CRIMSON PUBLISHERS
Wings to the Research

Examines in Physical Medicine and Rehabilitation: Open Access

ISSN 2637-7934

Research Article

Toward Generating Customized Rehabilitation Plan and Delivering to Engaged Patient

Weijia LU*

**The University of Aizu, Japan*

***Corresponding author:** Weijia LU, PhD. E.E and PhD. C.S, Biomedical Information Technology Lab, The University of Aizu, Fukushima Ken 965-8580, Japan

Submission: August 26, 2017; **Published:** March 28, 2018

Introduction

In 2014, Patrizia and his colleague published an interesting paper about the validated mobile smartphone application for body position measurement in rehabilitation [1]. The conclusion in that paper, they found 12 apps with promising features. Nowadays, this number has been increased to over 300 thanks to the booming of the smart phone and internet [2]. Normally a user community (rehabilitation ecosystem) can be also created to serve as the major feedback method, in which people can put comments on the rehabilitation plan, application itself, or any latest news. For the owner of this application, one may be interested about finding the most engaged user, and deliver customized plan with good quality. That wish can be fulfilled by data mining those comments. On the other hand, Yelp is a multi-dimension recommendation and review system, consisting several aspects of the data: users, business, and reviews. The purpose of this work is to use Yelp dataset to mimic the scenario in the rehabilitation ecosystem and answer following key question: Can insightful advertisement (the customized plan) be generated and deliver to those most engaged user? The insight of an advertisement means that advertisement should contain the key features of a particular domain of business belongs; here it's physical rehabilitation for certain category of patient. These key features should reflect the users' potential needs. And they could already exist in the top-level business model in the domain, or the general rehabilitation plan for certain type of patient in our case, and be pursued by the application owner. The engaged user means those members who may show great interesting to your business (your customized rehabilitation plan). And they show great need to get the service of good quality.

Methods

Load and preprocess the data

Yelp dataset contains 5json files: business, check in, review, tip and user. Most likely, the data from rehabilitation ecosystem should share the similar structure with business equal to rehabilitation plan, review equal to user comments imposed by mobile application. For the question raised before, we use the data belonging to the business, review and user, business, review can be summarized in following tables (list only columns we used), and we only use user_id and name in user data. It is noticed that each business belongs to a certain set of categories c, which can be taken as an observation from a population $\mathcal{C}$ with more than 783 possible entities. Meanwhile, each business originally has an array a_0 (length = 38) to describe the attribute. Moreover, it is noticed that some elements of the attribute array are data frame, thus we can generate a 'flatten' version a_1 (length = 78) from each one. Furthermore, we can 'unlist' the 'Accepts Credit Cards' attribute in a_1 thus to cast each element in a_1 into the basic data type. We notate this final attribute array as a_2.

Generate the insightful key phrase for advertisement

From a new business perspective, the owner can coarsely address the business into one of the category in the business table. And by leverage the existed business instance, they can easily find the common feature to follow by extracting the attributes mostly shared by these businesses. For example, "Eric Goldberg, MD" in dataset belongs to Doctors, Health & Medical. And for those businesses belonging to Doctors (1077 instances), there are 800 businesses only accept appointment patients over 164 who don't. Therefore, the owner probably may make the decision to follow the common trend, and broadcast a proper advertisement based on it to win more customers. Unfortunately, this situation could not always stand. For the same set of businesses, there is only 1 instance who claim they have the TV, while rest in dataset are unknown. In this situation, the owner will be lost. To deal this dilemma, in this section, we bring forward a methodology to extract insightful phrase from the reviews.

Firstly, we find all business instances, called buddies, which share the same category. For example, there are 1077 belonging to Doctors and 3213 belonging to Health & Medical, and the total number of the buddy is 3213 after removing the duplication. We notate the set of the buddy as $\mathcal{B}$. Then, we can find all relevant reviews to buddy set $\mathcal{B}$. And use the Part-of-Speech (POS) tag [3] to annotate each of them, following we show an annotation example,

```
Original tagged review:
 Great[NNP] friendly[JJ] staff[NN] ![.] ![.] Clean[NNP]
 teeth[NNS] and[CC] no[DT] cavities[NNS] ....[:]
 Great[JJ] exam[NN] and[CC] cleaning[NN] .[.] Thanks[NNS] .[.]
```

For our specific purpose (to extract insightful phrase), we may focus on the noun in these reviews. Since academically, in the sentiment analysis, to find the insightful phrase in the text is a typical problem of identification of the product features [4,5]. In an early study [5], the author makes a hypothesis that "Different customers usually have different stories. However, when they comment on product features, the words that they use converge". That heuristic statement leads to an association mining [6] to find all frequent itemsets in reviews. In this study, we move one step forward based on their methodology to extract not only the frequent item sets but also the syntactic structure. And hope the syntactic structure can remove fake nouns not belonging to the real product features. To do that, we manually mark 50 review samples, randomly sampled from corpus, and pick up insightful phrase with explicit opinion statement, e.g:

```
    N4    N3    N2 N1 C0   P1 P2    P3    P4    P5 YN ID
9  <NA><NA>    VB IN NN VBD JJ <NA><NA><NA>   Y  1
10 <NA><NA><NA> DT NN VBZ RB    JJ <NA><NA>   Y  2
                              Comments
9     Check In woman was professional
10 The billing is absolutely horrible
```

C0 in our output table is a noun (NN, NNP, NNS etc), which related to the product feature. The N1-N4 is 4 prefix word, and P1-P5 is 5 postfixes. NA in this table means an irrelevant word either belonging to the next sentence or adjuncts. The comments column in this table is the original text in the reviews.With this table, we use a prior algorithm in association mining [6] to learn the combination rules of the syntactic structure, as following code snippet,

```
#Learn the rule by association mining----
# kkPOSTag[,1:11] contains columns:N4 - YN
kkPOSTag <-kkPOSTag[,1:11]
rule <-apriori(kkPOSTag,
parameter =list(minlen=4),
appearance =list(rhs =c("YN=Y"),default ="lhs"))
ruleRHS <-subset(sort(rule,by="support",decreasing = T), subset = lhs %
pin% "C0=")
# Redundancy pruning as did in [Hu and Liu 2004]
subset.matrix <-is.subset(ruleRHS, ruleRHS)
subset.matrix[lower.tri(subset.matrix, diag=T)] <-NA
redundant <-colSums(subset.matrix, na.rm=T) >=1
rules.pruned <-ruleRHS[!redundant] # our final rules
```

Later on, these rules are used to select the candidate product feature phrase in all reviews related to B. During the selection, each candidate phrase is assigned a float number (weight), to reveal the important of this phrase in current text. This number is from the multiplication of the appearance count in a certain review by the review's star, then divided by the maximal star rank (=5). Finally, we can summarize the weight and provide the result to the user. The phrase with high score could indicate the high frequency of showing up in the relative high quality reviews. And we say these phrases can be used in composing insightful advertisement. More details are demonstrated in the result section.

Find the engaged user

From a new business perspective, or customized rehabilitation plan in our case, it is crystal important to find the engaged users, so that the designed advertisement based on the insightful feature phrase can be delivered. In this section, we try to use collaborative filtering [7] to find those engaged users, based on the similar business. Since people may pay more attention to the attribute of the business, e.g. preference of a club with parking lot, we need to clustering the buddy set B into groups based on the feature, and constrain ourselves on a single subset B_sub. For clustering operation, we need to further clean the business table. If we take a careful look at the business table, we can find that the attributes array a_2 usually contains quite a lot of missing data. So, we need to figure out a way to deal with them. The basic variable type in those arrays including logical, character and integer. For logical and character variable, we can replace NA by an empty string, and then cast each variable sequentially into the character, the factor, and finally the integer. For integer variable, we can use 0 to replace NA value. The next step is to clustering on the business table, and use the group mark to further down select the buddy business into B_sub. The clustering method in this study is K-means and the plot of the total within groups sums of squares against the number of clusters center is used to select the number of the total clustering center. Precisely speaking the bend of the trend suggests a reasonable choice. By using down selected business B_sub called neighbors, we can pickup relevant reviews from the review table. And sort them based on the issue date. Not all reviews is important to our interrogation business, we only look at the reviews belong to the same year of our interrogation. For each review, we have one user ID. Thus, we can goto the user table to find those users. And we only take those who has submit more than one reviews as our potential candidate.

Result

To demonstrate the methodology in previous section, we use first business in the business table as our interrogation target.

Step 1. Learn the syntactic structure from our labeled training samples

```
    lhs                      rhs    support   confidence lift
[1] {N1=DT,C0=NN,P1=VBZ}  => {YN=Y} 0.1764706 1          1
[2] {C0=NN,P1=VBZ,P2=RB}  => {YN=Y} 0.1470588 1          1
[3] {N1=DT,C0=NN,P2=JJ}   => {YN=Y} 0.1470588 1          1
[4] {C0=NN,P2=RB,P3=JJ}   => {YN=Y} 0.1176471 1          1
[5] {N1=DT,C0=NN,P2=RB}   => {YN=Y} 0.1176471 1          1
```

Each rule is expressed as A in lhs column =>B in rhs, the support (A=>B) is defined as P(AUB), confidence(A => B) is defined as a conditional probability P(B|A), and lift(A =>B) is confidence(A =>B) over P(B), since in our table event B is always happened, confidence and lift in the result are always 1. Thus, we only need to look at the

How to cite this article: Weijia LU: Toward Generating Customized Rehabilitation Plan and Delivering to Engaged Patient. Examines Phy Med Rehab. 1(4). EPMR.000518. 2018. DOI: 10.31031/EPMR.2018.01.000518

support, as the probability of the rule met by our training set. And use the top rules with relative high support value.

Step 2. Get the insightful feature phrase

Use the rules and was introduced in previous section, we get the insightful phrase like following, the horizontal axis is the top features with relative high scores (>10), and the vertical axis is the score, namely the sum of the weights. (Figure 1) Moreover, we can explore the original review (id=37) to double check the opinion of the reviewer on a certain topic (e.g. office).

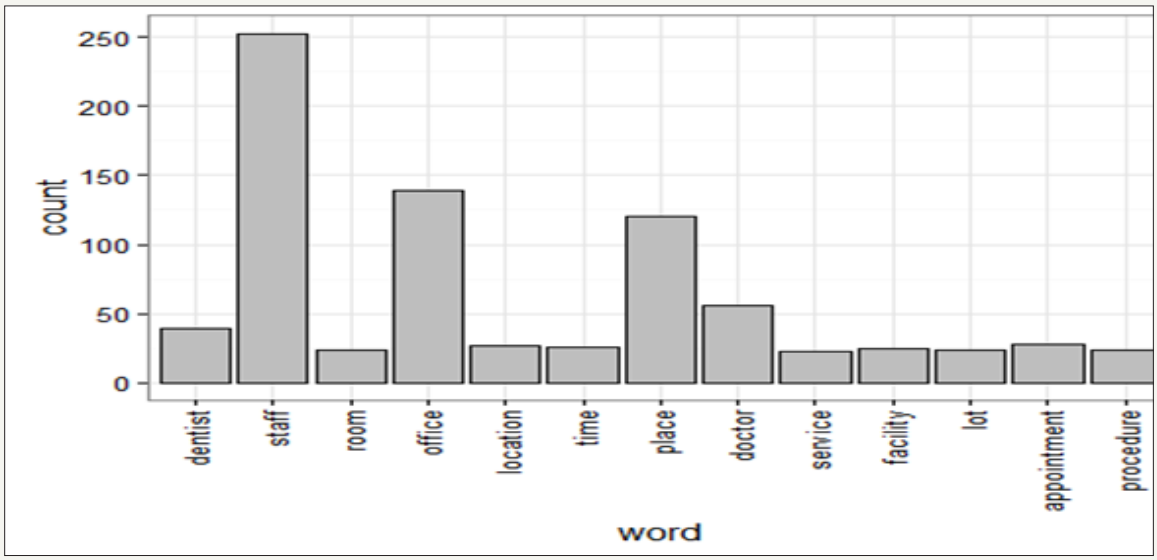

Figure 1: Generated insightful phrase.

```
 id                        phase
1 18     This office does not know
2 23      The office is very clean
3 24   this office are wonderful .
4 28         the office is great ,
5 37 The office is always spotless

The original tagged reviews said,
 I[PRP] love[VBP] Dr.[NNP] Dan[NNP] and[CC] Eric[NNP] ![.] ...
 The[DT] office[NN] is[VBZ] always[RB] spotless[JJ] ,[,] and[CC]
 thanks[NNS] to[TO] the[DT] warm[JJ] decor[NN] it[PRP] does[VBZ] n't[RB
] have[VB] that[DT] sterile[JJ] dentists[NNS] office[NN] feeling[NN]...
```

Step 3. Clustering the B into B_{sub} and find engaged user

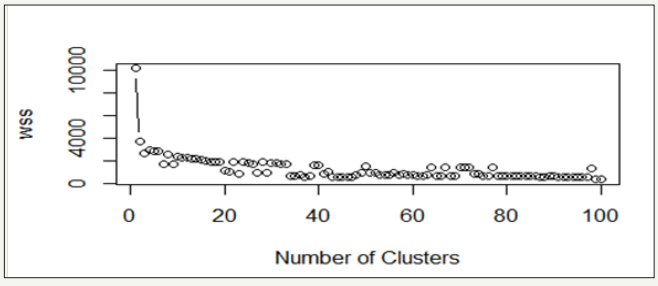

Figure 2: Total with-in cluster distance (wss) under different number of clustering center; This metric is calculated based on the sum of square distance in each clustering group, and smaller value indicate better aggregation of original number to each group center.

Then we clustering the buddy business, before clustering we pick up the total number of the clustering center from plot (=45). After the down selection of the buddy business by clustering result, we get 2576 business as our final neighbors B_sub. Using B_sub, we retrieve the relevant reviews, sort by date, filter by current year, and finally we have 256 reviews. These reviews are further used to select the users (total number = 248), and during which, we select 8 as our target users (Figure 2).

Discussion and Conclusion

In this study, we demonstrate a method to generate insightful advertisement, and find potential user to deliver this advertisement. In the physical rehabilitation, its equivalent task can be described as generate customized rehabilitation plan and deliver to engaged patient. During exploration of the key feature phrase, we use association mining to find syntactic structure. The training set for association mining is aggregated by sampling explicit opinion statement; the opposite side of this kind of statement may be an implicit statement with tremendous complex syntactic structure, or just a simple noun phrase. It is quite difficult to find and characterize this case, thus make it impossible to build a classifier. That's the reason we use association mining instead. The sample size we use is limited, since mark the sample is tedious and arduous. Hopefully, we can generate a larger sample corpus in the next phase of study. For finding the potential user, we use k-means clustering to down select business based on the features. And use the relevant review to find the candidate in a collaborative filtering way. In this step, we make a hypothesis that people, who recently post multiple reviews on a same kind of business, should show great interesting in them. We've explored other features like the statistical feature

How to cite this article: Weijia LU: Toward Generating Customized Rehabilitation Plan and Delivering to Engaged Patient. Examines Phy Med Rehab. 1(4). EPMR.000518. 2018. DOI: 10.31031/EPMR.2018.01.000518

on the interval between reviews in order to profile users. But we can't find good quantification in the raw dataset to support this kind of profiling, so we use a heuristic method during find the potential user. Our method presented in this study belongs to the supervised learning. Literately, there is a family of generative model, called LDA. Using latent distribution like Dirichlet to do statistical interference. And there is also proposal method to combine the syntactic analysis and LDA. Therefore, in the next stage of the study, we also hope to compare the performance and leverage the methodology from those researches.

References

1. Milani P, Coccetta CA, Rabini A, Sciarra T, Massazza G (2014) Mobile smartphone applications for body position measurement in rehabilitation: a review of goniometric tools PMR 6(11): 1038-1043.
2. Corlan, Alexandru Dan (2004) Medline trend: automated yearly statistics of PubMed results for any query. Reference Source (2016).
3. http://en.wikipedia.org/wiki/Brown_Corpus#Part-of-speech_tags_used
4. https://www.cs.uic.edu/~liub/FBS/Sentiment-Analysis-tutorial-AAAI-2011.pdf
5. Minqing Hu, Bing Liu (2004) Mining and summarizing customer reviews. ACM, New York, NY, USA, pp. 168-177.
6. Agrawal R, Ramakrishnan S (1994) Fast algorithms for mining association rules. Proceedings of the 20th VLDB Conference Santiago, Santiago, pp. 487-499.
7. https://en.wikipedia.org/wiki/Collaborative_filtering