Intelligent Visual Analytics Dashboard - *We*Care

Priyekant Aghi
School of Computing, Informatics &
Decision Systems Engineering,
Arizona State University, Tempe, AZ, USA
paghi@asu.edu

Sagar Patni
School of Computing, Informatics &
Decision Systems Engineering,
Arizona State University, Tempe, AZ, USA
shpatni@asu.edu

Chandni Shrivastava
School of Computing, Informatics &
Decision Systems Engineering,
Arizona State University, Tempe, AZ, USA
cshrivas@asu.edu

Kshitij Khandelwal
School of Computing, Informatics &
Decision Systems Engineering,
Arizona State University, Tempe, AZ, USA
kkhande1@asu.edu

1. INTRODUCTION

The concept of using pictures to understand data has been around for centuries, from maps and graphs in the 17th century to the invention of the pie chart in the early 1800s. Several decades later, one of the most cited examples of statistical graphics occurred when Charles Minard mapped Napoleon's invasion of Russia. The map depicted the size of the army as well as the path of Napoleon's retreat from Moscow – and tied that information to temperature and time scales for a more in-depth understanding of the event. We define Data Visualization as the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. Basically, huge chunks of data is graphically visualized so as to make the data look presentable , and can be easily interpreted by a layman.

In this project, we have proposed to visually present analytics of the dataset collected from WebMD over a period between 2010 and 2014. We will be doing this using visualization tools and techniques to show some useful statistics.

2. MOTIVATION

Data visualization is a very important technique in today's world. The data collected from real-time resources is usually very huge. To make sense out of it, we make use of different types of visualization tools and present them in various graphical and structured layouts using colors and pattern. The motivation behind this is to make people understand what the data is trying to say, and to find patterns (be it usual or unusual), so that a layman can understand the data easily without any problems. It also helps the decision-makers analyze the data pictorially.

This project is a part of the Data Visualization course we undertook. Our primary goal of implementing this project was to use our knowledge and understanding gained from the course lectures, and incorporate them in a real-life scenario. By undergoing each phase of the project, we learnt the importance of data visualization and got hands on experience with data collection, exploration, manipulation and representation.

We developed an intelligent and interactive system, ***We*Care**, which is a dashboard presenting analytics from the WebMD dataset. It provides information about similar health topics, relevant workplaces and departments associated with them, making use of simple interactive visualizations that can be easily interpreted by any user.

The main idea behind this implementation is to serve the people residing in remote and less developed regions where there is scarcity of healthcare resources [1]. A person can make use of this dashboard to know more about a particular health condition he is concerned about by easily accessing answers to questions related to the topic of his/her interest. It can provide the users with contacts and resources to appropriate information and services in a timely manner.

From research point of view as well, this tool is informative to researchers and analysts by providing them insights into various health conditions that were prevalent over a period of time. This can be considered as a medium to connect the general population with those in the world of medicine.

3. VISUALIZATION DESIGN

WeCare dashboard is designed with a simple and clear layout that is user-friendly and helps people with different levels of expertise to be able to interpret data easily. The components of the dashboard are selected in such a way that it serves both the groups of people - common man and researchers.

The entire layout follows a color scheme in the shades of blue and white. This was inspired primarily from the official website of WebMD itself, keeping in mind the association or extension our idea has with the original WebMD dataset. While blue is the color of trust and intelligence, white symbolizes purity and cleanliness [8]. This is another reason for our motivation to choose these hues as our dashboard background.

The intelligent and interactive dashboard comprises of six components:

Tile 1 -

- This component **[Figure 1]** consists of a table showing a list of all the topics collected from the dataset - be it a disease, a symptom, a drug etc.
- The topics and their counts are listed in the rows of the table that enables the user to search through any/all the topics and select the one that he/she is interested in.
- When the page reloads, the topics are listed in ascending order alphabetically. But depending on the nature of usage, the user has the option to sort the data in ascending/descending order alphabetically according to topic name or the count of questions.
- The body of the table is kept visually simple with shades of grey in white for easier readability of the contents to the user. [6]

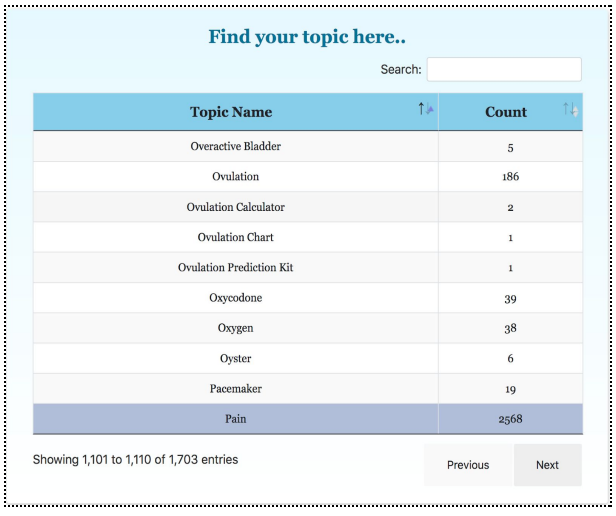

Figure 1 - List of all the topics to select from

Tile 2 -

- The next component **[Figure 2]** shows a bar chart with top 10 topics (in general) over a selected period of time.
- The user gets an option to select the time period and list out the most discussed topics (in descending order) in that timeframe. That can be narrowed down even further if the user selects a particular topic from the list.
- Bar chart presents a clear distinction between different topics/categories by quantifying the relative proportions and summarizing the statistics of a large data set in a concise manner.
- This is an interactive component of our dashboard from which a user can select a bar signifying one of the top trending topics and get more information about it.
- The bars have been enhanced by highlighting their selection as the cursor hovers over them. This is helpful for the user to ensure he has chosen the correct topic bar.
- The bar chart also helps the user to get a quick glimpse of the trending topics over a period of time.
- This visualization is particularly helpful for the researchers in the field of medicine. The visualization lets them easily study the top trending topics in a given time frame.
- Using this visualization makes it easier for them to collect the statistical data about the number of questions that were answered for a particular topic in the selected time range.

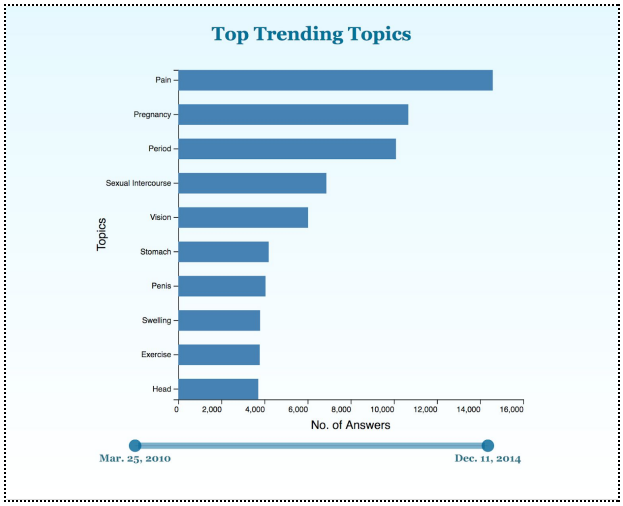

Figure 2 - Bar chart showing the top trending topics in the selected time period

Rest of the components are connected to visuals from first two tiles.

Tile 3 -

- We have used radar chart **[Figure 3]** as our intelligent component to show the similarities between topics [10].

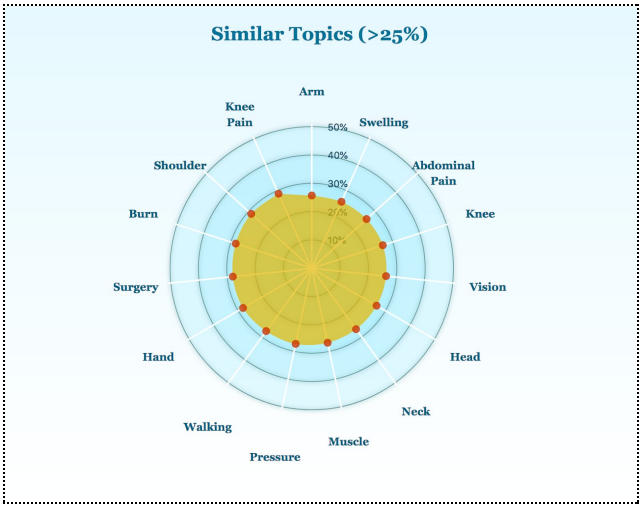

Figure 3 - Radar chart showing the similar topics to the currently selected topic 'pain'

- A radar chart uses a radial (circular) display with several different quantitative axes emerging like spokes on a wheel to create a unique shape of quantitative values.
- We used this to show what other topics might be similar to the selected topic. As we go radially outwards, the similarity increases and vice-versa.
- For similarity we chose 25% as the threshold value to select the topics to display in the radar chart.
- The user can also select one of the similar topics from the radar chart to see the effect on the rest of the components.
- Each axis of the radar/spider chart represents one similar topic. And the position of the red circle on the axis shows the quantitative value of the similarity in percentage. A bright red dot will be easy to spot in a complex figure, hence we chose it to mark the similarity percent points.
- Also showcasing the density of similarity, is the yellow web in the center. The color yellow has often been used to highlight things and draw user's focus to it. Even without knowing the actual percentage of similarity between topics, a person can get an overview by the area of the web.
- The radar chart makes it easier for the user to get the idea about relative order of similarity with the selected topic. If the user is interested in knowing the exact number, he/she can hover over the dots to display the exact similarity value.
- We could have shown the similarity values directly along with the red circles but it would have made the radar chart more congested and unpleasant to look at. We know that cramming the visualization space with too many details at once is not a good representation of data as it leads to difficulty in interpreting useful information.
- Furthermore, if the user does not find the answer he/she is looking for from the selected topic, he/she can look through the relevant questions of one of the similar topics displayed on the radar chart.
- This can be achieved by clicking on one of the red dots on the axes of the chart. This will transform all the visualizations including the radar chart to display information related to the newly selected topic.

Tile 5 & 6 -

- Furthermore, a user can get to know which (top)groups have answered the questions to the selected topic.
- We have two sections showing this - one is by workplaces **[Figure 4]**, and the other is by department or job ids **[Figure 5]**. They are visualized using bubble charts. Bubble charts can show the contribution of a component by the size of the bubbles, so the user can clearly perceive from the visualization the relative number of questions attended by an institute/workplace or department.
- We have used bright color schemes to represent both the bubble charts. Vibrant colors of the bubbles along with their size give a clear distinction among the objects and their contribution [9]. This may also prove helpful for people with low level of expertise.
- Like radar chart, we have refrained from showing answered questions statistics here as well. Because of differences in size of bubbles, displaying both department names and number of question they answered together does not sit well with a good visualization structure. It will overcrowd the space with texts and scramble the focus of the viewer making it less appealing.
- This visualization aspect of the bubble chart is helpful for people living in the remote areas in one more way - when they are not completely satisfied be answers already given to questions and seek expert advice for their problem. It is obvious that they would want to get the medical attention from the most experienced person. But due to lack of resources in the remote areas, it is possible that the most relevant workplace/department has a long waiting time for appointment.
- In this case our visualization comes to rescue by showing the less relevant alternate options so that they can get immediate assistance from them and visit an expert later to continue their treatment.

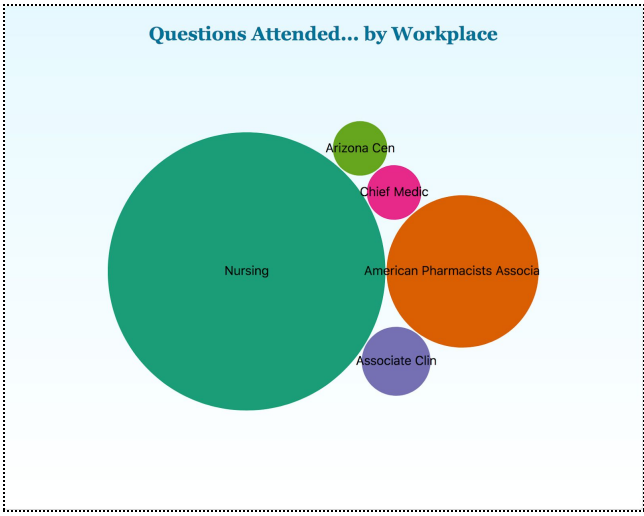

Figure 4 - Bubble chart showing the workplaces which answered questions related to currently selected topic 'pain'

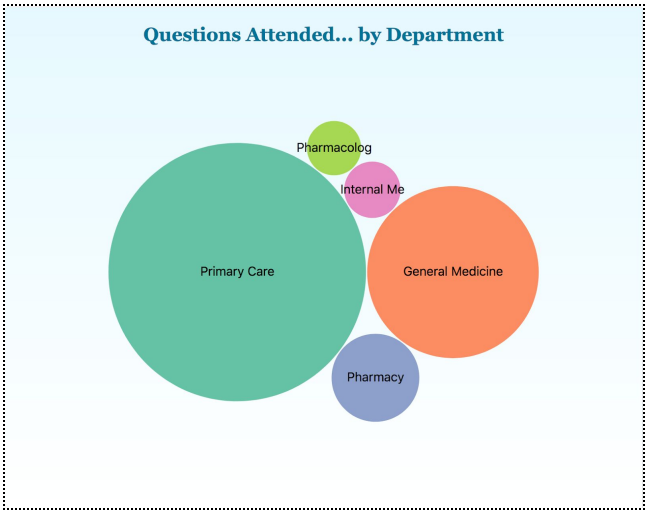

Figure 5 - Bubble chart showing the departments which answered questions related to currently selected topic 'pain'

- This visualization is also helpful for the researchers in the sense that they can get an idea about which institutes/departments they should contact in order to get more information regarding their research topic.
- As seen from the figures, the two bubble charts have different color schemes [5]. We have kept this to visualize workplace and department stats distinctly.

Tile 4 -

- Finally, we display a list of all the top useful questions **[Figure 6]** related to the selected topic.
- Like the table in the first tile, the table here too presents a clean and simple layout with greys and white. The user can skim through all the listed questions, find relevant ones and select them to refer to the answers.[6] It will take them to the specific question thread on actual WebMD portal.

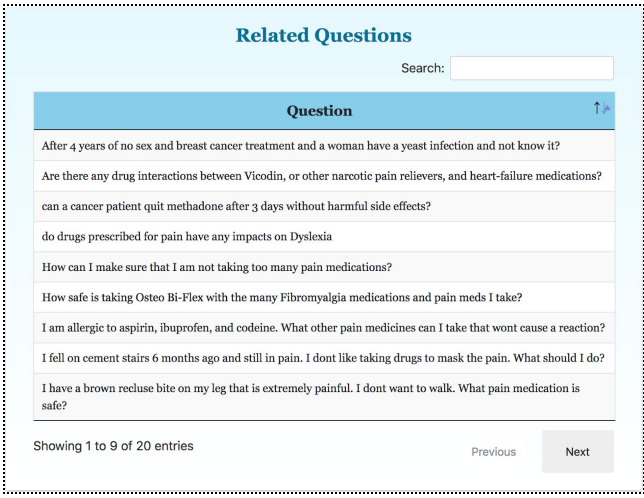

Figure 6 - List of questions related to the selected topic

4. METHODOLOGY

We have used several methods and concepts taught in the class. The technologies used include D3.js, HTML, CSS, Python, Django etc.

The visualization dashboard aims at answering the following research questions:

- What are the most trending topics over a particular period of time?
- What are the topics similar to the one's selected by the user?
- Which workplaces are answering the most number of questions related to a topic?
- Which departments are answering the most number of questions related to a topic?
- Which already answered questions might be useful for the user ?

The first step in the process of developing WeCare was data processing. We were provided with data collected from the WebMD website. The data included important attributes like users, questions, answers, workplaces, departments, date, upvotes etc. The details are as follows:

- Time period (Mar. 2010 - Dec. 2014)
- 1704 Topics
- 5402 Users
- 25319 Questions
- 74101 Answers
- 69 Workplaces
- 58 Departments

To process the data we combined all the data files into a single dataframe using 'questionId' as a key column. Combining all the data gave us the advantage to directly

find the characteristics of the dataset without using complex data operations.

The WebMD dataset as a whole has a lot of missing values. We ignored missing values at different stages of data processing as most of the missing values were categorical, for which default values cannot be added.

The intelligent algorithm which finds similar topics to the selected topic make use of the topic-id and related_topics. We have used the following algorithms:

TF-IDF : It stands for *term frequency-inverse document frequency*, and the tf-idf weight is a weight often used in information retrieval and text mining. [3]This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus, in our case we had a dataset of questions and answers and we found out the tf-weight for some of the words being repeated multiple times in the entire dataset. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

[4]One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated ranking functions are variants of this simple model. Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

Cosine similarity : In practice, cosine similarity tends to be useful when trying to determine how similar two texts/documents are. It is used for sentiment analysis, translation.Cosine similarity works in these use-cases because we ignore magnitude and focus solely on orientation.

In mathematics perspective, Cosine similarity is perfect. However, if we check it in text mining perspective, it may not always be reasonable.

Given two N-dimensional vectors A and B , the cosine similarity between them is calculated as follows:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 4 - Cosine Similarity [2]

For our study, we had dataset consisting documents which contained several questions and answers provided, with questionid and topic id.

We created a vector using the topics and related topics data. In this vector, each row and columns represents a topic and the row value represents the number of times this topic appeared in the topic corresponding to a row. Using this vector we calculated TF-IDF for each topic. The rationale behind TF-IDF is that we wanted to calculate the topics which have distinctive similarity with each other. We used euclidean distance with cosine similarity as distance metric. To find final similarity we performed 1 - distance and sorted the data according to distance to get the top similar topics.

5. EVALUATION PLAN

The project scheduled to begin in January 2018 was concluded in April 2018. We performed both formative and summative evaluation of our project [11]. For formative evaluation we concentrated on the planning and implementation stages of our project. Whereas for summative evaluation we focused on the outcomes of the project. Upon completion of each stage of the project we reviewed the progress and determined if we have successfully met all the requirements set for that stage of the project.

5.1 Project Planning

In this phase, we started with understanding the collected data. After a detailed and careful analysis of the data set, we did some research on how this data can be used with intelligence to present analytics of some usual health related facts in a pictorial format. Based on our research and ideas, we defined our problem statements and determined who our target users would be. We came up with a set of research questions that we aimed to find a solution for through our project. Based on these questions, we designed our visualization layout and components.

Once a cumulative structure of the design was decided, we went ahead to formulate different phases of the project in which we would proceed with implementation of the components one-by-one.

5.2 Project Implementation

In the project implementation phase, we started by building all the individual components of the visualization tool determined in the planning phase. Once all the components were implemented separately we started integrating different components and put them in a single dashboard. During the implementation phase we faced some challenges while integrating the components and connecting them with each other. The individual components developed initially were tested with dummy

data and worked without any issues with our data dump. However, it showed discrepancy while integrating the modules together. The input data was found to be in different forms across all the components. We had to take an additional step to process the data and transform it in a form suitable for different components. After successfully integrating the components together we came to a conclusion that some of the design decisions taken in the planning phase are not accurately answering all the research questions determined then. Therefore, we further explored a few more visualizations to incorporate in our final visualization design that could not only answer all of our questions but also looked appropriate from users' point of view.

5.3 Outcome Evaluation

In this final phase, we tested and evaluated whether all the components of the application dashboard were working smoothly. To verify the design from the user's perspective (common man or researcher), we took the help of our friends and classmates to review our application in terms of all the functional and non-functional aspects. After detail exploration of the dashboard by many peers, we found that our design met almost all the expectations of a first time user. One question that arose from a few users was the option to list departments and workplaces in nearby geographical locations of the user. We have addressed this concern in our discussions & future work section of this paper.

6. RESULTS

The data included important attributes like users, questions, answers, workplaces, departments, date, upvotes etc. The details are as follows:

- Time period (Mar. 2010 - Dec. 2014)
- 1704 Topics
- 5402 Users
- 25319 Questions
- 74101 Answers
- 69 Workplaces
- 58 Departments

We found out some important statistics from the given data set and were able to answer some very important questions:

- First, the question arises of how a person can find out the information about a particular topic; our tool makes it easy for a person to find that particular information while navigating through the dashboard, and he/she can also find topics relevant to the information they were looking for.

- By giving the access for this, the said person can find any desired information easily and swiftly, hence by making use of our visualizations and the search bar, it is very easy for a layman to search and find the required information of a particular disease, or what other people think about the solution to that particular disease

- Second, our tool could reach out to millions of people living in the rural areas. In regard to the research question of how people living in these areas would be benefited, our tool gives the people living in small towns/villages the ease of using and finding their desired information without spending any money.

7. DISCUSSIONS & FUTURE WORK

As just a prototype, we tried to build the interface as simple as possible, so it is easy to comprehend by other people and the user interface can be easily comprehended for searching the required information, as easily and swiftly as possible. For future purposes, if given more time and resources, several features can be added to the current interface.

Since ***WeCare*** aims to cater primarily to the users in the remote and rural areas with limited medical resources, we believe the dataset can be aggregated with geographical information of the workplaces and the people who are asking the questions.

We plan on keeping a map as the central component of our visualization, giving user the ability to select a state and then further selecting the city. Including this information will help in making our system providing information that is more relevant for the user. It would be more efficient if the dashboard navigation is governed by the particular geographic location. We also plan on including more workplaces in our dataset as currently only 69 workplaces are present in the dataset.

By doing this we can further display the top trending topics in the selected time period for the selected region which might be more useful for the user. The same is the case with the workplaces data. If the user is using our tool to get the information about the institutions he/she can visit then it makes sense to display the top workplaces based on the location of the user.

Including this information can also be useful for the researchers using our tool giving them the option to study the variation over time as well as geographical region.

Many different adaptations, tests, and experiments have been left for the future due to lack of time (i.e. the experiments with real time data are usually very time consuming, requiring even months to finish a single run). Future work concerns deeper analysis of particular mechanisms, new proposals to try different methods, or in our case simply curiosity.

The use of other types of individual representations and visualization functions with different types of designs and patterns could be investigated since they have an important influence on the results obtained at the end. New approaches in this direction can be induced from techniques described in the literatures and research done by different researchers..

8. ACKNOWLEDGEMENT

We thank our professor Dr. Sharon Hsiao for giving us the project to have hands on experience and making us realize the importance of data visualization. We also thank our teaching assistant Yihan Lu for giving prompt replies, solving our doubts and being available for all the questions we had.

9. REFERENCES

[1] https://www.ruralhealthinfo.org/topics/healthcare-access

[2] https://en.wikipedia.org/wiki/Cosine_similarity

[3] G. Salton and C. Buckley. "Term-weighting approaches in automatic text retrieval". Information Processing & Management, 24 (5). 1988.

[4] H. Wu and R. Luk and K. Wong and K. Kwok. "Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems, 26 (3). 2008.

[5]https://blog.graphiq.com/finding-the-right-color-palettes-for-data-visualizations-fcd4e707a283

[6]https://www.cmu.edu/gcc/handouts/handouts-new/data_viz_handout_gcc.pdf

[7]https://pdfs.semanticscholar.org/41ff/3934f40c32ac8643270822de1c763e16c71b.pdf

[8]https://optimized360.com/blogs/selecting-colors-for-your-medical-or-dental-website/

[9]https://visage.co/data-visualization-101-bubble-charts/

[10]http://bl.ocks.org/nbremer/21746a9668ffdf6d8242

[11]https://www.thehealthcompass.org/how-to-guides/how-develop-monitoring-and-evaluation-plan