Analyzing Self-Help Forums with Ontology-Based Text Mining: An Exploration in Kidney Space

Philipp Burckhardt, MSc, Rema Padman, PhD
Carnegie Mellon University, Pittsburgh, PA

Abstract

The Internet has emerged as a popular source for health-related information. More than eighty percent of American Internet users have searched for health topics online. Millions of patients use self-help online forums to exchange information and support. In parallel, the increasing prevalence of chronic diseases has become a financial burden for the healthcare system demanding new, cost-effective interventions. To provide such interventions, it is necessary to understand patients' preferences of treatment options and to gain insights into their experiences as patients. We introduce a text-processing algorithm based on semantic ontologies to allow for finer-grained analyses of online forums compared to standard methods. We have applied our method in an analysis of two major Chronic Kidney Disease (CKD) forums. Our results suggest that the analysis of forums may provide valuable insights on daily issues patients face, their choice of different treatment options and interactions between patients, their relatives and clinicians.

1 Introduction

In recent years, the Internet has emerged as a major source of health-related information, with more than 80% of the US Internet-user population, 113 million adults, having used the Web this way.[1] How health information on the Internet affects decision making has been studied thoroughly as well as the demographics of help seekers,[2,3,4,5] yet another area has received much less scrutiny: The proliferation of online self-help forums for chronic diseases that facilitate a social support network for patients.

The reason for this lack of attention is two-fold: On the one hand, the text amassed in these forums is generally too large to be hand-coded and subjected to a qualitative, human-guided analysis. On the other hand, automated text mining with the goal of discovering knowledge is a very challenging task and an active area of research.[6]

In natural language processing (NLP), there has been a lot of work done in areas such as sentiment analysis,[7] document summarization[8] and topic modeling.[9] Most of the current approaches are unsupervised learning methods, which utilize elaborate statistical methods and require a large amount of training data in order to generate sound results.

Although we are living in the era of 'Big Data', the data sets of these specialized self-help forums are much smaller than those of popular social networks. Confronted with this challenge, we opted to create a method for analyzing textual documents that deals with the relative data scarcity. In addition, popular topic modeling algorithms such as Latent Dirichlet Allocation (LDA) are mostly used to cluster documents into broad categories. Yet, our goal is to uncover hidden facts that would otherwise remain unknown.

In our proposed method, we use semantic ontologies to annotate natural language documents and mould them into a machine-readable format, with the goal of encoding the ontological knowledge a human might have. Unlike most statistical methods which use word counts as their atomic units, our representation is based on inter-related semantic concepts, which form an intermediate ontological layer representing a 'web of meaning'. This 'web of meaning', which consists of hierarchically organized conceptual nodes, is then analyzed by statistical means, the results of which can inform and ease content analysis.

Because of aforementioned challenges, not much research has been done in the analysis of self-help forum communication. Existing studies have investigated the impact on attitudes towards healthcare providers running such forums.[10] Others have studied the benefits and challenges arising from the usage of online communities by patients.[11,12,13]

Instead of engaging in the debate on the perceived benefits and disadvantages of such forums, we take the view that even erroneous conceptions and misunderstandings on the side of the patients provide valuable insights insofar as they manifest themselves in the minutiae of online discussion forums and can thus inform researchers on how chronic disease patients deal with their conditions. Taking this perspective, Liu et al. study how such patients use video blogs

to share their stories,[14] and Zhang et al. investigate the intent of people posting on health message boards.[15] There is also a small but growing body of research on how to evaluate the benefits and harms of different treatments and drugs by utilizing comments on online medical forums.[16,17]

The results of analyzing online forums can lead to a broadened perspective for the healthcare sector. Let us assume a doctor who treats a patient with a chronic disease. Not only has there been a decline in the time doctors can spend with their patients,[18] but due to the nature of the doctor-patient relationship, which is confined to single points in time, the doctor only gets a snapshot and not the whole picture of the patients' health. Therefore, it would be desirable for a clinician to learn how the patient feels, whether the treatment is successful and what other issues the patient might have. The analysis of forum communication of patients facing similar problems could extend the vision of the attending clinicians and increase their understanding of patients' needs and preferences.

Discerning communication patterns on self-help networks could also give rise to insights on how patients complement the information received from their doctors. These insights could lead to suggestions to improve the guidelines for doctor-patient communication and to integrate self-help networks into treatment programs.

In 2011, $\$49.2$ billion was spent in the United States on the treatment of End Stage Renal Disease (ESRD), the final stage of CKD.[19] It is estimated that more than 11% of the US adult population have some degree of CKD,[20] with recent projections suggesting that more than 50% of those aged 30 to 64 years will likely develop CKD.[21] In light of these trends, seeking improvement in the care of CKD patients has become a national issue as well as an economic necessity, both in terms of developing new treatments as well as preventive measures. This challenge is not restricted to the US, with CKD being called a "global challenge" requiring concerted action "to avoid a major catastrophe".[20] Although the gained insights might be very valuable, we are not aware of any studies learning communication patterns and insights from freely available CKD self-help forums - there has been only a small-scale intervention study targeted at adolescents.[22] In this study, which serves also as a use case for our methodology, we report on the results of an analysis of two major CKD self-help forums, DaVita and KidneySpace. The insights we have gained in this area render the pursuit of our methodology a promising endeavor.

2 Data

The basis of our analysis is formed by the communication acts from the following self-help forums about CKD: DaVita, a large forum on kidney disease and dialysis hosted by DaVita, one of the primary kidney care companies in the United States, and KidneySpace, a privately operated kidney support forum that was run by the Renal Support Network until it closed down on January 30, 2015.

Table 1: Example thread from KidneySpace forum in JSON format


```
{
 "posts" : [
        {
            "content" : "With all the attention around transplant,
            UNOS has opened up a number for people to call. (...)",
            "title" : "\# to call regarding concerns about your transplant center",
            "user" : "OZfan",
            "date" : " August 17, 2007, 03:17:18 PM PDT "
        },
        {
            "content" : "Hopefully I won't have to use this number.",
            "title" : "Re: \# to call regarding concerns about your transplant center",
            "user" : "Purple_Reign",
            "date" : " August 23, 2007, 12:57:49 AM PDT "
        }
    ],
    "title" : "\# to call regarding concerns about your transplant center",
    "url" : "http://www.kidneyspace.com/index.php/topic,40.1"
}
```


For each thread, our data set contains the following features: Its title, URL address as well as a list of posts, which

includes both the content of the posts as well as information on the author and date the post was written. An example thread is displayed in Table 1.

In total, our data set of the Davita forum consists of $6,501$ threads with $41,079$ posts. The time of the posts ranges from 2004 to the beginning of 2014, for a total of almost ten years. For KidneySpace, we have collected $3,715$ online discussions totaling $20,857$ posts from August 2007 until the year 2013, with 581 unique users present in the data set. For Davita, $5,603$ users are represented in the study. This number is striking insofar as the total number of registered users exceeds $180,000$, implying that a very large number of registered users did not write a single post. This trend continues if we look at the distribution of the number of posts per user, which are displayed for both forums in Figure 1. It seems as if most of the users are mere bystanders who do not actively communicate on the discussion forums, but instead are passive readers. This practice of *lurking* is commonly observed and has been studied in the literature.[23]

Figure 1: Post frequencies for users of both forums. The number of posts has been censored at a value of 25, with the last bar in both plots denoting users with more than 25 posts.

Comparing the two forums, it can be seen that KidneySpace, although a much smaller forum in total, seems to have a more active community: The average number of posts per active user is 2.588 for Davita and 6.196 for KidneySpace.

Before our data analysis, we pre-process the raw forum data and annotate it with ontological information, turning it into a structured, machine-readable and statistically analyzable corpus. The necessary background information and the proposed methodology is presented in the next section.

3 Methods

Traditionally, statistical models for text analysis use word counts as their atomic units. Such a representation, which is devoid of meaning, is ill-suited to the characteristics of languages, which often display *polysemy* and *synonomy* of words.

To illustrate the problem of synonymy, we may use a mundane example: Every native speaker of English knows that the expression "Give me a buck" is equivalent to "Give me a dollar", which in return is equivalent to "Give me a clam". Any word-based model will treat the three words buck, dollar and clam as different entities, and there would be no inherent connection between them. A human listener though recognizes that these three words are synonyms and share a common meaning. However, not only can multiple words refer to the same concept, but a word can have different meanings, too. This polysemy may be illustrated by the word 'lemon' which denotes a fruit as well as a defective automobile.

When we as human beings assign the proper meaning in a given context, we make use of our ontological knowledge. To train an algorithm to make such an assignment, there is a need for a digital ontology. For this project, we have evaluated three commonly used ontologies: ConceptNet, the UMLS Semantic Network and the WordNet taxonomy.[24,25,26] In order to use these ontologies, we wrote APIs for each of them and made them publicly available under an open-source license[1].

Although our proposed methodology is ontology-agnostic, we have settled on WordNet as it does not have some of the

[1] UMLS: `http://bit.ly/1BjRuiM`, WordNet: `http://bit.ly/1MlW421`, ConceptNet: `http://bit.ly/1C2tmDD`

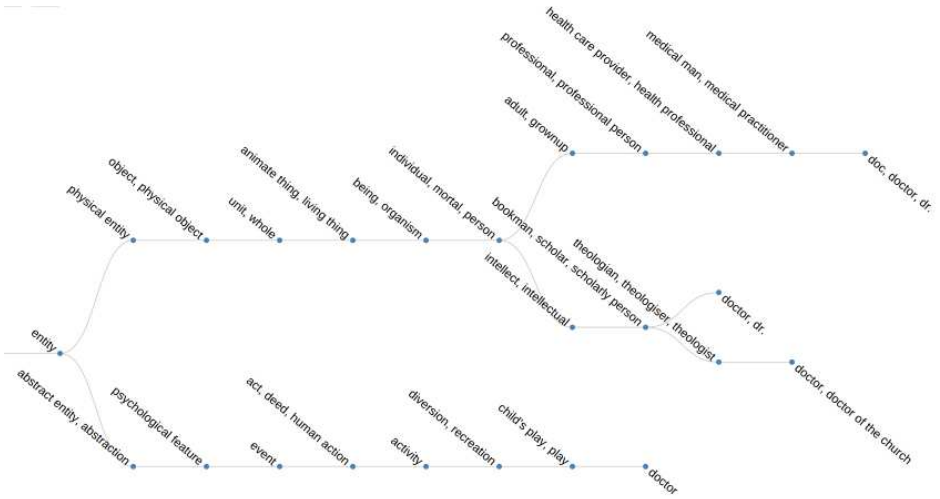

Figure 2: Synsets of the noun 'doctor' and their hypernyms

disadvantages of the other two options. ConceptNet, developed at the MIT Media Lab, aggregates information from various sources, at the cost of structural inconsistencies. The *Unified Medical Language System* (UMLS) contains a vast dictionary of more than two million biomedical and health related concepts (called the *Metathesaurus*), but its general ontology is very flat, consisting of merely 133 inter-related entities, whereas WordNet establishes relationships between all of its $117,000$ concepts, called *synsets*, an abbreviation for *sets of synonyms*. However, a potential issue with WordNet might be its missing specialized medical concepts. Therefore, we thought of enhancing it by UMLS. Since UMLS does not contain colloquial language needed in our context, we considered the *Consumer Health Vocabulary*[2], which provides a mapping from colloquial terms to the relevant concepts in UMLS. But since we found that approximately 84% of words could be assigned to a WordNet synset, the marginal benefit from incorporating UMLS information into WordNet seems limited.

The advantage of mapping words to their respective ontological concepts is that the relationships between these concepts enable us to attach background information to our data. We focus on the `Is-A` relationship between concepts. This results in a taxonomy that can be represented as a tree structure with very general concepts on the higher layers of the tree and very specific concepts as leafs. For two concepts sharing an `Is-A` relationship, we refer to the broader concept as the *hypernym* of the more specific concept. In Figure 2, the possible synsets for the noun 'doctor' are displayed as well as their hypernyms. Because of this tree of knowledge, we can deal with the linguistic problems of polysemy and synonymy.

Pre-Processing: We start by tokenizing each input document and then run a standard Part-Of-Speech (POS) tagger to infer the lexical class of each extracted token. We remove interpunctuation, convert characters to lower case and strip off extra whitespace occurring in the documents. We then exclude all stop words from the analysis. Our system uses a stopword list of 571 words, including conjunctions such as *and*, auxiliary verbs such as *should* and pronouns like *they*. After removal of stop-words, we use a custom-implementation of WordNet's *morphy* algorithm for morphological processing in order to map inflected forms back to their base forms. One of the input parameters of the *morphy* function is the POS tag, which identifies words as a *noun*, *adjective*, *verb* or *adverb*. Because of this intermediate step, it is possible to associate each word with all candidate synsets from the ontology, i.e. all possible meanings the word could have. Words for which no base form could be inferred are removed from further analysis. The steps of the pre-processing stage are visualized in Figure 3.

Word Sense Disambiguation:

The goal of *word sense disambiguation* is to choose the appropriate sense of a word given its context. For each word, we wish to select the correct synset c from the candidate set $\mathcal{C}$ of possible synsets. For example, when the word

[2]See: `http://www.consumerhealthvocab.org/`

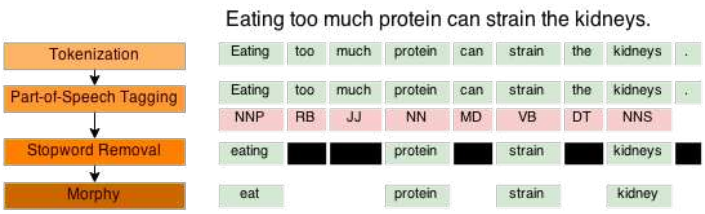

Figure 3: Pre-Processing Steps

'lemon' occurs in a text about fruits, we would like to choose the synset with the definition *yellow oval fruit with juicy acidic flesh*, whereas in a text about used cars it is highly probable that 'lemon' refers to the synset {lemon, stinker} denoting *an artifact (especially an automobile) that is defective or unsatisfactory*.

Similarly to the approach by Fodeh et al.,[27] for any given word w_i and its context $\mathcal{W}(i)$ (in our case, the other words in the sentence of the i-th word), we choose the synset c_i as

$$\arg\max_{c_1 \in C_i} \sum_{w_j \in \mathcal{W}(i)} \max_{c_2 \in C_j} \operatorname{sim}(c_1, c_2), \tag{1}$$

where C_i and C_j are the sets of possible synsets for words w_i and w_j and $\operatorname{sim}(c_1, c_2)$ is some similarity measure between the synsets c_1 and c_2. The idea behind this equation is familiar to anybody: That the narrative context gives each word its special meaning.

In the literature, many different similarity measures have been proposed:[28] some depend on the taxonomy and deem two synsets to be more semantically similar the closer the two nodes are together in the hypernym tree,[29] other so called *Information Content* metrics take probabilistic information of synset occurrence into account.[30] Jurafsky and Martin argue[28] that of the similarity or distance based information measures, the Jiang-Conrath similarity measure[31] performs most consistently. This measure takes both taxonomic and probabilistic information of synset occurrence into account. Inserting the definition of the Jiang-Conrath similarity measure into Formula (1), we estimate the synset for word w_i as

$$\arg\max_{c_1 \in C_i} \sum_{w_j \in \mathcal{W}(i)} \max_{c_2 \in C_j} \left[-IC(c_1) - IC(c_2) + 2 \times IC(lcs(c_1, c_2))\right], \tag{2}$$

where $IC(c)$ denotes the self-information of concept c, a quantity defined as $IC(c) = -\log P(c)$, $P(c)$ being the probability to observe an instance of concept c. We choose the context $\mathcal{W}(i)$ as the sentence in which word w_i appears. In the formula, lcs denotes the *lowest common subsumer*, the lowest node in the hypernym tree which is an ancestor for both c_1 and c_2. This formula reflects two forces: On the one hand, concepts are more likely to be chosen the more frequent they are. On the other hand, the information content of the lcs has the effect of favoring synsets that share a very specific lcs: For example {nurse} and {doctor} have {medical practitioner} as lcs. Its information content is higher than in the case where the lcs for both synsets would be just {entity}. Thus, in co-occurence with 'nurse', we conceive the 'doctor' as a {medicial practitioner} and not as an {intellectual}.

In order to compute $IC(c)$, we estimate the probabilities P via their empirical frequencies. Obtaining synset counts of a document corpus is often prohibitively costly, as it would require to manually annotate each word with its correct synset. A common practice is not to insist on obtaining a sense-tagged corpus and instead calculate the empirical frequencies by summing up the number of occurrences of words belonging to synset c.[30] This method is undesirable for disambiguation, as its only effect will be that synsets with multiple words end up having a higher probability of being chosen. Instead, we make use of the Brown corpus, one of the few corpora for which the occurence of synsets has been determined by annotating its documents by hand. However, the corpus size is only 500 documents, so the Brown corpus has the drawback that many synsets are assigned a count of zero. To increase disambiguation performance, a straightforward approach is to manually annotate a training set from the domain of interest, in our case

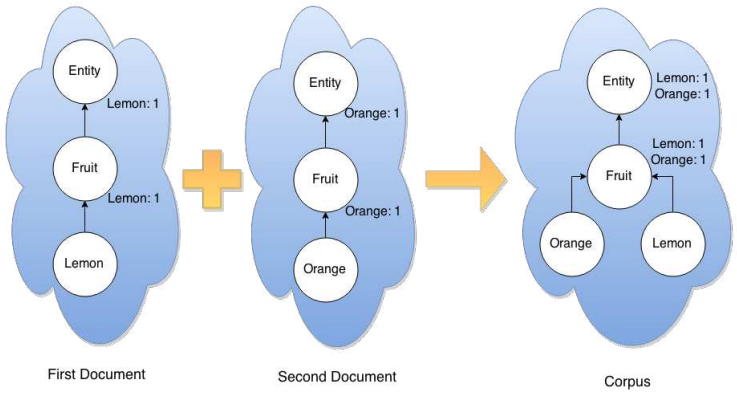

Figure 4: Simplified Illustration of the Merging Process

medical self-help forums on CKD. In order to make this process easier, we have developed a command-line tool where users can manually annotate words in training sentences with their correct synsets. This allowed us to increase the percentage of correctly disambiguated synsets to around 75%.

Document Merging and Word Propagation: Each document is now represented as a set of disambiguated synsets. Since we have augmented each synset with ontological knowledge in form of its hypernyms, it has now the form of a document tree. Next, we merge all document trees together into a single corpus tree (see Figure 4). In doing so, we take all words that have occurred in the corpus and attach them not only to their respective synset, but also to all its hypernyms. To make this clear, consider the example of a 'nurse' which belongs to the $\{$nurse$\}$ synset. However, a $\{$nurse$\}$ `Is-A` $\{$health professional, health care provider, caregiver$\}$, and the word propagation algorithm ensures that the word 'nurse' is also attached to the hypernym synset. This allows us to pose questions such as 'Which words denoting health professionals appear in our corpus?' and obtain answers that also include all words which belong to children of $\{$health professional$\}$.

Thresholding: Many synsets appear only once or twice in a document corpus and could be the result of random noise. We want to discard these synsets and only keep those that are reflective of the inherent characteristics of the document collection. As a default choice, we remove all synsets from the tree which appear in less than $\sqrt{n}$ of the documents, where n denotes the total number of documents.

Analysis: The analysis of the generated synset tree proceeds in three different steps. First, we have a global view on the univariate frequencies of the concepts. From this most distant zoom level, we have a bird's eye view that allows us to detect the most probable and conspicuous synsets. The ontological tree structure in its visual form allows us to easily explore the corpus, for example by starting from highly abstract concepts drilling down to more specific ones. To give an example, we might encounter the abstract synset $\{$emotion$\}$. Looking what kind of emotions appear, we would see that $\{$cravings$\}$ appears surprisingly often, and that the word most often disambiguated to this synset is 'appetite'. A likely hypothesis might be that articulated desires are almost exclusively linked to nutrition.

For the identified synsets (in above example, nutrition or cravings), we take a detailed view investigating their Pearson correlation with other synsets. This allows us to identify interaction fields between different synsets and sharpen our hypotheses. Because we are looking for real correlations and not chance occurrences, we have to carry out hypotheses tests. Here, we also have to account for multiple testing since for any given synset, all possible pairwise correlations with other synsets (for which there is no word-overlap) are calculated. We do this by controlling the false discovery rate (FDR), i.e. the expected percentage of correlations wrongly considered significant, at level 5%.[32] Specifically, we use the procedure by Benjamini-Yekutieli[32] as it is applicable even under arbitrary dependencies among tests. In our example, we find to our surprise that the most correlated synset with $\{$cravings$\}$ is $\{$nausea,sickness$\}$, raising the question whether the patients are not actually complaining about a loss of appetite.

In a final step, we check the formulated hypotheses by tracking the proportion of posts in which they can be established.

To do this, we draw a simple random sample from the documents in question (a size of 100 is chosen as a default), in our example documents containing remarks about 'appetite'. Selecting a random sample helps in eliminating potential biases by giving all documents an equal probability to be selected.

A human judge then goes through automatically generated summaries of the selected documents. This human-guided step is crucial as there are often multiple explanations for the observed frequencies and correlations, and only a look at the actual documents can reveal which of them apply.

4 Results

CKD patients eventually have to face a decision regarding which treatment to undergo. The following options exist: Hemodialysis (HD), peritoneal dialysis (PD) and kidney transplantation. The different treatment options all have their own risks and benefits, so patients might not be sure which suits them best. In addition to that, patients might have to change treatment because a transplant is rejected and has to be replaced by PD or HD. They might also have to switch between dialysis types as a result of of high blood pressure, fluid overload or infections.

Hence, it is no wonder that discussions on all three options make up a considerable amount of our corpus. In the opening posts, the synset {haemodialysis, hemodialysis} is mentioned 785 times, and the synset {organ transplant, transplant, transplantation} even appears in a total of 1,409 documents. Table 2 displays the list of synsets most correlated with the respective treatments. To provide some perspective: the average correlation coefficient of the pairs formed by the 1,876 distinct synsets in the corpus above the threshold level is approximately 0.0435 with a standard deviation of 0.0598.

Table 2: Table displaying synsets correlated with treatment options. CIs are adjusted to control the false coverage-statement rate (FCR) at level 5%.[33] To save space, each synset is identified by one representative synonym, e.g. {dark, night, nighttime} is displayed as {night}.

transplant			**hemodialysis**			**pd**		
Synset	Corr	95% CI	Synset	Corr	95% CI	Synset	Corr	95% CI
donor	0.34	0.32, 0.36	peritoneal	0.25	0.23, 0.27	catheter	0.28	0.26, 0.30
kidney	0.33	0.31, 0.34	intervention	0.18	0.16, 0.20	exchange	0.19	0.17, 0.21
time period	0.28	0.26, 0.3	patient	0.16	0.14, 0.18	time period	0.16	0.14, 0.18
list	0.24	0.22, 0.26	home	0.15	0.13, 0.17	way	0.15	0.13, 0.17
recipient	0.24	0.22, 0.26	discipline	0.13	0.11, 0.15	begin	0.14	0.12, 0.17
rejection	0.23	0.21, 0.25	modality	0.13	0.11, 0.15	nurse	0.14	0.12, 0.16
wait	0.23	0.21, 0.25	nephrology	0.12	0.09, 0.14	bag	0.12	0.10, 0.14
message	0.22	0.2, 0.24	change	0.11	0.10, 0.13	commutation	0.12	0.10, 0.14
years	0.21	0.2, 0.23	occurrence	0.11	0.09, 0.13	dialysis	0.11	0.08, 0.13
experience	0.21	0.19, 0.23	experience	0.11	0.09, 0.13	home	0.11	0.08, 0.13
donate	0.21	0.19, 0.23	therapy	0.11	0.09, 0.13	night	0.11	0.08, 0.13
information	0.21	0.19, 0.23	death	0.11	0.09, 0.13	weaken	0.10	0.08, 0.13
create	0.21	0.19, 0.23	failure	0.11	0.09, 0.13	fluid	0.10	0.08, 0.13
transfer	0.20	0.19, 0.22	rate	0.10	0.08, 0.12	feel	0.10	0.08, 0.12
possession	0.20	0.18, 0.22	years	0.10	0.08, 0.12	peritoneal	0.10	0.07, 0.12

Whereas forum members mentioning transplants seem to be most interested in discussing the modalities of receiving a kidney transplant from a *donor*, a large amount of discussions on PD and HD seem to involve comparisons between the two options, e.g. the most highly correlated synset with {haemodialysis, hemodialysis} is {peritoneal}. And indeed, 10% of discussions on the two treatments mention both HD and PD.

A commonly raised objection concerning HD is the need to visit a dialysis center, with many patients preferring to undergo dialysis at home, either in the form of home HD or PD. {home} frequently co-occurs with both PD and HD. PD patients or those considering it seem to be mostly discussing its practical aspects: Synsets like {catheter}, {exchange}, {bag} and {dark, night, nighttime} all refer to the immediate process of PD. Because of this practical

bias in the results, we assumed that a large proportion of users discussing PD have practical experience. Manually checking a random sample of 100 opening posts mentioning PD, we find that 67 did already undergo PD treatment while 33 did not, so significantly more patients belong to the former group (95%-CI: $0.57, 0.76$).

Besides considering the three different treatment options, some of our other observations are the following:

1. We found a disproportionate occurrence of the synset {hubby, husband, married man} compared to {married woman, wife}, with only 122 first threads mentioning the latter and 433 mentioning husbands.
2. Of all the $3,174$ first threads, 617 reference an ancestor, most of the time either mother or father.
3. A very large number of discussions centered around food, with {food, nutrient} appearing in $2,623$ of $10,216$ opening posts. Besides several ingredients, we found {formula, recipe} to be highly correlated with the food synset (Corr: 0.195, 95% CI: $[0.176, 0.215]$).

The first observation could mean that women constitute the majority of forum members or at least of those who care for a significant other. To check this, we manually labelled the gender of 200 users by looking at their opening threads. We positively identified 94 members, of which 67 were women and 27 men. The 95% CI for the percentage of female users is $(0.58, 0.76)$, indicating that significantly more women than men are active in the considered CKD discussion forums. Because of the second observation, we might guess that up to 20% of all opening threads were created by someone caring for a relative. Out of a random sample of 100 threads, we have manually identified 79 posts coming from someone caring for an ancestor, in most cases father or mother, and five threads alluding to a question on the hereditary nature of the disease. 16 threads were of an unspecified nature. This confirms our hypothesis that offspring of patients are concerned about the health of their parents and seek help online. Notice that in both cases, we restricted our view to the first threads created by each user. We decided to do this as it ensures that heavy users do not bias the results when inferring a characteristic of the forum population. The third observation triggered the hypothesis that while patients may *in theory* know what their diet should look like, they could be missing recipes that can help them to practically adapt to the new dietary requirements. In the documents, {formula, recipe} appeared 220 times, and in a random sample of 100 of opening posts containing both food and recipe, 40 posters specifically asked for recipes, 36 shared recipes and 24 discussed some other recipe-related topic. These results suggest that 3% of users who discussed food specifically asked for recipes. (95%-CI: $0.026, 0.042$).

5 Discussion

We found Internet self-help forums to be an unique and potentially fruitful medium for learning about the daily issues CKD patients face. Our results indicate that there might be a need for stronger embedding of relatives in the treatment process. We also detected a gender gap, with significantly more women active on the forums than men. Further research might provide insights whether this finding applies only to the two considered forums or generally, and if so, what measures could be taken to increase men's participation. A bit surprising was our observation that many people seem to discuss alternatives to classical hemodialysis such as *home hemodialysis* or *peritoneal dialysis*, even though the percentage of patients on these treatments combined is only around 10% in the US. It is unclear to us whether the clientele opting for these more unpopular treatment options is particularly contracted in the two considered forums or whether in general there exists a much larger demand for PD and *home hemodialysis*. Another explanation for this might be that patients engaged in these forums are highly motivated and more active in their disease management. Finally, a significant number of patients inquired about the types of food they were still allowed to eat, asking for recipes satisfying their nutrition requirements.

As we have seen, self-help forums accumulate a lot of information over time. This makes the forums data hubs that patients can query to answer their questions. However, not all of the accumulated information is accurate, and the shared first-hand experiences might not be representative. Therefore, it is necessary to investigate the potential biases existing in the opinions and experiences expressed on these forums. In addition, analyzing the differences in patient's reactions towards first-hand experiences of fellow patients compared to the advice of their doctors might shed further light on how patients' needs are addressed best.

Analyzing online communication data comes with a lot of challenges, requiring the development of new approaches for text analysis. The results of our proposed methodology are encouraging, but there are several potential improvements.

One of them would be to enhance the disambiguation algorithm, which is an active research topic in itself. One limitation of our current algorithm is that the similarity score for words of different parts of speech is always zero due to the WordNet hierarchy. Hence, if there was only one word of a given part of speech in a context, the most frequent synset is chosen without taking the neighboring words into account. A potential remedy could be to leverage WordNet's synset definitions in the similarity calculation. This would be similar to the Lesk algorithm[34] which was adpated to WordNet previously.[35]

Alternatively, disambiguation could be improved by extending the Brown corpus or by integrating ontologies such as UMLS to compensate WordNet's deficiencies in the medical domain. Another limitation of our current method is that it does not take syntactic information into account and thus relies on a human-guided final step. If this could be incorporated, one would not only pertain semantic information, but could also identify the interactional relations: Who acts and who is acted upon? However, given the complexity of such a task, this might be a long-term project. From a short-term perspective, it might be more fruitful to develop methods for calculating synset similarity and clustering techniques of synset trees in order to more easily identify discussed topics. Since our methodology is not tied to a single medical domain, one could think of applying it in other areas. In particular, we think of forums where due to the nature of the problem one would expect high veracity (chronic or lethal diseases) or where due to a high social stigma there is a lack of reliable information (e.g. depression).

6 Conclusion

Given the persistent rise in costs related to CKD, developing new preventive measures to curb disease progression has become an economic necessity. In this study, we have developed a process to systematically analyze online self-help forums. We encountered a few surprising phenomena: The large gap in user activity, the gender gap, forums as knowledge hubs and the desire for practical information. The study of online self-help forums broadens the perspective and allows medical practitioners to gain insights into the daily routine of chronic disease patients. Incorporating the insights drawn from such analyses might allow doctors to take a more holistic view of their patients and promote more personalized treatments. Information extraction and analyses of communication of chronic disease patients might therefore play an important role in developing new preventive care approaches.

Acknowledgement

We wish to thank David Choi for his contributions and valuable feedback provided in many discussions about the project.

References

[1] Fox S. Online Health Search 2006. Search. 2006;209-419-45:1 – 15. Available from: `http://www.pewinternet.org/Reports/2006/Online-Health-Search-2006.aspx`.

[2] Baker L, Wagner TH, Singer S, Bundorf MK. Use of the Internet and e-mail for health care information: results from a national survey. JAMA. 2003;289:2400–2406.

[3] Bundorf MK, Wagner TH, Singer SJ, Baker LC. Who searches the internet for health information? Health Serv Res. 2006;41:819–836.

[4] Laurent MR, Vickers TJ. Seeking health information online: does Wikipedia matter? J Am Med Inform Assoc. 2009 Jan;16(4):471–9.

[5] Feufel MA, Stahl SF. What do web-use skill differences imply for online health information searches? In: J. Med. Internet Res.. vol. 14; 2012. .

[6] Stavrianou A, Andritsos P, Nicoloyannis N. Overview and semantic issues of text mining. ACM SIGMOD Rec. 2007;36(3):23.

[7] Pang B, Lee L. Opinion Mining and Sentiment Analysis. Found Trends Inf Retr. 2008;2:1–135.

[8] Erkan G, Radev DR. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. J Artif Intell Res. 2004;p. 457–479.

[9] Blei D. Probabilistic Topic Models. Commun ACM. 2010 Nov;p. 77–84.

[10] Nambisan P, Gustafson D, Pingree S, Hawkins R. Patients' sociability and usability experience in online health

communities: Impact on attitudes towards the healthcare organisation and its services. Int J Web Based Communities. 2010;6:395–409.
[11] Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Primary care Health related virtual communities and electronic support groups :. 2004;328(May):1–6.
[12] Kraut R, Patterson M, Lundmark V, Kiesler S, Mukopadhyay T, Scherlis W. Internet Paradox. Am Psychol. 1998;53(9):1017–1031.
[13] White M. Receiving social support online: implications for health education. Health Educ Res. 2001 Dec;16(6):693–707.
[14] Liu LS, Huh J, Neogi T, Inkpen K, Pratt W. Health Vlogger-Viewer Interaction in Chronic Illness Management. Proc SIGCHI Conf Hum factors Comput Syst CHI Conf. 2013;2013:49–58.
[15] Zhang T, Cho J, Zhai C. Understanding User Intents in Online Health Forums. IEEE J Biomed Heal informatics. 2015 Mar;Available from: http://www.ncbi.nlm.nih.gov/pubmed/25823052.
[16] Cho JHD, Liao VQZ, Jiang Y, Schatz BR. Aggregating Personal Health Messages for Scalable Comparative Effectiveness Research. In: Proc. Int. Conf. Bioinformatics, Comput. Biol. Biomed. Informatics - BCB'13; 2007. p. 907–916.
[17] Wang S, Li Y, Ferguson D, Zhai C. SideEffectPTM: An Unsupervised Topic Model to Mine Adverse Drug Reactions from Health Forums. In: Proc. 5th ACM Conf. Bioinformatics, Comput. Biol. Heal. Informatics. BCB '14. New York, NY, USA: ACM; 2014. p. 321–330.
[18] Dugdale DC, Epstein R, Pantilat SZ. Time and the patient-physician relationship. J Gen Intern Med. 1999 Jan;14 Suppl 1:S34–40.
[19] ECLRJC P, AMKC P, AMSPQCQ P. Costs of end-stage renal disease; 2013. Available from: http://www.usrds.org/2013/pdf/v2_ch11_13.pdf.
[20] El Nahas aM, Bello AK. Chronic kidney disease: The global challenge. Lancet. 2005;365:331–340.
[21] Hoerger TJ, Simpson SA, Yarnoff BO, Pavkov ME, Ríos Burrows N, Saydah SH, et al. The Future Burden of CKD in the United States: A Simulation Model for the CDC CKD Initiative. Am J Kidney Dis. 2015 Mar;65(3):403–11.
[22] Nicholas DB, Picone G, Vigneux A, McCormick K, Mantulak A, McClure M, et al. Evaluation of an Online Peer Support Network for Adolescents with Chronic Kidney Disease. J Technol Hum Serv. 2009 Feb;27(1):23–33.
[23] Nonnecke B, Preece J. Why lurkers lurk. AMCIS 2001 Proc. 2001;p. 1–10.
[24] Liu H, Singh P. ConceptNet - a practical commonsense reasoning tool-kit. BT Technol J. 2004;22:211–226.
[25] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32:D267–D270.
[26] Fellbaum C. WordNet. In: Theory Appl. Ontol. Comput. Appl.; 2010. p. 231–243.
[27] Fodeh S, Punch B, Tan PN. On ontology-driven document clustering using core semantic features. Knowl Inf Syst. 2011 Jan;28(2):395–421.
[28] Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. vol. 21. Prentice Hall; 2000.
[29] Wu Z, Palmer M. Verb Semantics and Lexical Selection. 1994 Jun;.
[30] Resnik P. Semantic Similarity in a Taxonomy: An Information Based Measure and Its Application to Problems of Ambiguity in Natural Language. J Aritificial Intell Res. 1999;11:95–130.
[31] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv Prepr C. 1997;(Rocling X). Available from: http://arxiv.org/abs/cmp-lg/9709008.
[32] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under depencency. Ann Stat. 2001;29(4):1165–1188.
[33] Benjamini Y, Yekutieli D. False Discovery RateAdjusted Multiple Confidence Intervals for Selected Parameters. J Am Stat Assoc. 2005;100(July 2012):71–81.
[34] Lesk M. Automatic sense disambiguation using machine readable dictionaries. In: Proc. 5th Annu. Int. Conf. Syst. Doc. - SIGDOC '86; 1986. p. 24–26.
[35] Banerjee S, Pedersen T. An adapted Lesk algorithm for word sense disambiguation using WordNet. In: Comput. Linguist. Intell. text .... vol. 2276; 2002. p. 136–145.