Intelligent Recommendation System for WebMD Forum Using Text and Visual Analytics

Jayashree Subramaniam
Graduate Student
Computer Science (MCS)
Arizona State University
jsubram5@asu.edu

Krithika Mahalingam
Graduate Student
Computer Science (MCS)
Arizona State University
kmahalin@asu.edu

Mahathi Srinivasan
Graduate Student
Computer Science (MCS)
Arizona State University
msrini22@asu.edu

Vivek Ranjan Panigrahy
Graduate Student
Computer Science (MCS)
Arizona State University
vpanigr1@asu.edu

ABSTRACT

Visual analytics plays a pivotal role in today's business operations for carrying out decisions, driven by insights gathered from visual reports. It is an incredibly powerful technique and is one of the core components of business intelligence. Rather than just conveying information to the audience with no domain knowledge, it personally and financially benefits them and also benefits the company. In this project, data is gathered from a popular medical forum known as WebMD which serves as an online community where people post questions regarding medical problems that are answered by experts. Hello Doctor, goes a step ahead and helps a layman understand what the solution to the medical problem is, by performing textual analytics on the WebMD data to create stunning visualizations which are easy to interpret and conveys the required information

KEYWORDS

Visual Analytics, WebMD, Text Analytics, Stemming, Lemmatization, TF, DF, IDF

1. INTRODUCTION

The essential purpose of a visualization is to effectively communicate information to an audience with varying familiarity of the topic presented through as a story in such a way that everyone in the audience comfortably understands the information conveyed and that there is no ambiguity raised for the content presented.

Nowadays, with the importance of physical and mental health brought into light, web applications have been created for the same. Such apps collect data from disparate medical forums that not contain a list of diseases and their associated symptoms but also recommendations and solution from health experts, thus providing quality service and benefits to the community. In this project, we have created a web application for data collected from the WebMD forum that displays the statistics of a disease based on the topic searched by the user.

WebMD being a question and answer forum only gives us the responses of various people to questions. As a user, this might be challenging because not every user has the same level of understanding of the diseases which leads to more confusion among the users. Also, some answers in the forum contain technical medical jargons whose meanings are not easily understood by everyone.

To tackle this problem, Hello Doctor introduces the user to a set of visualizations which ensures that the user irrespective of his knowledge of the disease can interpret various aspects of the disease. We help the user predict what kind of symptoms he or she might encounter later the basis of current disease by using text analysis and the chord diagram. We help the user find information of the expert or contributor who is the most influential person pertaining to that topic which helps the user to directly contact the person. The time series visualization showcases the time period wherein the spread of the disease is the most which could help user to predict and prepare for it. Lastly the word cloud for the topic showcases not just the most frequent words used in describing the topic but also solutions (medications) which can help treat the disease. This helps the user to make an informed decision without having to hassle between reading lengthy answers and struggling to make sense from it.

2. MOTIVATION

In the past, people had very limited access to health care, due to limitations in transport and financial problems. Finding treatments for ailments was still in progress and the death rates were significantly higher. After ground-breaking research in science and microbiology, people were able to obtain the right prescription for an exhaustive range of diseases or symptoms, both known and unknown, and the death-rate reduced greatly. With the advent of the internet and big data, traditional health-care models had to evolve to better ones to accommodate usage of data from disparate sources and application of analytics, thus bridging the gap between people and health care significantly. However, in times of crisis or unforeseen circumstances, people will need to have enough knowledge about ailments or diseases to take swift actions when hospitals or any in-person professional medical assistance is not accessible.

A study conducted by Vagelis Hristidis [2] analyses how different social-media users accessed medical data from Web forums. The study came up with approaches to also lend a better hand to

underserved communities while minimizing the spread of misinformation.

In order to serve people who haven't made use of means to medical forums properly or to lend a hand in time of emergency, we decided to design and deploy an application that is easy to use. It takes medical data from the WebMD forum and provides instant solutions to medical problems, both from novices and experts. We chose the WebMD forum to be the source of data since it's the most popular medical forum with an active community of dedicated participants. As for trusted and reliable data, the answers are rated based on the expertise level of the user and number of votes.

3. VISUALIZATION DESIGN

Hello Doctor, is an intelligent interactive system which serves as a one stop shop for users to get solutions for all their medical problems without having to leave the comfort of their room. Our system collates the data from the WebMD medical forum and visualizes 4 different aspects of a particular disease helping the user understand the disease in a better way

Our Model showcases the WebMD information using 4 different charts or views which helps in exploring the data in a more specific manner.

- Bubble Chart
- Time Series
- Word Cloud
- Chord Diagram

Chord Diagram is chosen over heatmap or any other visualization for visualizing the similarities between the topics because it is instantly used to check the similarity with the help of thickness between the topics. On hovering over each topic gives a clear view than using a rectangular view of the Heat Map. It is visually appealing, and we can instantly infer the relationship between the topics. We have performed text analytics for our data which includes the following.

- We have the list of topics /diseases for the WebMD data
- We remove the stop words and perform stemming for the topics and perform the TDIDF for the given data
- We then build a symptom cross symptom matrix where each value in the matrix gives us the similarity between the topics, which is used for plotting the chord diagram.
- We also map the topics across time frames and get the most frequently used words along with the topics using world cloud visualization
- We give a bubble chart for the member contribution and a deeper view to analyse among the two types of members Experts/Contributors answering the questions

3.1. Incorporating the Visualizations

3.1.1 Member Expertise with Bubble Chart

Bubble charts are similar to scatter plots, with bubbles in place of data points. It is artistic, useful in depicting the relation between entries using various dimensions [4]. The bubble is very effective in demonstrating the relative comparison of values with its radius. Hence, we are trying to use a bubble chart to represent the expertise level of a member who has answered the questions on the forum, to help the user determine the reliability of a member's answers in that topic. This will help the user make an informed decision.

We can find top contributors throughout the forum for each disease or topic of data and make sure the effective answers are readily available to users with the least amount of search or navigation.

Primarily, the bubble chart depicts a cluster with the size of the bubbles representing the number of questions answered by that person relating to the selected topic. Also, there are variances shown in terms of colors [5]. A light color, coral, has been used for 'Contributor' and a brighter green color is used for highlighting the 'Experts'. Thus, color coding helps the user in reliable perception of averages and sizes [6]. As the person answers more questions in a domain, it could be derived that they are more influential and knowledgeable in that. Each bubble, on hovering shows the details of the members such as their name, number of questions they have answered in that topic, number of followers, profession and other details. There are also two buttons, 'split' and 'combine' to split or combine the cluster of bubbles respectively. This aids in effective comparison of the size of the bubbles among a single group, say 'Experts'. So, if a user wants to consider only the experts, then the user may easily do so by splitting the bubbles. The overall comparison of the sizes of the bubbles could be done with the combined cluster.

.

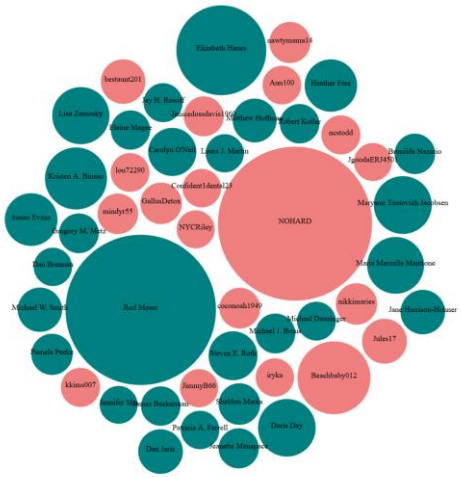

Figure 1. Bubble Chart (Combined View)

It can be seen from the above chart that the number of experts who have answered the questions is more than the number of contributors. The light share of red color denotes the contributors and green denotes the experts. A darker shade is used for experts since they are larger in number than the contributors. We can also do a member type wise comparison, by splitting the cluster. Thus, there are a greater number of bubbles in the experts. Also, the deeper view enables the user to analyse within the member types, who answered the maximum number of questions related to the topics. It splits the view as Experts and Contributors.

Research Question Addressed

- How to effectively find the best contributor for each topic or disease with their additional information?

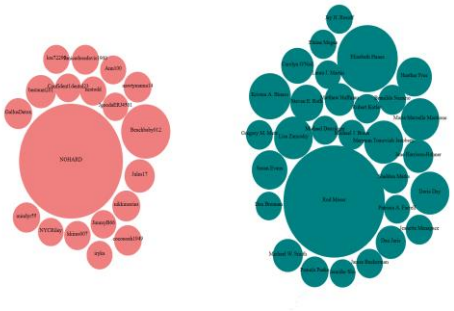

Figure 2. Bubble Chart (Deeper View)

3.1.2 Time Series Visualization

The time series chart demonstrates the frequency of questions being asked in a particular topic over time [7,8]. This helps in identifying trends in the data, hence plays a vital in finding out the diseases that were predominant in a time period.

This time series chart was chosen to give the user a simple and concise visualization of the number of questions over time. The data being dealt with has only one dimension that varies based on time [9]. On hovering, the values of the peaks are displayed and lets the user to slide over the years and months.

The user can select whether he wants to see a Yearly view or a Monthly View and then also reset the entire view with the help of the buttons displayed on the top. For ease of access, there is also a sliding bar that is given, if the user wants to navigate to and view a year's disease trends. This visualization could provide the necessary information in case of a disease break-out, to analyse its trends in the past and for the medical team to be prepared to treat the patients. This visualization of ours helps user to zoom and filter the frequency of the topics by adjusting the slider across the year. We have also provided a reset button to get back to our overall view of the time series chart. It is user friendly and convenient for the user to view only a part or section of the topic frequencies. Thus, it effectively visualizes the historical data.

Research Question Addressed

- How to show the trend of the topic over the years and months based on their frequencies?
- How to predict an upcoming disease by analyzing the frequency of the topics for that disease?
- What is the most common disease across the year and what is the driving cause for the spread of the disease in that year?

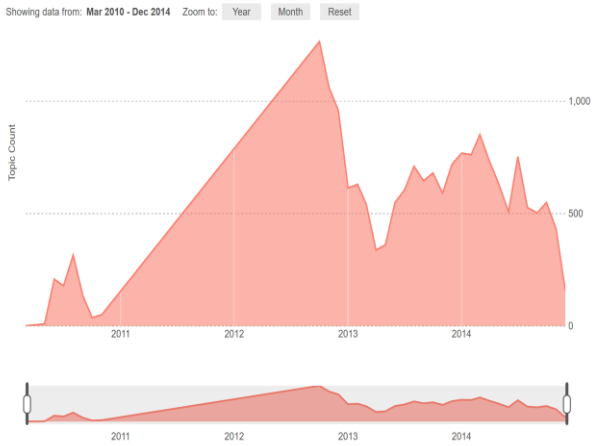

Figure 3. Time Series Chart

3.1.3 Word Cloud

This type of chart gives an analysis over the relationship among various keywords that are mentioned in a topic while being visually appealing [10]. This chart can be taken as the best method for text summarization, as it depicts the relation among words with various dimensions like colors. The properties that affect the user the most are the size, color and weight. Space between two words can also be taken as a measure of correlation [11].

Research Question Addressed

- How to determine the predominant words being spoken in a particular topic and the associated words?

Figure 4. Word Cloud Based on Topic

3.1.4 Similarity Distribution Using Chord Diagram

The chord diagram visualizes inter relationships among given objects or entities [12]. The connections denote that the entities have something in common. This is a useful graph to examine the similarities among samples in a dataset or between different groups of data.

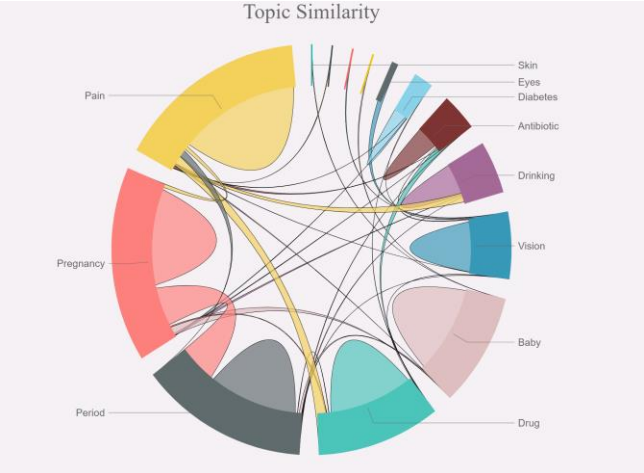

Figure 5. Chord Diagram for Similarity Distribution

The most predominant topic is obtained with the help of the chord diagram as described in the previous section. In the given dataset, the topic names for each question are identified and tagged by matching the question topic ID and the common topic ID. It is noticed that some questions belong to more than one topic. After tagging the topic names for each question, a topic name-question matrix is created. The matrix contains only binary values (1 for question belonging to a topic and 0 otherwise). The similarity matrix is computed by multiplying the topic name-question matrix with its transpose. From the similarity matrix, the top ten highest topic-topic values and connections (similarity with other topics) are extracted and viewed using the chord diagram. When the cursor hovers over ach topic, the topic's label and the topics it has an association with are displayed. In the given figure 5, pregnancy has a strong connection with period, baby and pain although it has a very flimsy connection with vision. This observation is because, when few questions are parsed, they belonged to multiple topics have are not related to each other.

Research Question Addressed

- How to determine which is the most predominant topic being discussed among the users and its similarity among other topics?

3.1.5 Text Analytics on the Question-Answer WebMD Data

Text analytics methods are used to get the most relevant questions and their answers based on the searched topic. The topic the user types in the search bar is stemmed to get the root word of its main topic. The questions in the dataset are lemmatized to remove stop words such as am, the, is, a, etc.[13] The word set obtained after lemmatization is then stemmed to get the desired root words for which the root word of the searched topic is matched against them. If a match is found, it is added along with its corresponding answer in the list of questions and answers to be displayed After the list of questions and answers have been identified, the same is sorted based on the number of votes of the question and the expert level of the user who has answered the question. The top five related questions and answers are displayed. On hovering over each question, the details of the person who answered it are displayed. When the user types a topic in the search bar, the mechanism is similar and the top five questions belonging to all the topics typed in the search bar are displayed.

Research question addressed:

- How to suggest the most relevant questions and answers to users based on the topic selected?

3.2. Overview of the System

The dashboard (Figure 6) shows the overall view which has the two solutions we came up for the WebMD forum data:

- The first being the overall view which gives the user a visual depiction of the problems and helps them predict any upcoming problems they might face in future
- We display all four of our visualizations after clicking the overall view which follows the display on demand mantra. These include all the implementation of various charts helping us answer all the relevant research questions
- Coming back to our overall view, helps us navigate to the Question and answer thread which our second solution to the WebMD forum
- We have provided a Question and answer thread view, which allows the user to filter the questions and answers with respect to a specific topic. We have performed extensive textual analysis and provide the most relevant answers which have been processed and displayed to the user.

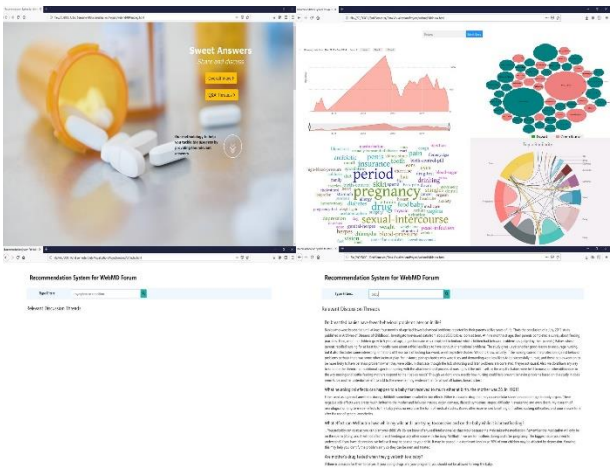

Figure 6. Overview of Entire System

4. METHODOLOGY

4.1.1. Data Collection and Pre-Processing

The data was collected from the WebMD forum. All the datasets are initially JSON files. The JSON files were then converted into .CSV files [14]. The missing values were filled based on the trends observed in the data. In some cases, the dates of the posted questions were missing, hence the latest date was used as the default value for imputation. Every question had a questionID, Title, and TopicID. For some questions the TopicID was missing, hence we discarded those without a TopicID.

4.2 Technology used

Text Analytics: NLTK, Gensim

Charts: Bubble, Chord, Word Cloud, Time Series

Web Framework: Django

Visual Analytics: D3.js, Power BI

5. EVALUATION PLAN

More charts less Words

It has more visual elements (charts) when compared to text and conveys the user the story in a more efficient way.

The concept of visual seeking mantra is applied to the visualizations. The first page of the dashboard gives an overview first, then we let the users to zoom and filter for the options by choosing the top 10 topics from the drop-down menu and then the time series allows us to zoom and filter for the topics over the required time frame.

Details on demand measure is implemented in the Q and A thread section where it enables the user to search for a topic and displays the relevant Questions and Answers to those topics.

Intuitive

The Chord diagram is chosen for c the similarly between the topics over heap map as it precisely defines the relationship between topics with just viewing at once. The thickness between the topics determine the amount of similarly between each of these topics. From this visualization we can clearly understand that people who talk about pregnancy also talk about periods. And this emerges as the one with highest frequency among the topics.

Interactive and Intelligent

The text analytics concepts like topic modeling, stemming, stop words removal, building document term matrix and identifying the most frequent topics are implemented and word cloud is used to give an overall view of the words related to a particular topic.

Time Series Modeling

The time series modeling concept is used to view the trend of the most probable topic over time. There are two views provided, a yearly and a monthly view. It also allows user to zoom and slide over the years.

Data Ink Ratio

The data ink ratio concept is wisely used to display the required information in a simple and a clear manner covering all the features.

Simple, clear and easy to understand

Easy to understand and helps to convey the entire idea with lesser and powerful choice of visualizations.

6. DISCUSSION & FUTURE WORK

Thus, in our project we recommend the relevant questions and answers to the user based on the topic chosen by them. This will benefit the user by picking out the relevant answers that might help the user with fewer number of interactions. It is also very easy to understand from the naive user point of view. The time series visualization aids the user in analyzing the most probable topics across time. The user can also predict the most influential person in answering the questions pertaining to that topic and get to know the member information. From the system, we can make the following observations.

- Pregnancy is the most frequently discussed topic followed by period, sexual inter-course and drug related topics.
- People who talk about pregnancy also talk about periods, sexual inter-course, pain and out of these topics, periods and pregnancy emerge with the highest similarity among them
- On an overall view, experts answer 56% of the questions whereas the contributors answer 44%. People answering the maximum number of questions belong predominantly to Nursing, followed by American Pharmacists Association, New York University, St. John's University College of Pharmacy, Health Sciences & Beverly Hills Plastic Surgery & Skin Care institutions and Health Coach, WebMD
- The trend of topics discussed increases gradually over the years and is highest between 2012 – 2013 (Max for Oct 2012) and contributes to 56% of the total topics discussed over time.

More problems related to mental health and body fitness can be added to the database so that the system will be able to provide relevant solutions in those fields as well. Food and nutrition-based data can be incorporated as well. A network visualization of a topic and its connections to the other similar topics can be implemented and viewed. The search query can also be less restrictive in nature when questions belonging to not all the topics typed in the search bar are displayed as well.

Thus, we have incorporated the visual seeking mantra, “Overview First, Zoom and Filter” [15] in the visualization system to achieve an effective visualization that serves the users to the fullest.
We can utilize word2Vec, Machine learning models and neural networks for building a complex model. The advanced text analytics concepts can be used to extract and provide more efficient results.

7. REFERENCES

[1] Edward Tufte, "The Visual Display of Quantitative Information"(2001)..

[2] Vagelis Hristidis et al. A Study of the Demographics of Web-Based Health-Related Social Media Users. Journal of Medical Internet Research, August 2015.

[3] https://www.webmd.com/

[4] https://support.office.com/en-ie/article/present-your-data-in-a-bubble-chart-424d7bda-93e8-4983-9b51-c766f3e330d9

[5] Christopher G. Healey, "Choosing Effective Colours for Data Visualization" IEEE March 2009

[6] Michael Correll, Danielle Albers,Michael Gleicher,Steve Franconeri, "Comparing Averages in Time Series Data".

[7] Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala, "Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations"

[8] https://doc.arcgis.com/en/insights/enterprise/latest/create/time-series.htm

[9] James Walker, Rita Borgo, Mark. W. Jones, "Time Notes: A study on Effective Chart Visualization and Interaction Techniques for Time-Series Data" IEEE 2018. https://ieeexplore-ieee-org.ezproxy1.lib.asu.edu/document/7192735

[10] R. Brath; M. Peters; R. Senior,"Visualization for communication: the importance of aesthetic sizzle", IEEE 2005. https://ieeexplore-ieee-org.ezproxy1.lib.asu.edu/document/1509153

[11] Florian Heimerl, Steffen Lohmann, Simon Lange, Thomas Ertl, “Word Cloud Explorer: Text Analytics Based on Word Clouds”. IEEE 2017. https://ieeexplore-ieee-org.ezproxy1.lib.asu.edu/document/6758829.

[12] https://datavizcatalogue.com/methods/chord_diagram.html

[13] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O’Reilly Media Inc.

[14] http://convertcsv.com/json-to-csv.htm

[15] J.J. Thomas; K.A. Cook, "A visual analytics agenda", Vol. 26 IEEE issue 1. https://ieeexplore.ieee.org/document/1573625