds a structured layer on top of the unstructured text of OSG posts which can be utilised for OSG analytics. As shown in Figure 6, each post can be represented using five dimensions that enables researchers to conduct OSG analytics and gain insights. It provides unprecedented access to OSG data from different viewpoints.

Researcher use case

In order to demonstrate the OSG analytics capability, we performed several analyses on patients who report the symptom *heartburn*. Note that this attempt is solely to showcase the analytical capabilities and not a comprehensive medical research on *heartburn*.

Dimensional analysis: In this analysis we combine age and gender dimensions shown in Figure 6 and present demographic distribution of patients who report the symptom *heartburn*. Figure 7 shows the demographic distribution of posts that mention the symptom *heartburn*. This type of analysis is useful to identify the age groups that are more affected by a particular symptom and also observe potential demographic biases.

Association mining: The OSG analytics layer is also useful for association mining. It can be used to analyse associations between different symptoms in order to identify co-existing symptoms. Table 7 presents the top five other symptoms that co-exists with the symptom *heartburn* in different age-groups.

Temporal analysis: The Date dimension can be used to perform temporal analysis to identify seasonal patterns in the OSG. Figure 8 shows the temporal distribution of the post counts that report the symptom *heartburn* drawn for each month for a period of three years. It shows that over the three-year period reported *heartburns* are relatively high during March and April.

These examples highlight the importance of the proposed method for medical research, which can be effectively used to conduct large scale analytics investigations on OSG participants. Furthermore, proposed methods have been advanced into an oncology care framework for the analysis of patient reported outcomes and emotions (Ranasinghe, et al., 2017; Bandaragoda et al., 2018).

Evaluation

This section evaluates the three knowledge extraction modules: (i) narrative type classification, (ii) age resolution, and (iii) gender resolution. We obtained the services of qualified domain experts for manual classification of test datasets. Narrative type classification is evaluated using a labelled set of posts as advice or experience. Age and gender resolution is evaluated using a labelled set of OSG post authors using their published posts.

Narrative type classification performance

Narrative type classification is evaluated using 500 posts labelled by domain experts as experience or advice. Note that we ignored the sub-classification of experience (experience: first person and experience: second person) for this evaluation, because *experience: second person* is relatively rare in the dataset. It is evaluated as a classification problem where the two classes are *Experience* and *Advice*. Table 8 presents the evaluation results.

The results show that both *Experience* and *Advice* are identified with a precision above 0.9. Recall of *Advice* is relatively low mainly because some advising posts are mixed with the authors experience and therefore hard to identify them as advice.

Age and gender resolution performance

Age and gender resolution was evaluated using a set of 300 labelled author profiles. Posts of each author were examined to identify age or gender of the author if such information is present. Each author profile is annotated based on the identified age and gender.

The labelled data is then compared to the output of age and gender resolution modules. We employed precision and recall statistics for this evaluation. Note that age is often mentioned in incremental values for some authors as a result of prolonged contribution to the OSG over several years. Therefore, age resolution is considered correct if it falls within two integer values of the labelled age.

Similar to the previous evaluation, performances of the gender and resolution modules are evaluated as classification problems. For gender, the classes are *Female*, *Male*, and *Unknown* and for age, classes are *Age mentioned* and *Age not mentioned*. Note that, in *Age mentioned* class the classifier has to correctly resolve the age value (within two integer values to the labelled age value) in order to be a *true positive*.

Table 9 and Table 10 present the gender and age classification results respectively.

Both age and gender resolution have average precision and recall over 0.85. Recall for *Male* is relatively low. Most of those misses are classified as *Unknown* as the classifier misses the gender specific clues. This is mainly because males tend to expose very few clues about their gender compared to females. Precision in *Age mentioned* class is relatively low because when a profile does not have actual age mentions, the classifier tends to pick up low confidence age mentions that are often age mentions in past incidents or age mentions about other people. Same issue results a relatively low recall in *Age not mentioned* class as well.

Conclusion

As widely emphasised in medical literature, sustainable and inclusive information cycles that extend across researchers, healthcare professionals, policy makers and patients are a key element in the transition towards patient-centred care. OSGs play an informal yet active role in addressing this void. In this paper, we present a novel method that imposes a relevant and reliable knowledge extraction layer on to large volumes of unstructured text in OSGs used by patients for information exchange and emotional support. This knowledge extraction layer is developed by extending current text mining techniques to retrieve age, gender and narrative type from text; and domain ontologies to retrieve symptom and medication mentions. Age, gender and narrative type extraction modules show high precision and recall when evaluated with labelled datasets.

As future work, the knowledge extraction layer can be extended with further knowledge extraction modules to extract more information from unstructured text such as demographics (ethnicity, weight, height etc.) and personal health information (family history, allergies, surgeries etc.). Moreover, other crowd sourced information sources such as Facebook and Twitter can be integrated to increase the coverage on public patient health information.

The proposed method was tested on a collection of 800,000 OSG posts, which demonstrates its scalability. The capacity to handle information needs of patients with high relevance and reliability as well as the capacity to provide aggregated OSG analytics to researchers were also demonstrated. In summary, this novel approach for personalised knowledge extraction from crowdsourced OSG data elucidates a broad spectrum of individual and aggregate knowledge that makes a noteworthy contribution towards patient empowerment, patient-centred care and 'health information for all'.

References

Aronson, A. R., Lang, F.-M., Aronson, A., Aronson, A., Rindflesch, T., Browne, A., … Divita, G. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, *17*(3), 229–36. [CrossRef][10.1136/jamia.2009.002733][Mismatch]

Baldwin, B. (1997). CogNIAC: High precision coreference with limited knowledge and linguistic resources. *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Text, Madrid, Spain, July 1997*, 38–45.

Bechtel, C., & Ness, D. L. (2010, May). If you build it, will they come? Designing truly patient-centered health care. *Health Affairs*.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research, 32*(90001), 267D–270. [CrossRef][10.1093/nar/gkh061]

Cheng, N., Chandramouli, R., & Subbalakshmi, K. P. (2011). Author gender identification from text. *Digital Investigation, 8*(1), 78–88. [CrossRef][10.1016/j.diin.2011.04.002]

Cho, J. H. D., Liao, V. Q. Z., Jiang, Y., & Schatz, B. R. (2007). Aggregating Personal Health Messages for Scalable Comparative Effectiveness Research. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'13* (pp. 907–916). New York, New York, USA: ACM Press.

Cho, J. H. D., Liao, V. Q. Z., Jiang, Y., & Schatz, B. R. (2013). Aggregating Personal Health Messages for Scalable Comparative Effectiveness Research. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, 907–916.

Epstein, R. M., Fiscella, K., Lesser, C. S., & Stange, K. C. (2010, August). Analysis & commentary: Why the nation needs a policy push on patient-centered health care. *Health Affairs*.

Ferrucci, D., Lally, A., Verspoor, K., & Nyberg, E. (2009). Unstructured Information Management Architecture (UIMA) Version 1.0. OASIS Standard.

Finch, D. J., Bell, S., Bellingan, L., Campbell, R., Donnelly, P., Gardner, R., … Jubb, M. (2013, June). Accessibility, sustainability, excellence: How to expand access to research publications. Executive summary. *International Microbiology*.

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: A systematic review. *Journal of the American Medical Informatics Association, 23*(5), 1007–1015. [CrossRef][10.1093/jamia/ocv180]

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association : JAMIA*, *1*(2), 161–74. [CrossRef][10.1136/jamia.1994.95236146][Mismatch]

Godlee, F., Pakenham-Walsh, N., Ncayiyana, P. D., Cohen, B., & Packer, A. (2004). Can we achieve health information for all by 2015? *Lancet*.

Gupta, S., Maclean, D. L., Heer, J., & Manning, C. D. (2014). Induced lexico-syntactic patterns improve information extraction from online medical forums. *Journal of the American Medical Informatics Association*, *21*(5), 902–909. [CrossRef][10.1136/amiajnl-2014-002669]

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10. [CrossRef][10.1145/1656274.1656278]

Herdağdelen, A., & Baroni, M. (2011). Stereotypical gender actions can be extracted from web text. *Journal of the American Society for Information Science and Technology*, *62*(9), 1741–1749. [CrossRef][10.1002/asi.21579]

Institute of Medicine. (2001). Crossing the Quality Chasm:a New Health System for the 21st Century. *Institute of Medicine*, (March), 1–8.

Kim, M. Y., Xu, Y., Zaiane, O., & Goebel, R. (2013). Patient information extraction in noisy tele-health texts. In *Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013* (pp. 326–329). IEEE.

Kononenko, O., Baysal, O., Holmes, R., & Godfrey, M. W. (2014). Mining modern repositories with elasticsearch. In *Proceedings of the 11th Working Conference on Mining Software Repositories - MSR 2014* (pp. 328–331).

Ku, Y., Chiu, C., Zhang, Y., Chen, H., & Su, H. (2014). Text mining self-disclosing health information for public health service. *Journal of the Association for Information Science and Technology*, *65*(5), 928–947. [CrossRef][10.1002/asi.23025]

Kumar, M., Mostafa, J., & Ramaswamy, R. (2017). Federated health information architecture: Enabling healthcare providers and policymakers to use data for decision-making. *Health Information Management Journal*.

Kummervold, P. E., Gammon, D., Bergvik, S., Johnsen, J.-A. K., Hasvold, T., Rosenvinge, J. H., & Kummervold, E. (2017). Social support in a wired world Use of online mental health forums in Norway.

Lappin, S., & Leass, H. J. (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, *20*(4), 535–561.

Lee, K., Hoti, K., Hughes, J. D., & Emmerton, L. (2014). Dr google and the consumer: A qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. *Journal of Medical Internet Research, 16*(12), 1–13. [CrossRef][10.2196/jmir.3706][Mismatch]

Li, K., Lin, Z., & Wang, X. (2015). An empirical analysis of users' privacy disclosure behaviors on social network sites. *Information and Management*, *52*(7), 882–891.

Liu, Y., Chen, Y., Tang, J., & Liu, H. (2015). Context-Aware Experience Extraction from Online Health Forums. *2015 International Conference on Healthcare Informatics*, 42–47.

Liu, Y., Xu, S., Yoon, H.-J., & Tourassi, G. (2014). Extracting patient demographics and personal medical information from online health forums. *AMIA Annual Symposium Proceedings, AMIA Symposium.*, *2014*, 1825–1834.

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., & Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association : JAMIA*, *18*(4), 441–8. [CrossRef][10.1136/amiajnl-2011-000116][Mismatch]

Oh, S. (2012). The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments. *Journal of the American Society for Information Science and Technology, 63*(3), 543–557. [CrossRef][10.1002/asi.21676]

Park, K. C., Jeong, Y., & Myaeng, S. H. (2010). Detecting experiences from weblogs. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Proaño, A., Ruiz, E. F., Porudominsky, R., & Tapia, J. C. (2016). The dream of health information for all. *F1000Research, 40*, 1–6. [CrossRef][10.12688/f1000research.6950.2][Mismatch]

Qiu, L., Kan, M.-Y., & Chua, T.-S. (2004). A Public Reference Implementation of the RAP Anaphora Resolution Algorithm. *Proceedings of the 4th Language Resources and Evaluation Conference (LREC 2004), 2*(Lrec), 291–294.

Ranasinghe, W., Bandaragoda, T., De Silva, D., & Alahakoon, D. (2017). A novel framework for automated, intelligent extraction and analysis of online support group discussions for cancer related outcomes. *BJU International, 120*, 59–61. [CrossRef][10.1111/bju.14036]

Ranasinghe, W., De Silva, D., Bandaragoda, T., Adikari, A., Bolton, D., Lawrentschuk, N., … Persad, R. (2018). The PRIME framework for investigating emotions and other patient factors in low-intermediate risk prostate cancer patients based on online cancer support group discussions. *Annals of Surgical Oncology*.

Savova, G. K., Masanz, J. J., Ogren, P. V, Zheng, J., Sohn, S., Kipper-Schuler, K. C., … ODIE. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, *17*(5), 507–13. [CrossRef][10.1136/jamia.2009.001560][Mismatch]

Tang, B., Wu, Y., Jiang, M., Chen, Y., Denny, J. C., & Xu, H. (2013). A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association : JAMIA*, *20*(5), 828–35. [CrossRef][10.1136/amiajnl-2013-001635][Mismatch]

Tanis, M. (2008). Health-Related On-Line Forums: What's the Big Attraction? *Journal of Health Communication*, *13*(7), 698–714. [CrossRef][10.1080/10810730802415316][Mismatch]

Vydiswaran, V. G. V., Mei, Q., Hanauer, D. A., & Zheng, K. (2014). Mining Consumer Health Vocabulary from Community-Generated Text. *AMIA Annual Symposium Proceedings*, 1150–1159. [Mismatch]

Zhang, Y. (2015). Understanding the sustained use of online health communities from a self-determination perspective. *Journal of the Association for Information Science and Technology*

Zhu, H., Ni, Y., Cai, P., Qiu, Z., & Cao, F. (2012). Automatic extracting of patient-related attributes: disease, age, gender and race. *Studies in Health Technology and Informatics*, *180*, 589–93.

Table 1: Examples for narrative type resolution

Post	**Pronoun resolved post**	**Human mention nouns wit**
My doctor … yesterday. He … daily. I took it … and … I felt … Is this normal …? I was … because … I haven't … and I was … with my fiancée. I got paranoid about her … even though I knew … she …	My doctor … yesterday. **<doctor> …** daily. I took it … and … I felt … Is this normal … ? I was … because … I haven't … and I was … my fiancée. I got paranoid about **<fiancée> …** even though I knew … **<fiancée>** …	I -> 7 Doctor -> 2 Fiancée -> 2

My mother is diagnosed … She is suffering with … I know that … her heart. An endocrinologist suggested her to … Also, he recommended her to … But now she had … She is suffering …	My mother is diagnosed … **<mother>** is suffering with … I know … **<mother>** heart. An endocrinologist suggested **<mother>** to … Also , **<endocrinologist>** recommended **<mother>** to … But now **<mother>** had … **<mother>** is suffering …	Mother -> 5 Endocrinologist->2 I->1
You should consider … such …. You may even want to …. Try to .., but if you've … your other options you may ….	You should consider … such …. You may even want to …. Try to …, but if you've … your other options you may ….	You -> 5

Table 2: Positive and negative samples of age related phrases found in OSG posts

Age Phrases (positive examples)	**Non Age Phrases (negative examples)**
when i was in my 20s	My SVT started about 10 years ago
In my 41 years of life	I had a 48 hr monitor

I am 42 yrs old	I am using atenolol 50 mg
my 25 year old son	doc prescribes 10 tablets 1 twice a day
operated on at 10 days old	Currently I weigh 16 ¾ stone

Table 3: A Sample of features extracted from a text chunk

Feature	**Applied segments**	**Representing terms**
First-person pronouns	L	I, Im, Iam
Possessive pronouns	L	my, our
Family relationship mentions	L,R	mother, mom, father, dad, brother, sister, son, daughter, …

State of being verbs	L	am, is was, were, will be
Age related prepositions	L	at, until, around, under, about, abt
Year Mentions	MI R	year, years, yo, yrs, y/o, s
Time mentions	MI R	days, hours, hrs, minutes, min, weeks, wks, …
Dose mentions	MI R	mg, dose, tablets, mgs, micrograms, ug

Table 4: Examples for gender resolution

Post	**Pronoun resolved post**	**Narrative type**	**Inference of gender**
My doctor … yesterday. He … daily. I took it … and … I felt … Is this normal …? I was … because … I haven't … and I was … with my fiancée. I got paranoid about her … even though I knew … she …	My doctor … yesterday. **<doctor> …** daily. I took it … and … I felt … Is this normal … ? I was … because … I haven't … and I was … my fianc…e. I got paranoid about **<fiancée> …** even though I knew … **<fiancée> …**	experience: first-person	Gender specific word: **fiancée** Patient is a **male**
My mother is diagnosed … She is suffering with … I know that … her heart. An endocrinologist suggested her to … Also, he recommended her to … But now she had … She is suffering …	My mother is diagnosed … **<mother>** is suffering with … I know … **<mother>** heart. An endocrinologist suggested **<mother> to …** Also , **<endocrinologist>** recommended **<mother>** to … But now **<mother>** had … **<mother>** is suffering …	experience: second-person	Gender specific word: **mother** Patient is a **female**

A friend … when he …. He suddenly … hurt his back. He fell down …. Despite … him to his feet and made him walk to … away.	A friend … when **<friend> ….** **<friend>** suddenly … hurt **<friend>** back. **<friend>** fell down …. Despite … **<friend>** to **<friend>** feet and made **<friend>** walk to … away.	experience: second-person	Gender specific word: **friend** **'he'** resolved to **friend,** thus friend is male Patient is **male**

Table 5: Statistics about extracted age, gender and narrative type

Posts	**Count (percentage)**
Total	797,438 (100%)
Age resolved	446,137 (56%)
Gender resolved	443,323 (56%)
Age and gender resolved	371,517 (47%)
Discuss experience	637,881 (80%)
experience: first-person	623,106(78%)
experience: second-person	14,775 (2%)
Provide advice	159,557 (20%)

Table 6: Excepts from the top five results obtained using the three search approaches (including the proposed method) to retrieve similar experiences for the query “I’m a 40 year old woman taking Nexium for heartburn”

Querying method	Excerpts from top five results

The proposed method Breaks the query into the following structure: {symptom- heartburn, medication- Nexium, age-40, gender- female}	*(female,40)*: How I cured my gastritis… side effect of fish oil is **heartburn**… my doctor said I could try **Nexium** as well I decided not to… **My husband** made gluten-free banana bread… I am **40**… *(female,40)*: I had a very severe attack during a 24 hour ph probe test. I used to have **heartburn**… I am on nexium, ranitadine and donperidone… I too am **40** years old but I feel 80… *(female,43)*: My …it's been in a long time and the heartburn seems to be easing off… inhibitors are medication for reducing acid in your tummy like nexium and zoton… I'm only **43** and feel my life is… *(female,50)*: I'm a **50yo** Aussie female with Barrett's… I still have **heartburn** if I eat/drink the wrong things or forget to take my medication for a few hours… I was on **Nexium** 40 forever until I saw a different GP… *(female,30)*: i am bloated and get **heartburn** all the time…iam **30yrs** old wiv 4kids…take them today along with **nexium**. **Nexium** in my opinion is of no use at all…
Full-text search using "**heartburn Nexium woman 40**" (searched in the actual forum search of patient.info website)	**Nexium** and side effect, anyone here while taking **Nexium** suffer diarrhea… I was taking another PPI tablet for **heartburn** for about 2 years(**Nexium**) Constant burping …no pain or **heartburn**. I have been taking **nexium** for the past two weeks to relieve constant burping… just wondering if … if you have been on **Nexium** for a long time? **Women** would be the ones who might find their iron levels low… I am a **40** year old **woman** with no notable health problems aside from the nephrotic syndrome. At 10 years old I was diagnosed with primary focal segmental glomerulosclerosis
Full-text search using "**heartburn AND Nexium AND woman AND 40**" (searched within the collected posts using Elasticsearch)	the meds slowly over two years improved symtoms, during the symptom phase I started with different pain, chronic **heartburn**, leading to gall bladder removal…bought **nexium** as I'd read theyre same as omp, they are expensive… they use **40**-80gm…. but tomorrow will try cabbage juice, just told **woman** on

	cfs site who has chronic nausea to try it

Table 7: Top five symptoms co-exists with the symptom ‘heartburn’ in different age-groups

Age group	**Top five symptoms co-exist with ‘heartburn’**
< 20	less sleep, depressed, tiredness, anxiety, stress, living alone
21 to 40	reflux, anxiety, less sleep, nausea, stress
41 to 60	anxiety, stress, depression, reflux, indigestion
61 to 80	reflux, indigestion, anxiety, constipation, less sleep

Table 8: Performance statistics of the narrative type resolution module

Label	# posts	Precision	Recall
Experience	329	0.92	0.96
Advice	171	0.91	0.81
Combined	**500**	**0.92**	**0.89**

Table 9: Performance statistics of the gender resolution module

Label	# profiles	Precision	Recall
Female	109	0.91	0.87
Male	35	0.90	0.77
Unknown	156	0.87	0.94
Combined	**300**	**0.90**	**0.86**

Table 10: Evaluation of the age resolution module

Label	Number of profiles	Precision	Recall
Age mentioned	131	0.78	0.89
Age not mentioned	169	0.95	0.84
Combined	**300**	**0.86**	**0.87**

Figure 1: (a) existing and (b) the proposed methods of information retrieval from OSGs

Figure 2: The proposed method with the extensible knowledge extraction layer

Figure 3: Age resolution process

Figure 4: A text chunk and its segments L, M, and R

Figure 5: Total time taken for age, gender and narrative type extraction against number of posts processed.

Figure 6: Five dimensions of a OSG post

Figure 7: Demographic distribution of the posts that mention symptom 'heartburn'

Figure 8: Temporal distribution of the posts that mention symptom 'heartburn' over a three-year period