Know Your Diseases!

Kashyap Bhansali

Graduate Student  
Computer Science (MCS)

Arizona State University

[Kashyap.bhansali@asu.edu](mailto:Kashyap.bhansali@asu.edu)

Saloni Shah

Graduate Student  
Computer Science (MCS)

Arizona State University

[Saloni.shah@asu.edu](mailto:Saloni.shah@asu.edu)

Raj Dalvi

Graduate Student  
Computer Science (MCS)

Arizona State University

[rsdalvi@asu.edu](mailto:rsdalvi@asu.edu)

**ABSTRACT**

Visual analytics is a kind of analytics in which the data is showcased visually displaying analytical results and insights extracted from large, complex dataset. Intelligent Visual Analytics is not only about showing statistics using charts but also involves text processing using text mining and Information Retrieval to represent important information visually. Know your diseases! is one such kind of intelligent visual analytics. In this paper, we describe the analytical system in it’s entirety, ranging from the visualizations displayed and the principles  involved behind its selection. The data from Q/A portal called WebMD is processed, cleaned and represented to not only look aesthetically pleasing but also to convey the information visually. The goal of the system is to convey as much information as possible while the user is interacting with the system.

**Keywords**

Health data visualization, Visual Analytics.

# INTRODUCTION

We visualize information to meet a very basic need - to tell a story. Vision is a single most important faculty we use to communicate information. A Visual can communicate more information than a table, especially when we have large-scale data.  It is important for visualizations to answer questions, support decisions and communicate information [<http://www.fusioncharts.com/whitepapers/downloads/Principles-of-Data-Visualization.pdf>]. Today, one of the richest internet sources of information on public opinion are discussion forums.  [<http://jmlr.org/proceedings/papers/v11/trampus10a/trampus10a.pdf> ]  We developed an Intelligent Visual Anayltics for one such forum called WebMD.

Over the years, there is an increase in the amount of users who have started using the question and answer forums to gain basic information about their health. Users find it convenient to find answer to their questions just in time.These systems allow people to interact with other users and share their knowledge about anything related to health. These forums also claim that sharing helps benefit those who anticipate similar medical procedures. There have being plethora of such websites that provide information about diseases, symptoms and drugs, like <http://www.thepatientforum.com/> , [www.ehealthforum.com](http://www.ehealthforum.com), <http://www.healthcaremagic.com/>  and many more.

WebMD is primarily an online publisher of news and information pertaining to human health and well-being[1]. It is widely used as a health forum inundated with plethora of symptoms, health concerns and expert advice. We’ve implemented an analytical system which goes through a broad expanse of question topics, question-answers, WebMD community member details , and disease details. Know your Diseases! lists the most prevalent topics and based on the selection of topic displays an interactive visualization which shows the relationships between various diseases and symptoms, related topics, the prevalence of disease over the past years, and reveal the affected region of human anatomy affected. Overall the idea is to allow the user to get a good overview of what is being discussed and how things are related to each other and also have the ability to get more details on demand from Q&A discussed on the WebMD forums. This analytical system aims at providing a complete analysis of a given disease topic discussed on the forum.

In section 2, we describe what motivated us to build such a system. Section 3 discusses the visual design of the system followed by Section 4 discussing about the implementation details. We then conclude by presenting some unusual findings and discussing some future goals.

# MOTIVATION

# “Health is wealth” - There is nothing in our life that is more valuable than good health[2]. A healthy person is completely free from any illness or injury. Therefore, staying healthy and sound is very crucial.

The question and answer portals are modern form of an information retrieval system. A study conducted by M. Dolores et al in[5] collects hundreds of definitional question from two health related websites WebMD and MedQA. The retrieved answers were tested for four evaluation measures like precision, mean reciprocal rank, total reciprocal rank and first hit success to check the quality of the answers given by users on this portal. It was seen that these systems are useful to retrieve basic healthcare related information. Some of the answers from MedQA also gave links to future research articles. This gives us a simple insight into how a question-answer forum can be a tool providing realm of information access.

Recently, the software firm Intel developed a way to provide personalized care[6] to individuals. the main idea behind the project was to provide unique health care information to a patient keeping in mind their situation and needs.  They promise to help individuals to be proactive in deciding which patients are at high risk for readmission into hospital, admission to ICu and similar conditions. All this is powered by their technology and health related DATA.

This inspired us to design an analytical system ‘Know Your Diseases!’ which would assist in speculating health problems based on symptoms and would present an insight into the disease prevalence and related problems. We aim to provide answers to following research questions: 1. What kind of questions are asked on the forum? What are the most commonly discussed Topics & Diseases? 2. What are the Symptoms people with particular Disease have? What parts of the human body are affected by this Disease? 3. Prevalence of the Diseases in the population over the years? Hence, a better health quotient was what is aimed at, while designing and implementing ‘‘Know Your Diseases!

# VISUALIZATION DESIGN

Our visual analytical system know your diseases! consists of interesting visuals that helps us answer the research questions better.

We designed our visualizations in a way to induce viewers to think more about the substance and what questions are we trying to answer by visualizing the data and knowledge that we extracted from the WebMD Q&A discussions.

With the amount of information that we use and want to understand visually, it becomes important and challenging to use the right visuals and colors.

We have selected our choice of visualizations based on the Visual Information seeking mantra:

* Overview first
* Zoom & filter
* Details on demand

Following are the design principles that we follow:

* Correlation doesn’t imply causation
* Exaggerate or Simplify
* Generalizations improve clarity
* Emphasize important information or de-emphasize irrelevant details

It is important to have a way of revealing several levels of details , from a broad view to a fine structure. It is important in visual analytics to give the user the greatest number of ideas in the shortest time with the least amount of ink in the smallest space [].

Choosing the right colors improves the perception and makes the understanding the data much simpler. It is also necessary to think how the whole interaction will work. In our implementation we crosslinked the charts. That means we have coordinated views where an interaction in one chart will drive the changes to other charts to show the corresponding data.

It analyses different group of diseases into an interactive chart showing a hierarchy of health problems. Selecting one bubble subsequently alters the chord diagram to show related symptoms and health diagrams. This helps the user explore his related health problems which he is not aware about and hence, helps him to analyze his situation better. To facilitate better understanding of diseases under consideration, our human physique highlights the human body area infected and the line chart explains how the disease has been prevalent in the past. This implementation provides a complete analysis of a disease from the WebMD dataset in its entirety, ranging from symptoms, related problems, and its prevalence in the world.

## Zoomable Bubble Chart

Zoomable bubble charts are basically charts that have nested bubbles in them. This sort of visualization is aesthetically pleasing and quite intuitive when we need to demonstrate a Hierarchy in or data. As the name says, this kind of chart will have multiple levels to zoom in. The nature of the bubble also provides one more dimension, that is, the radius of the bubble, which can be used to determine some quantity.

Using the same idea we decided to use a zoomable bubble chart to represents a hierarchy of topics discussed on the WebMD forums.

We clustered a set of topics to build a hierarchy which would help a user to get a good overview and also get to the smallest detail if required. This chart is the driver for the rest of our visual analytics system.

The bubble chart implementation provides a wide taxonomy[3] of health problems grouped into following categories:

* Gender Specific.
* Infants
* Pain
* Hair and Skin
* Drugs
* Hands and legs
* Obesity
* Respiratory Problems
* Food and Drugs
* Sensory Organs
* Stomach, Kidney and Intestines
* Heart Problems
* Sexual Problems
* Surgery
* Blood Problems
* Brain and Spine

The groups have been intentionally named in layman terms so as to facilitate effortless selection. Each big bubble encompasses smaller bubbles, which zoom in on the selection of Parent bubble. Size of the bubble symbolizes the relative number of health complications it categorizes, or the number of.

The motivation of choosing a prescriptive bubble chart for taxonomy was to maintain uniformity in parent and child (representing all the diseases in form of bubbles one inside the other), and to provide easy navigation from one group of diseases to the other.

The radius of the bubble indicates which topics have the most questions. This could help the user to get an overview as to what kind of topics are discussed the most on the forum. It makes it easy to drill down to see comparison between individual topics belonging to the same cluster.

Color choices have been made taking into consideration the visualization principles of color and vision. A clear sequential scheme has been used to color the bubbles, cyan ,blue and white (going down the category tree). [4] White color has been used to enhance visualization of the sub-most category in each parent category. Shades of blue have been used because blue symbolizes trust, peace and comfort.

For example, by only looking at the Zoomable bubble chart, we can see and tell the following things:

* there is a lot of discussion about Women health problems accompanied by Q&A’s for General Symptoms like headache, nausea, burns, etc.
* Pain and infection are the most common complications associated with a gamut of diseases.
* Complicated issue like Cancer and Surgery have been less discussed and talked about on the forum, since they’re critical enough to directly consult a doctor in-person.

## Chord Diagram

Chord Diagram is a graphical method of displaying the inter-relationships between data. The data is arranged radially around a circle with the relationships between the points drawn as arcs connecting the data together [<https://en.wikipedia.org/wiki/Chord_diagram>]. It is a beautiful way of representing relationships in the data.

This kind of a chart was popularized in 2007 by a post on The New York Times that discussed “Close-Ups of the Genome, Species by Species by Species” [http://www.nytimes.com/imagepages/2007/01/22/science/20070123\_SCI\_ILLO.html].

An intelligent chord diagram demonstrates how symptoms and health problems can be linked to other health problems, i.e. the problems which can be simultaneously present or in other words, prevalence of which can make the present symptom into consideration, more probable. All the datasets, especially questions and answer sets have been analyzed to extract all the relations between a gamut of health issues. Figure below is a sample of chord diagram when smaller bubble muscle pain in parent bubble ‘Pain’ is selected.   \*\*\* Picture \*\*\*. <Desc of the figure>

Nature has created our body in a way that all the organs work together in a  highly coupled fashion. A little malfunctioning in one, results in a complication in other system. For example, a small infection in oesophagus, if left unattended,  might result in ulcers of stomach infection. Or in case one has a throat infection, then it may increase the probability of having ulcers. Therefore, for an analysis of a disease, it is very important to obtain related health problems, since it opens new avenues to understand the disease and to find ways to treat it, since treating one problem may help cure the other. Also, in the diagram, a certain chord connects all the problems which have occurred simultaneously together in some other instances. Different contrasting color options have been selected to display related yet different problems.

## Human Physique

Knowing symptoms, related health problems or broad classification of your problem is not enough if you want to treat your health complication. The most important thing lies in the region of your body where the root of problem resides. For example, patches or rashes in skin, can be caused by the disease dengue, which is a viral attack, not on your skin.

Also it would be wrong to assume that the user of our system would know about all the diseases and symptoms and what areas of the body internally or externally it affects. This gave us the idea of add Human Physique diagram to our analytics system which gives a general overview of what areas might be affected with a particular disease.

Therefore, this particular visualization displays the region in body where the selected bubble effects and realizes pain. The following two figures shows the <Figures>

## Line Chart

A line chart is a chart that plots a series of data points and these data points are connected by straight line segments [Burton G. Andreas (1965). *Experimental psychology*. p.186]. A line chart is often used to visualize a trend in data over intervals of time[Neil J. Salkind (2006). *Statistics for People who (think They) Hate Statistics: The Excel Edition*. page 106.]. Specially in case of the time series data. Thus the line is often drawn in a chronologically order.

In our project we make use of a simple line chart to demonstrate a trend in the prevalence of disease over the years or months. When a topic is selected, this line chart is updated based on the number of questions asked on the forum related to that particular topic over months or years.

A user using our intelligent analytics system could quickly answer if the questions related to certain topics are asked more or less and at around what time of the year. This could help in identifying the trend in symptoms that occur due to change in weather and seasons.

## Relevant Questions and Medical Tests

This section of the analytics dashboard displays the most recent questions asked related to a particular topic. The answers to these questions can be read by clicking on it. The question links directly to the WebMD website where this question was asked, along with providing extra details like Name of the user, User ID, question post date, etc.

Along with symptoms and drugs, we see that relevant medical tests also play an important role in providing information to users. Thus, in our visualization we mention the most relevant medical tests that are discussed amongst the people in question and answers.

# METHODOLOGY

Text Mining involves many different techniques to extract important data from the dataset. In this case, we scraped the data from WebMD datasets and from some other places like Mayo Clinic to extract the relevant close matches from the questions and answers using sequence matching. Sequence matching has been implemented in python, with a library called ‘difflib’. Sequence matching can be used to match sequences in strings which are hashable. This was preferred over TF-IDF because it reduced the occurrence of meaningless data and also took care of typos which could be there because of the complex spelling of diseases and symptoms. TF-IDF (**t**erm frequency–inverse document frequency) only looks for words in a corpus which it thinks are important but we are not just going for important words. With a bag-of-words approach, some clustering was performed using some manual inputs. A hierarchical structure, was then formed, which has been used to construct a packed zoomable bubble chart.

The bubble chart has various topics and on clicking any of the child bubble, you can see the diseases and connected symptoms in a chord diagram. The bubbles are generalized into topics to improve overall clarity. [Data-Ink and Graphical Redesign, In The Visual Display of Quantitative Information. (2001) Tufte.]. The size of bubble indicates the amount of times the topics and disease are discussed. For instance, there are maximum number of questions and answers related to Women Health problems followed by Sensory Organs in human body and general symptoms. This helps the user of the system to get a detailed overview of most commonly discussed topics and diseases. The zoom and filter capability help the user to know about the disease and symptoms in the linked chord diagram. More details are provided as the user zooms in to the innermost level of the bubble chart. Thus, the visualization provides complete justice to Visual Information Seeking Mantra taught in class. [Design principles for visual communication. (2011).Agrawala, Maneesh, Li, Wilmot, & Berthouzoz, Floraine. Commun. ACM, 54(4), 60-69. doi: 10.1145/1924421.1924439].

The chord diagram was created by mapping the topics and their presence in questions and answers datasets. Difflib has a tuning factor which when increased, increases the accuracy but reduces the number of results. Thus, we had to decide a proper tuning factor for both question and answer datasets. We went with 0.6 because it was the ideal factor. The aim in designing the chord diagram is to make large data sets coherent. The chord has multiple colors to show different categories of diseases and a disease and its symptoms are linked together with the same color. Ultimately, as discussed in class, the chord diagram conveys multiple complex ideas with clarity, precision and efficiency. In some cases, the data is not large enough to show a chord diagram in which case a message is put up informing the user that data is not enough to show a chord diagram. At the same time, there might be some topics which are present in multiple questions such as pain whose chord diagram appears really dense.

When a bubble in the chart is clicked on, the chord diagram displays the diseases and symptoms. Apart from that, the affected body part is highlighted in the image of a human body. Here, the diseases affecting either men or women are also present so the bodies of both a man and a woman are present and the affected body parts are highlighted. Majority of the body parts were extracted from the dataset but in some cases, manual inputs were needed to improve the accuracy. Another challenge in this human body usage was to highlight diseases such as diabetes which are not exactly centered on a particular organ of the human body. In that case, the whole body is highlighted or the closest possible body part is highlighted.

The trend of discussion over a particular topic was highlighted using a line chart. The line chart changes for every disease and is interactive with a tooltip present to show the values. This line chart was created by using D3.js and a csv file containing the topic name, the month-year and the number of times the topic was present at that month-year. The same analysis was done in the question and answer set to get a comprehensive set of information. To generate the csv, python was used. A progressively rising line chart indicates that the  prevalence of the disease/symptoms increased over the year.

# EVALUATION PLAN

Because visual analytics and data visualization are designed for humans, we focus on visual interfaces and presentations, which are well understood, based on human biases and how users interact with computers.  [<http://data-informed.com/7-tips-for-evaluating-data-visualization-software/> ]. Our results are completely based on the questions and answers posted by humans on this forum. These humans are variety of groups ranging from patients, physicians to medical experts. Our results not only will help common people to gain basic knowledge but also health care professionals as they use symptoms and signs as clues that can help determine the most likely diagnosis for a disease. However, proper testing performed by a doctor is recommended to render a diagnosis because many patients experience some symptoms and not others.

[<http://www.medicinenet.com/symptoms_and_signs/article.htm> ]

We apply bottom-up method of processing as discussed in [<http://idl-bnc.idrc.ca/dspace/bitstream/10625/49286/1/IDL-49286.pdf> ] . Firstly, a user extracts information just by having a glance at the analytics dashboard and knowing it showcases medical information without any other thought. The efficiency of this process is related to the large amount of neurons, up to five billion, which are simultaneously processing different features

# DISCUSSION & FUTURE WORK

Upon successful implementation of the visual analytics, we see that some critical disease like Cancer, Blood problem and the likes are discussed less as compared to common diseases. This indicates that users do not like to discuss critical topics on a forum and would prefer seeing a doctor in person. We also notice that there is a lot of discussion about Women Health problems followed by Sensory Organs of human body. Pain and infections are the most common complication associated with a gamut of diseases.

After the implementation of showing how diseases and symptoms were related and their behavior over time, we also wanted to display the areas in the world affected by these diseases. For this, we digged into the dataset to find data related to geographical location of a user. However, we were not able to implement this due to vague or blank output we received for almost all diseases. If some concrete data related to location of the users posting question and answers can be found, we would add a map visualization showing the occurrence of disease in various parts of the world. We would use choropleth maps showing the country highly affected by the disease in darkest shade and lightening its shade with countries affected relatively less.

We would also want to extend this project to a visual recommender system for health related dataset. After the user enters a symptom, we would recommend the diseases which commonly see that symptom and vice versa as well. We would also recommend the commonly occurring diseases based on the users’ geographical location. After knowing about the basic principles of data visualization and analysis, we would be able to do complete justice to a recommender system.

# ACKNOWLEDGMENTS

We thank our professor Dr. Sharon Hsiao for her reviews about our ideas and guidance in this project. Her presentation slides and weekly reading of research papers were very helpful in forming a basis for this project. We would also like to thank Arizona State University’s School of Computing, Informatics and Decision Systems Engineering for providing us with an opportunity to learn this course of Data Visualization and broaden our horizons and ultimately, improve our skill set.

# REFERENCES

1. Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= <http://doi.acm.org/10.1145/161468.16147>.
2. Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.
3. Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands, April 01 - 06, 2000). CHI '00. ACM, New York, NY, 526-531. DOI= <http://doi.acm.org/10.1145/332040.332491>.
4. Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.
5. Sannella, M. J. 1994. *Constraint Satisfaction and Debugging for Interactive User Interfaces*. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.
6. Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.
7. Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM, New York, NY, 1-10. DOI= <http://doi.acm.org/10.1145/964696.964697>.
8. Yu, Y. T. and Lau, M. F. 2006. A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions. *J. Syst. Softw.* 79, 5 (May. 2006), 577-590. DOI= <http://dx.doi.org/10.1016/j.jss.2005.05.030>.
9. Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= <http://doi.acm.org/10.1145/90417.90738>.

Columns on Last Page Should Be Made As Close As Possible to Equal Length