Semantic Visual Analytics of WebMD

CSE 591: Topic: Data Visualization Report

Aditya Gnaneshwar, Rakesh Prabhu, Sahana Sekhar PC
MCS, CIDSE
Ira A. Fulton Schools of Engineering
Arizona State University
Email: {agnanesh, rrprabhu, schand40}@asu.edu

Abstract—**One of the important benefits of Data Visualization is that it allows visual access to huge amounts of data. Interactive visualization turns number and letters into aesthetically pleasing visuals, making it easy to recognize patterns and find exceptions. This project deals with mining, organizing and visualizing the data obtained from the Wed-MD website. We intend to study the data, transform it and derive the not-so-obvious information and facts that could otherwise be obtained by just viewing the data. We adopt various modeling techniques such as Topic Facet Modeling, Latent Semantic Indexing and other data processing techniques in order to convert the data to a viewable format. For visualization we embrace different libraries and techniques such as D3, Word Clouds and basic HTML elements. The user, after viewing the entire dashboard with different visualizations will be acquainted with a lot of information about both the website as a whole and its consumers.**

Keywords—Data Visualization, Topic Facet Modelling, Latent Semantic Indexing, Word cloud, D3, Data Mining.

I. INTRODUCTION

WebMD is an online health information services website that provides valuable health details and health managing tools to its consumers. It's mainly a publisher of human health and well-being information, serving it' s consumers with accurate, easy to understand answers to their most pressing health related questions. The consumers of the website include physicians, healthcare professionals, employers along with health plans. Everyday thousands and thousands of consumers use the website for - posting questions seeking accurate reliable services, answering diagnosis questions after proper comprehensive understanding and providing health tips. All this accords to huge amount of data being stored on the website. This data collected on the go can be mined, analyzed and visualized in several ways to gather important generalized facts such as consumer's health life, trending topics of discussion and how over the years the data collected has changed reflecting the changes in lifestyles of the people. This project proposes a semantic visual analytic system to explore the above online Q/A forum. The visual analytics system delivers a text mining technique with interactive visualization to present a unique way of understanding the semantics of questions asked, answers posted and the consumers associated.

II. MOTIVATION

With such a vast number of discussion boards and forums constantly increasing in volume and variety of topics, the idea of analyzing the content and providing visually interactive interface gives totally a new dimension towards the data. Our goal is to provide simple yet effective Visualization in such a way that even a non-technical individual will be able to gain useful information. By analyzing the discussions based on content we believe we can associate a certain post with the user query which normally would not have been possible. Analyzing the content in the discussion forum and the coherence between them for a certain inter related topics would result into more valuable results. This would provide users with additional contents / knowledge that the user could utilize. When deciding on the visualization it was decided that topic and coherence visualization, sentiment Visualization be presented in such a way that we could see the connection between them. Since adding a user model visualization contained in the Web-MD dataset would distract the users using the web application from their goal, we decided to separate user based visualization in a separate module. With our visualization we intent that the user should not only get their required piece of information but in fact he/she should be able to also gain insight to the trends, link between various topics/questions and would get a better understanding of the users, contents and the topic associated with it.

III. PROJECT PERSPECTIVE

The entire purpose of mining data and visualizing, here, is three-fold. First we focus entirely on the topics, it's questions and answers, and judge how well the certain questions are related to the topics. We also propose a sentimental analysis model of questions and answers. All this is separately presented according to year-wise data. Second, we focus mainly on the consumers, study their activity and publish results showing the popular ones, by considering certain attributes. Lastly, we provide infographic reports after thoroughly studying the trending topics and state various facts related to those topics for better understanding.

IV. Data Selection

Due to the vast number of data present in Web MD it is not feasible to take into consideration all the data. So we did some data purification process to remove un-necessary data from the data sets. Initially there was nearly 31000 unique questions posts present on web MD dataset. We first removed all the data which had very few helpful votes, then we created 4 different groups of data based on the year the question was asked, after this we clustered the documents into different cluster based on the document similarity matrix. From each of those clusters we took into consideration the top 20 question from each from each of those question. We did the same for the user based data model. After many trials and revisions, the set containing all users that had answered more than 150 questions was chosen as the visualization data source. This was determined by a number of factors such as the number of users in the set, the number of helpful votes the user received, the number of questions answered by the user, the frequency by which the user answers a question etc.

V. Data Analysis

For each of the visualizations, the data has been processed and analyzed differently by incorporating different data mining techniques and models. This section gives an in-depth overview of how the processed data that is fed into each of the visualizations was obtained.

a) Topic Facet Modelling – Topic Value

We make use of Topic Facet Modelling to automatically extract topics from the large pool of discussion forum dataset corpus. Topic Facet Modelling is basically built on top of LDA and SLDA and it tries to extract the semantics of the text by associating facets. Also in SLDA if a particular word belongs to two different topics that it would associate the question with the topic which has more number of related topics. But this certainly has its own disadvantage as it does not take into consideration the core topic that the question was depicting. This disadvantage is taken care of in TFM, so if a question has a word 'cramp' it is more likely to be classified as "Muscle related issue" even though it has more keywords related to Topic "Pain". So this way by taking into consideration the facets it helps us to efficiently classify the questions into its topics.

We did our own research to filter down the facets that are associated to each topic. After careful observation and analysis, the number of facets range from 10-15 which we categorized as a temporary mapping (Alternative if Plan A does not work). The number of topics to be considered for our analytics was chosen by using a clustering technique on the Web MD dataset. The First step includes deriving the corpus features by making use of TFIDF and reducing it to two dimensions. We made use of the python library genism to transform the corpus features to two dimensional co-ordinate space. It transforms the corpus obtained from TFIDF to a 2D coordinate space object by making use of the Latent Semantic Indexing technique. We then do K-Means Clustering to the 2D coordinate space obtained from the previous step to get an optimal value k by changing the values of k from 1 to 13. Next we find the Inertia of the cluster all of the for the k values to find the best trade-off point which is k=8 in our case. After getting the value for k, we wanted to confirm which topics we need to select from those selected keywords from clusters. Finally, we came up with the following Facets Heart related, Women Related, Kidney and Lungs related, Pain related, Blood related, Heart Related, Infection related, Diabetes related, Muscle Related.

b) Topic based Visulaization – Sentiment Value

Sentiment Analysis in the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral. There are two kinds of Sentiment Analysis – Subjectivity and Polarity. Subjectivity refers to how someone's judgment is shaped by personal opinions and feelings instead of outside influences. In this project we have employed the subjectivity module in order to study the sentiment analysis of the answers posted. Subjectivity is a value that is between 0 and 1. To perform the sentiment analysis we have made use of the Text Blob package available in Python. We study this sentiment analysis to see how productive an opinion of a contributor is. This will help gain the confidence of the reader. A value of 0 indicated that the response is objective and a response of 1 indicates that the response is subjective. After this value is calculated for all the selected questions, a csv file is generated containing the question ID, the sentiment value and the color associated to it. The question ID with the lowest sentiment value is given the darkest color indicating that the response is highly subjective. Lightest color indicates that the response is objective.

c) User based Visualization – Top 8 users.

For the user model the data is obtained for the top 8 contributors in the following categories: Number of Topics Addressed, Number of Helpful Votes gathered, Number of followers and Number of questions answered. The members JSON file was parsed to extract the information for each user based on the specific properties. A minimum threshold of 200 was set to be in the top 8. On this refined information, the list was sorted in descending order to obtain the top 8 contributors. The total number of each property is then divided by the total participants in this group of top 8 and multiplied by 10. This is done so as to score each user on a scale of 10.

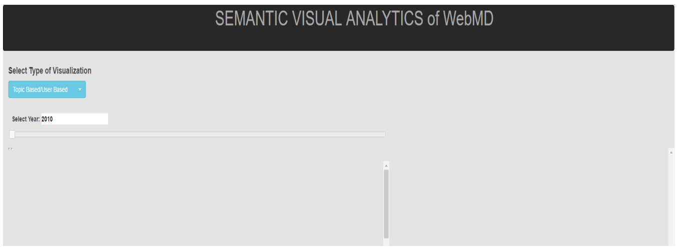
d) Infographic Reports

These reports give details about the four trending topics of discussion in the forum. Each report contains a graph showing the trends in the no of questions over the years. The data for this was obtained by parsing the WebMD wesbite using Selenium and Phantom.js to obtain the dates for all the questions posted. For all the questions that that did not have a date in the questions.json file, the correct date was appended to it after crawling the website. The final json file was then cleaned and processed to obtain a csv file containing the topic name and respective years. This data was then filtered to obtain the graph for each topic inidivually. Excel was used to plot the graphs.

Next was the word cloud display. To obtain the word cloud of all the answers related to a articular topic, for the infographic report we used the nltk library in python to eliminate the stop words. The resulting data was further refined using the technique of tokenization to eliminate the numbers and gather only the important keywords. This set was sorted in descending order to obtain the most occurring words in the given set. The data collection is based on the trending topics of interest. Using the above two visual analytics, conclusion was drawn and stated.

VI. Design Rationals

The visualizations proposed are divided into two separate categories. Topic–based and User-based.



SEMANTIC VISUAL ANALYTICS of WebMD

Select Type of Visualization

Topic Based/User Based

Select Year: 2010

Figure 1. Dashboard

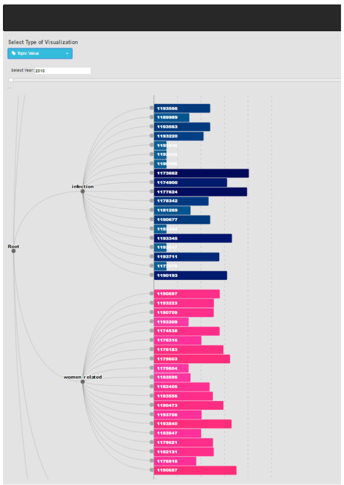
The dashboard designed provides an interface to the user to toggle between different visualizations, by selecting the appropriate option. It also consists of a slider, where the user is allowed to select the year and the visualizations are implemented in such a way that they dynamically change and load data for that year and present the model. The entire screen is divided into two sections, one for the visualizations and the other for displaying relevant information as we will see in below descriptions.

VII. Visualizations

A. Topic Based

This visualization basically maps the question to its eight generalized topics and quantifies the association based on two values – the topic value and sentiment value.

1) Topic Based Visualization- Topic Value



Select Type of Visualization

Topic Value

Select Year: 2010

Figure 2. Topic Value based Tree Visualization

*a) Choice of Visualization:*We have chosen a tree structure to depict the mapping of questions and topics . There is a main root node called "Root" that branches into eight generalized topics. Each of these topics later branch out into question ID's belonging to the respective topics. There is an estimation bar just above rating the relevance/sentiment depending on the value associated with every leaf-node.

b) Data: The data for this visualization is a csv file containing the id (Topic), value (topic value) and a color associated to every rectangular bar.

c) Size: As it can be seen from the visualization, the Root node the topics, and the question IDs are each represented by a circle and each leaf node is a rectangular colored bar of a definite size. The size of each of the bar containing the ID's represent the value associated with that question.

d) Color: Each of the topics is given one categorical color and each of the child nodes of these topics (topic ID's) are colored by varying the saturation level of the assigned categorical color. Under each topic, the rectangular bar with the highest value is given the darker color.

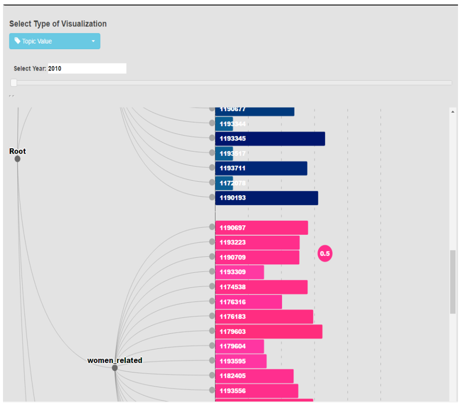
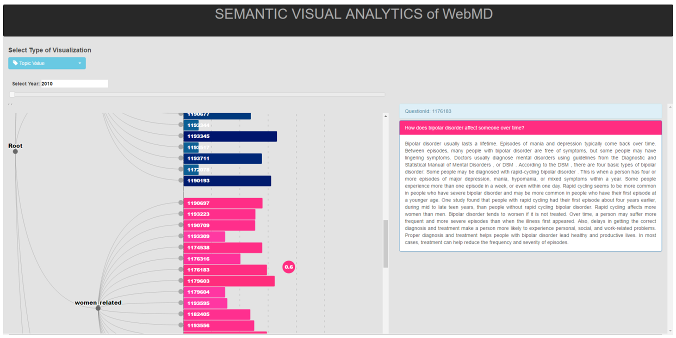


Figure 3. Interactivity showing the topic value.

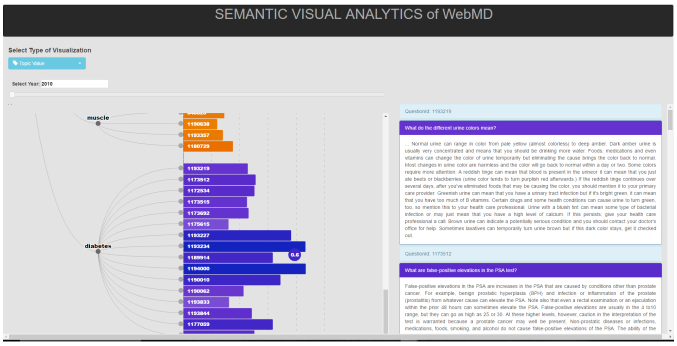

SEMANTIC VISUAL ANALYTICS of WebMD
Select Type of Visualization
Topic Value
Select Year: 2010
Root
women_related
0.5

QuestionId: 1176183

How does bipolar disorder affect someone over time?

Bipolar disorder usually lasts a lifetime. Episodes of mania and depression typically come back over time. Between episodes, many people with bipolar disorder are free of symptoms, but some people may have lingering symptoms. Doctors usually diagnose mental disorders using guidelines from the Diagnostic and Statistical Manual of Mental Disorders , or DSM . According to the DSM , there are four basic types of bipolar disorder: Some people may be diagnosed with rapid-cycling bipolar disorder . This is when a person has four or more episodes of major depression, mania, hypomania, or mixed symptoms within a year. Some people experience more than one episode in a week, or even within one day. Rapid cycling seems to be more common in people who have severe bipolar disorder and may be more common in people who have their first episode at a younger age. One study found that people with rapid cycling had their first episode about four years earlier, during mid to late teen years, than people without rapid cycling bipolar disorder. Rapid cycling affects more women than men. Bipolar disorder tends to worsen if it is not treated. Over time, a person may suffer more frequent and more severe episodes than when the illness first appeared. Also, delays in getting the correct diagnosis and treatment make a person more likely to experience personal, social, and work-related problems. Proper diagnosis and treatment helps people with bipolar disorder lead healthy and productive lives. In most cases, treatment can help reduce the frequency and severity of episodes.

Figure 4. Clicking on question ID displays the respective question and answer


SEMANTIC VISUAL ANALYTICS of WebMD
Select Type of Visualization
Topic Value
Select Year: 2010
muscle
diabetes
0.6

QuestionId: 1193219

What do the different urine colors mean?

... Normal urine can range in color from pale yellow (almost colorless) to deep amber. Dark amber urine is usually very concentrated and means that you should be drinking more water. Foods, medications and even vitamins can change the color of urine temporarily but eliminating the cause brings the color back to normal. Most changes in urine color are harmless and the color will go back to normal within a day or two. Some colors require more attention: A reddish tinge can mean that blood is present in the urineor it can mean that you just ate beets or blackberries (urine color tends to turn purplish red afterwards.) If the reddish tinge continues over several days, after you've eliminated foods that may be causing the color, you should mention it to your primary care provider. Greenish urine can mean that you have a urinary tract infection but if it's bright green, it can mean that you have too much of B vitamins. Certain drugs and some health conditions can cause urine to turn green, too, so mention this to your health care professional. Urine with a bluish tint can mean some type of bacterial infection or may just mean that you have a high level of calcium. If this persists, give your health care professional a call. Brown urine can indicate a potentially serious condition and you should contact your doctor's office for help. Sometimes laxatives can temporarily turn urine brown but if this dark color stays, get it checked out.

QuestionId: 1173512

What are false-positive elevations in the PSA test?

False-positive elevations in the PSA are increases in the PSA that are caused by conditions other than prostate cancer. For example, benign prostatic hyperplasia (BPH) and infection or inflammation of the prostate (prostatitis) from whatever cause can elevate the PSA. Note also that even a rectal examination or an ejaculation within the prior 48 hours can sometimes elevate the PSA. False-positive elevations are usually in the 4 to10 range, but they can go as high as 25 or 30. At these higher levels, however, caution in the interpretation of the test is warranted because a prostate cancer may well be present. Non-prostatic diseases or infections, medications, foods, smoking, and alcohol do not cause false-positive elevations of the PSA. The ability of the

Figure 5. Clicking on the topics, displays the question and answers for that topic

e) Interactivity: The visualization has some interactive elements incorporated. Hovering on any of the rectangular bars, displays the value associated with it, right next to it. With the help of the estimation bar above, the user will be able to easily judge as to which node has a better review in terms of relevance. The rectangular bars with just question-IDs, don't really give the user enough information about the actual question being evaluated here. To solve this, we provide the user the option to view the selected question and the answer. On clicking the rectangular bars, the corresponding question and answer is displayed next to the visualization. On clicking the topic node, all the associated, leaf-nodes question and answers gets displayed.

2) *Topic Based Visualization- Sentiment Value*

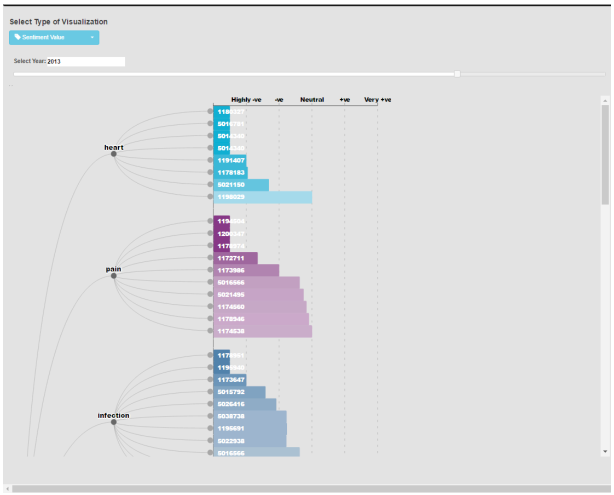


Figure 6. Sentiment Value Based Tree Visualization

a) Choice of Visualization: The choice of visualization is same the one described in the previous section.

b) Data: The data for this visualization is a csv file containing the id (Topic), value (sentiment value) and a color associated to every rectangular bar.

c) Size and Color: Here too, the same description above, holds.

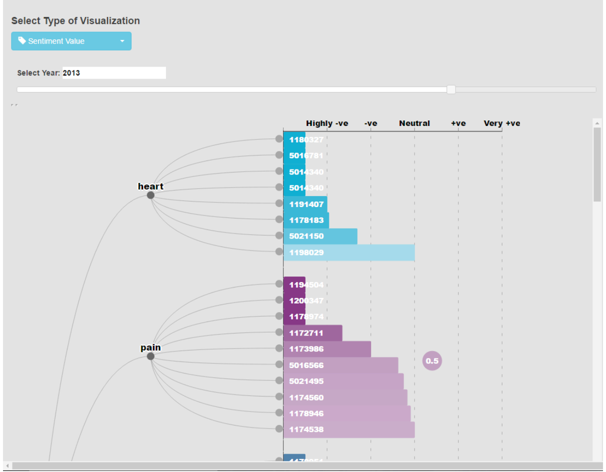
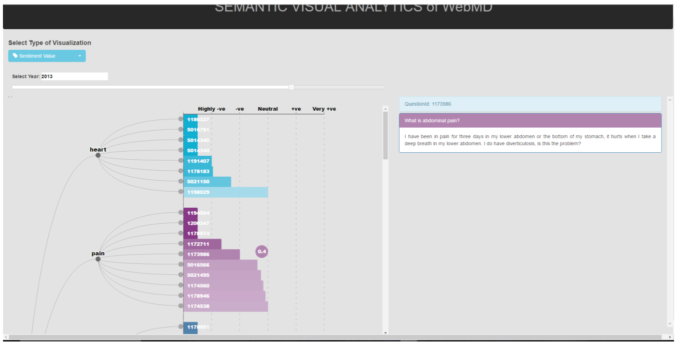


Figure 7. Interactivity showing the sentiment value.

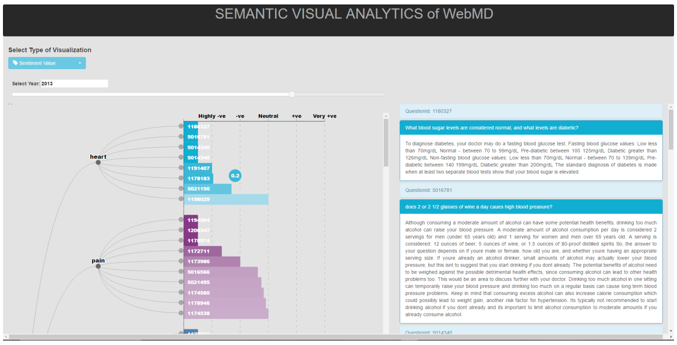


QuestionId: 1173986

What is abdominal pain?

I have been in pain for three days in my lower abdomen or the bottom of my stomach, it hurts when I take a deep breath in my lower abdomen. I do have diverticulosis, is this the problem?

Figure 8. Clicking on question ID displays the respective question and answer



QuestionId: 1180327

What blood sugar levels are considered normal, and what levels are diabetic?

To diagnose diabetes, your doctor may do a fasting blood glucose test. Fasting blood glucose values: Low less than 70mg/dL Normal - between 70 to 99mg/dL Pre-diabetic between 100 125mg/dL Diabetic greater than 126mg/dL Non-fasting blood glucose values: Low less than 70mg/dL Normal - between 70 to 139mg/dL Pre-diabetic between 140 199mg/dL Diabetic greater than 200mg/dL The standard diagnosis of diabetes is made when at least two separate blood tests show that your blood sugar is elevated.

QuestionId: 5016781

does 2 or 2 1/2 glasses of wine a day caues high blood preasure?

Although consuming a moderate amount of alcohol can have some potential health benefits, drinking too much alcohol can raise your blood pressure. A moderate amount of alcohol consumption per day is considered 2 servings for men (under 65 years old) and 1 serving for women and men over 65 years old. A serving is considered: 12 ounces of beer, 5 ounces of wine, or 1.5 ounces of 80-proof distilled spirits So, the answer to your question depends on if youre male or female, how old you are, and whether youre having an appropriate serving size. If youre already an alcohol drinker, small amounts of alcohol may actually lower your blood pressure, but this isnt to suggest that you start drinking if you dont already. The potential benefits of alcohol need to be weighed against the possible detrimental health effects, since consuming alcohol can lead to other health problems too. This would be an area to discuss further with your doctor. Drinking too much alcohol in one sitting can temporarily raise your blood pressure and drinking too much on a regular basis can cause long term blood pressure problems. Keep in mind that consuming excess alcohol can also increase calorie consumption which could possibly lead to weight gain, another risk factor for hypertension. Its typically not recommended to start drinking alcohol if you dont already and its important to limit alcohol consumption to moderate amounts if you already consume alcohol.

QuestionId: 5014340

Figure 9. Clicking on the topics, displays the question and answers for that topic

d) Interactivity: The interactive elements remain the same, as described above. This visualization helps the user understand and judge the sentiment of the answers associated with the questions.

B. *User Based*

This visualization depicts a many to many mapping between the users and the topics he/she addressed.

1) *User-based visualization – User to Topic Mapping*

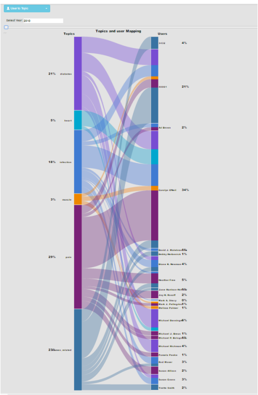


Figure 10. Many to Many user to topic mapping

a) Choice of Visualization: In order to depict the mapping between the user and the topic we have chosen a bipartite graph, which helps us show the association clearly. The visualization has two columns, one showing the topics and the other showing the users. Each topic is mapped to several users and vice versa. From the topics perspective, the visualization maps all the users who have addressed questions related to that topic and the from the users perspective, it shows all the topics that particular user is proficient in. The percentage values next to each of the labels at first depicts accurately, the relative comparison between the contribution of each user and how popular each topic is when compared to other. The values change appropriately when interactivity is introduced.

b) Data: The data fed into the visualization, is a json file containing the topic, the user and the topic value . Year wise json files are fed into the visualization, depending on the slider selection and visualization is displayed.

c) Size: Under the topic column, the size of each topic is proportional to the percentage value next to it. With respect to the user, if he/she has answered questions related to a particular topic, a color bar (color related to that topic) with the size proportional to the number of questions he has answered in that topic gets added. Hence for a user, it's a combination of different color bars of different sizes.

d) Color: We have chosen categorical colors to represent the eight generalized topics. This helps the user easily differentiate between the topics. Each user gets a color bar of a topic to this side if he/she has addressed any question on that topic, as explained above.

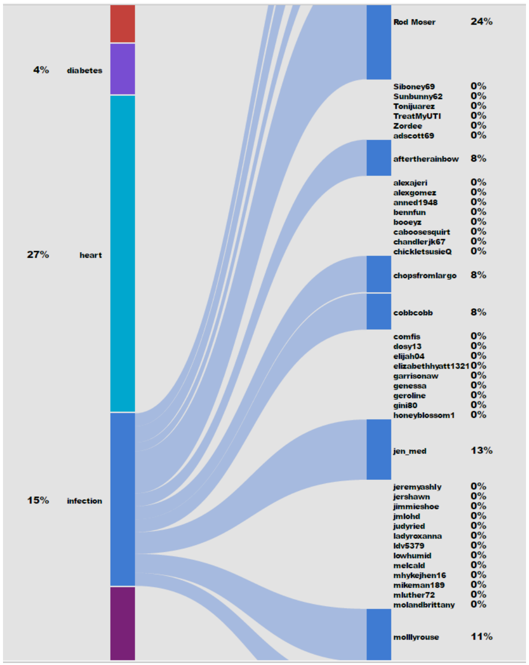


Figure 11. Hovering over the topics, displays the mapped users who have addressed questions on that topic.

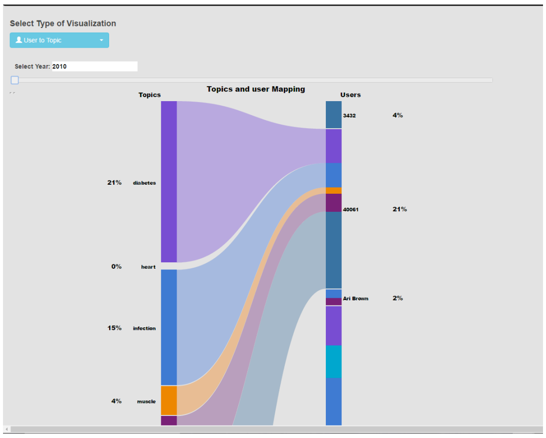
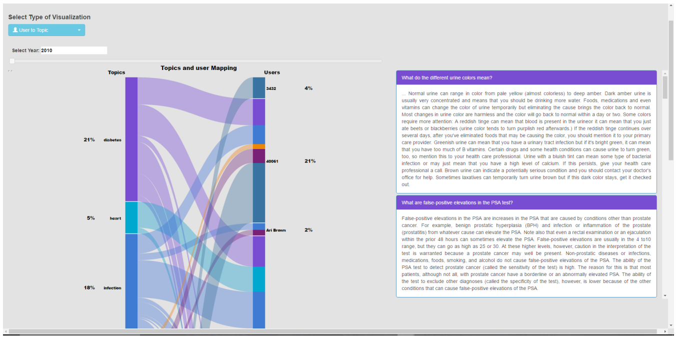


Figure 12. Hovering over the users displays all the topics he/she is proficient in.



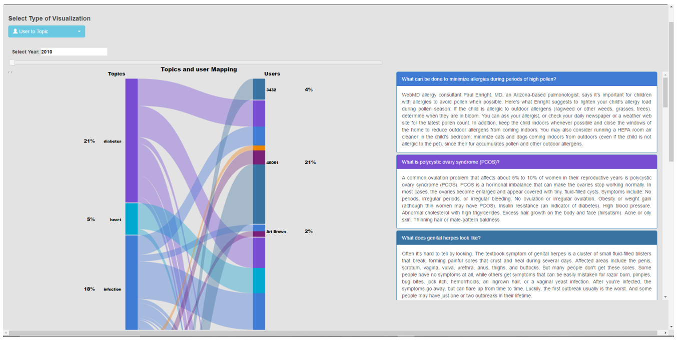
What do the different urine colors mean?

... Normal urine can range in color from pale yellow (almost colorless) to deep amber. Dark amber urine is usually very concentrated and means that you should be drinking more water. Foods, medications and even vitamins can change the color of urine temporarily but eliminating the cause brings the color back to normal. Most changes in urine color are harmless and the color will go back to normal within a day or two. Some colors require more attention: A reddish tinge can mean that blood is present in the urineor it can mean that you just ate beets or blackberries (urine color tends to turn purplish red afterwards.) If the reddish tinge continues over several days, after you've eliminated foods that may be causing the color, you should mention it to your primary care provider. Greenish urine can mean that you have a urinary tract infection but if it's bright green, it can mean that you have too much of B vitamins. Certain drugs and some health conditions can cause urine to turn green, too, so mention this to your health care professional. Urine with a bluish tint can mean some type of bacterial infection or may just mean that you have a high level of calcium. If this persists, give your health care professional a call. Brown urine can indicate a potentially serious condition and you should contact your doctor's office for help. Sometimes laxatives can temporarily turn urine brown but if this dark color stays, get it checked out.

What are false-positive elevations in the PSA test?

False-positive elevations in the PSA are increases in the PSA that are caused by conditions other than prostate cancer. For example, benign prostatic hyperplasia (BPH) and infection or inflammation of the prostate (prostatitis) from whatever cause can elevate the PSA. Note also that even a rectal examination or an ejaculation within the prior 48 hours can sometimes elevate the PSA. False-positive elevations are usually in the 4 to10 range, but they can go as high as 25 or 30. At these higher levels, however, caution in the interpretation of the test is warranted because a prostate cancer may well be present. Non-prostatic diseases or infections, medications, foods, smoking, and alcohol do not cause false-positive elevations of the PSA. The ability of the PSA test to detect prostate cancer (called the sensitivity of the test) is high. The reason for this is that most patients, although not all, with prostate cancer have a borderline or an abnormally elevated PSA. The ability of the test to exclude other diagnoses (called the specificity of the test), however, is lower because of the other conditions that can cause false-positive elevations of the PSA.

Figure 13. Clicking on the topics displays all the questions and answers of that topic



What can be done to minimize allergies during periods of high pollen?

WebMD allergy consultant Paul Enright, MD, an Arizona-based pulmonologist, says it's important for children with allergies to avoid pollen when possible. Here's what Enright suggests to lighten your child's allergy load during pollen season: If the child is allergic to outdoor allergens (ragweed or other weeds, grasses, trees), determine when they are in bloom. You can ask your allergist, or check your daily newspaper or a weather web site for the latest pollen count. In addition, keep the child indoors whenever possible and close the windows of the home to reduce outdoor allergens from coming indoors. You may also consider running a HEPA room air cleaner in the child's bedroom; minimize cats and dogs coming indoors from outdoors (even if the child is not allergic to the pet), since their fur accumulates pollen and other outdoor allergens.

What is polycystic ovary syndrome (PCOS)?

A common ovulation problem that affects about 5% to 10% of women in their reproductive years is polycystic ovary syndrome (PCOS). PCOS is a hormonal imbalance that can make the ovaries stop working normally. In most cases, the ovaries become enlarged and appear covered with tiny, fluid-filled cysts. Symptoms include: No periods, irregular periods, or irregular bleeding. No ovulation or irregular ovulation. Obesity or weight gain (although thin women may have PCOS). Insulin resistance (an indicator of diabetes). High blood pressure. Abnormal cholesterol with high trigylcerides. Excess hair growth on the body and face (hirsutism). Acne or oily skin. Thinning hair or male-pattern baldness.

What does genital herpes look like?

Often it's hard to tell by looking. The textbook symptom of genital herpes is a cluster of small fluid-filled blisters that break, forming painful sores that crust and heal during several days. Affected areas include the penis, scrotum, vagina, vulva, urethra, anus, thighs, and buttocks. But many people don't get these sores. Some people have no symptoms at all, while others get symptoms that can be easily mistaken for razor burn, pimples, bug bites, jock itch, hemorrhoids, an ingrown hair, or a vaginal yeast infection. After you're infected, the symptoms go away, but can flare up from time to time. Luckily, the first outbreak usually is the worst. And some people may have just one or two outbreaks in their lifetime.

Figure 14. Clicking on the users displays all the questions he/she has answered, spanning across various topics.

e) Interactivity: The visualization does have a couple of interactive elements as shown above Hovering over the topics, shows all the users who have answered those topics along with the percentage contribution of each user. Hovering over the users, shows all the topics he/she is proficient in, along with a percentage estimation.We also display question and answers, related to the both the topic and the user. Clicking on the topics, displays all the questions on that topic and clicking on the users, displays all the question that user has answered.

2) *User-based visualization – Top Eight Users*

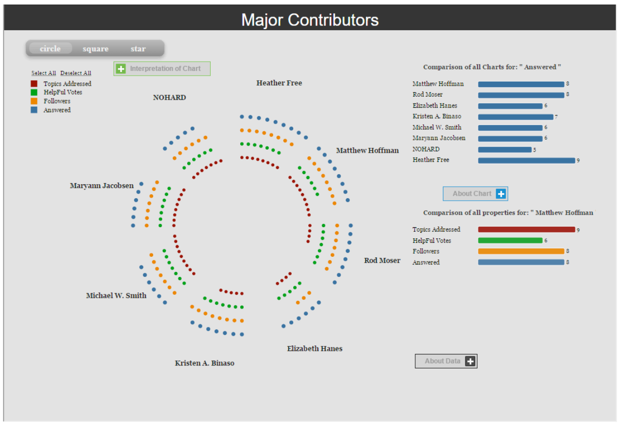


Figure 15. Chart wheel depicting top 8 users

a) Choice of Visualization: In order to depict the top 8 contributors, the chart wheel model was chosen with the number of circles indicating the score for a particular property. This allows the viewer to perform a visual comparison of the users based on the property. The properties form a concentric visual for each user. The properties are arranged in a sorted order with the least important property "Topics Answered" being inside the concentric view. The comparison amongst the users is depicted using a horizontal bar chart. For a single user the properties can be compared again by using a horizontal bar chart.

b) Data: The data for this visualization is obtained based on the analysis as stated earlier.

c) Size: The size of the concentric increases as one diverges out. This indicates the importance of the specific property with Number of Questions answered getting the highest importance and hence being displayed on the periphery on the visualization. The size of the horizontal bar graphs is as chosen to depict only the points scored for a user or a property.

d) Color: The color utilization is based on categorical selection tied to the importance of a property.

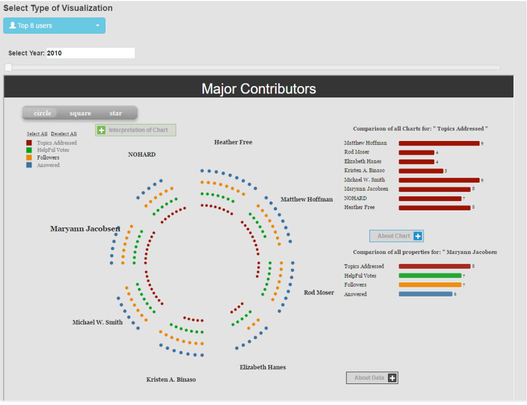


Figure 16. Chart Wheel showing details for Maryann Jacobsen and also comparing the "Topics Addressed" attribute for all users.

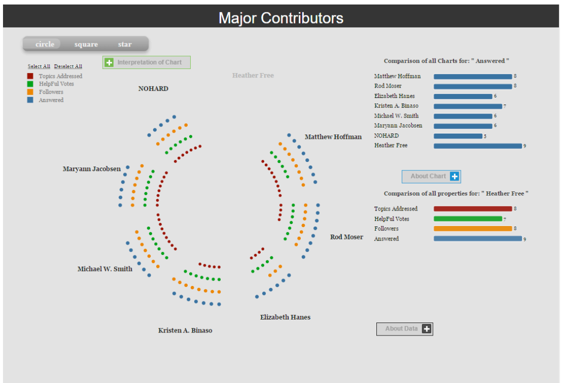


Figure 17. Clicking on the user's names, hides his/her details, making way for better comparison of the remaining users.

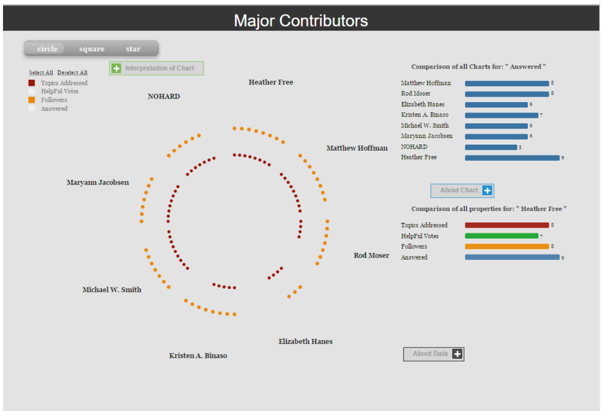
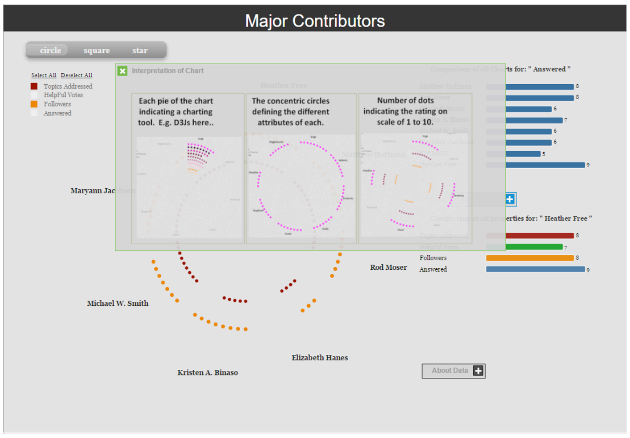


Figure 18. Clicking on the attributes, hides the respective attribute circle, enabling better comparison.



Each pie of the chart indicating a charting tool. E.g. D3Js here..

The concentric circles defining the different attributes of each.

Number of dots indicating the rating on scale of 1 to 10.

Figure 19. Details of the chart being depicted in the visualization itself

e) Interactivity: Hovering over each user will give a visual comparison of the 4 properties in the form of a horizontal bar graph. Hovering over a specific property displays a comparison of the top 8 users. Clicking on a user will hide his/her concentric circle scores. This will help a viewer to compare the results on the chart wheel itself. Clicking the property will hide the property out so that a comparison can be made with the rest of the properties.

C. *Infographic Reports*

Infographic reports by definition is a representation of information in a graphic format designed to make the data easily understandable at a glance. In this project, we have produced infographic reports for the four trending topics that is Pain, Pregnancy, Vision and Intercourse. Each of these reports contains various facts and information about the topic that the user can read and get acquainted to.

a) Data: The infographic report mainly contains a graph showing the progressive trend in the increase of the number of question over the years. It also contains a word cloud of all the answers relating to a particular topic. Upon studying these, certain points and statistics have been stated.

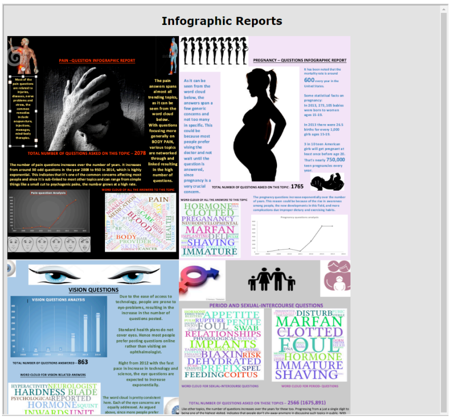
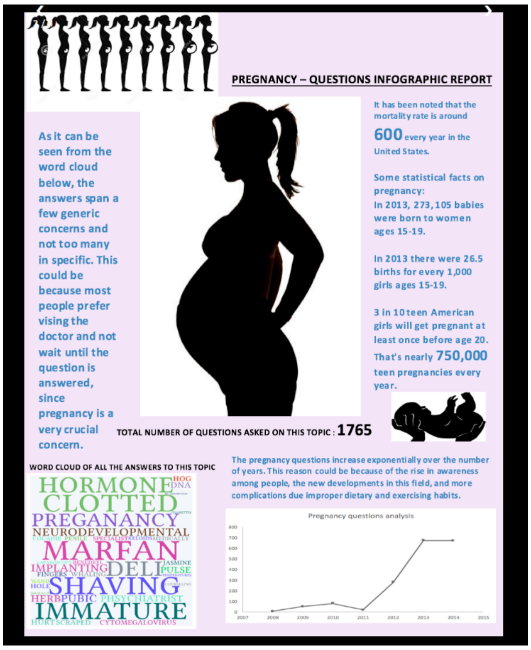

Infographic Reports
PAIN - QUESTION INFOGRAPHIC REPORT
PREGNANCY - QUESTIONS INFOGRAPHIC REPORT
VISION QUESTIONS
PERIOD AND SEXUAL-INTERCOURSE QUESTIONS

Figure 17. Infographic Reports being displayed.



PREGNANCY – QUESTIONS INFOGRAPHIC REPORT

As it can be seen from the word cloud below, the answers span a few generic concerns and not too many in specific. This could be because most people prefer vising the doctor and not wait until the question is answered, since pregnancy is a very crucial concern.

It has been noted that the mortality rate is around 600 every year in the United States.

Some statistical facts on pregnancy:
In 2013, 273,105 babies were born to women ages 15-19.

In 2013 there were 26.5 births for every 1,000 girls ages 15-19.

3 in 10 teen American girls will get pregnant at least once before age 20. That's nearly 750,000 teen pregnancies every year.

TOTAL NUMBER OF QUESTIONS ASKED ON THIS TOPIC : 1765

WORD CLOUD OF ALL THE ANSWERS TO THIS TOPIC

The pregnancy questions increase exponentially over the number of years. This reason could be because of the rise in awareness among people, the new developments in this field, and more complications due improper dietary and exercising habits.

Figure 18. Clicking the reports, zooms them for better reading.

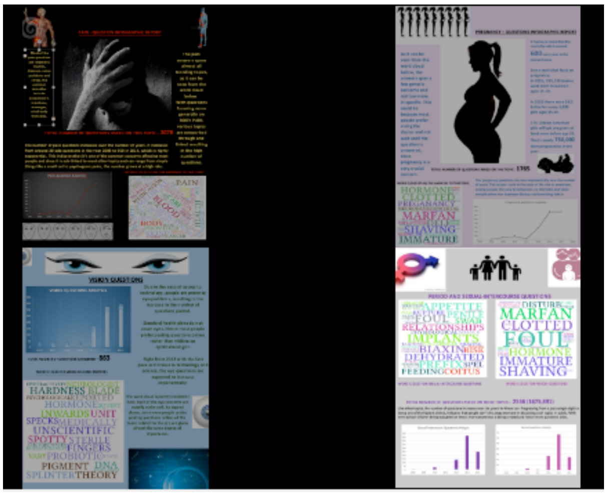


Figure 18. List of reports present below the zoom image available for selection.

Conclusion

In conclusion of this project we have achieved the following tasks: Semantic Data Analysis of WebMD and Visual display of semantic data analysis. A detailed analysis is performed on the data and this is being displayed in an interactive manner which is user-friendly and easily graspable.

With the implementation of Topic Facet Modeling, the user is easily able to judge which question fits perfectly with the topic and choose to view those questions accordingly. The sentiment value helps the user to view the best fitting answer. This saves a lot of time for the website user and aids him his/her look for relevant question and answers in a timely manner. The chart-wheel information about the user provides the website user with the information as to which consumer has the highest votes and who has answered most number of topics. This helps the user choose who's answer he/she wants view. Lastly, the infographic reports help the user understand each topic in detail.

All the visualization's together basically helps the user find his/her answers quickly which are accurate, relevant and apt.

Future Scope

One of the major limitations of this visualization is if the volume of users and questions increases, it would not be capable of holding that much data. So we would like to come up with improved scalable visualization to that would be robust and give quick response. The other limitation we would like to highlight is we would want to associate a backend processing engine that would keep track of user interaction so that we would be also able to analyze data on the go. The topic models that we have used do not provide accurate results on topic as well as the sentiment values associated with a question and answer, since with every iteration it was giving different facets. So, we would like to improve and come up with an ideal approach for conversational text based topic modelling that can be applied for discussion forums posts.

Acknowledgment

We would like to take this opportunity to express our gratitude and thanks to professor Dr. Sharon Hsiao and TA Yihan Lu for their constant guidance and support throughout the course of the project.

References

[1] Social Visualization Encouraging Participation in Online Communities (2006) Lingling Sun, Julita Vassileva, Groupware: Design, Implementation, and Use, Lecture Notes in Computer Science Volume 4154, 2006, pp 349-363

[2] What attributes guide the deployment of visual attention and how do they do it? J. M. Wolfe and T. S. Horowitz, Nature reviews. Neuroscience, vol. 5, no. 6, pp. 495-501, Jun. 2004.

[3] What attributes guide the deployment of visual attention and how do they do it? J. M. Wolfe and T. S. Horowitz, Nature reviews. Neuroscience, vol. 5, no. 6, pp. 495-501, Jun. 2004.

[4] What attributes guide the deployment of visual attention and how do they do it? J. M. Wolfe and T. S. Horowitz, Nature reviews. Neuroscience, vol. 5, no. 6, pp. 495-501, Jun. 2004.

[5] Discovering interesting usage patterns in text collections: integrating text mining with visualization, Anthony Don, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman, and Catherine Plaisant (2007)

[6] Narrative Visualization: Telling Stories with Data. E. Segel and J. Heer. IEEE InfoVis 2010.

[7] Hsiao, I.-H., & Awasthi, P. (2015). Topic FacetModeling: Visual Analytics for Online DiscussionForums. Paper presented at the The 5th international Learning Analytics & Knowledge Conference, Marist College, Poughkeepsie, NY, USA.

[8] http://christopheviau.com/d3list.