EVALUATING THE READABILITY OF SCIENTIFIC WEB PAGES USING INTELLIGENT ANALYSIS TOOLS

A Thesis
by
SEENA SUKUMARAN MENON

Submitted to the Graduate School
Appalachian State University
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

December 2010
Major Department: Computer Science

EVALUATING THE READABILITY OF SCIENTIFIC WEB PAGES USING INTELLIGENT ANALYSIS TOOLS

A Thesis
by
SEENA SUKUMARAN MENON
December 2010

APPROVED BY:

Rahman Tashakkori
Chairperson, Thesis Committee

Cindy A. Norris
Member, Thesis Committee

Carl R. Russell
Member, Thesis Committee

James T. Wilkes
Chairperson, Computer Science

Edelma D. Huntley
Dean, Research and Graduate Studies

Copyright by Seena Sukumaran Menon 2010
All Rights Reserved

ABSTRACT

EVALUATING THE READABILITY OF SCIENTIFIC WEB PAGES USING INTELLIGENT ANALYSIS TOOLS

Seena Sukumaran Menon, B. E., University of Mumbai

M. S., Appalachian State University

Chairperson: Rahman Tashakkori

The World Wide Web (WWW) is a primary resource of information. However, due to its exhaustive and complicated nature, verification of the relevancy and quality of information on the WWW presents a major problem. A user has to search for an appropriate document, verify the relevancy, read and comprehend the information provided. This is more complicated in the case of scientific web pages. Scientific web pages often include text content, tables, graphs, charts, images and mathematical formulae that are difficult to represent in a legible manner. Readability of a web page is an indicator of how easy it is to view, read and understand the contents. There are multiple factors that affect the readability of web pages – for example, consistency of fonts, use of background colors and formatting.

Our study involved creating a sample scientific website along the lines of a conventional scientific website. Users had to browse through the sample website and answer a survey questionnaire to record their experience with the website. The collected data was then analyzed using the data mining techniques of the SAS Enterprise Miner to determine the main factors affecting readability of the website. Visualization techniques in SAS Miner

were utilized for data analysis. In the future, this analysis may be used in developing an algorithm to redesign a web page for better readability.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the vision, guidance and support of Dr. Rahman Tashakkori. I am grateful to Dr. Cindy Norris and Dr. Ray Russell for their advice and encouragement. Thanks to the faculty and staff at the Department of Computer Science, the Office of Student Research and the Office of Research and Graduate Studies. I am indebted to all the students and faculty at Appalachian State University who took my readability survey and provided valuable feedback for my research.

Thanks to Shailendra, my husband and best friend, for believing in me and encouraging me to outperform myself. Thanks to my parents and in-laws for their blessings, well-wishes and prayers. A special thanks to all my other friends and family who helped me through my small loses and little victories.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Statement of Problem

An important aspect of Human Computer Interface (HCI) is the evaluation of interactive systems and determining how different factors affect usability or readability of such systems. This analysis may have a significant impact on the way these systems are designed. At the advent of the World Wide Web (WWW), very few guidelines existed for the design of websites and the arrangement of content on web pages. However, with the growth of the WWW, it is becoming clear that simply having a web presence is not sufficient. A lot of research has involved evaluating the traditional operational usability of a website [Joshi, 1999]. Not much significance has been associated with the aesthetic appearance of a website or the components of HCI. However, the appearance, design and user interface of a website can have a tremendous influence on users' perception of its readability. This study considers a subset of such aesthetic design factors and quantitatively measures their effectiveness through user survey questionnaires. The collected data is analyzed using data mining techniques. The analysis results reveal preferred values for the subset of readability factors under evaluation. The results also group readability factors that receive similar ratings from the participants and impact readability of the website in a similar way.

1.2 Literature Review

Literature provides numerous definitions for usability. Whitehead attempted to consolidate the definitions of usability presented by several researchers [Whitehead, 2006]. Whitehead indicated that usability was user and task dependent and related to how well the users were able to accomplish what they set out to do, how efficiently the users could do this

and how satisfied the users were during and after the process. Evidently, usability was complex and user-centered. Ivory et al. asserted that usability evaluation consisted of methods and procedures to measure the usability aspects of a system's user interface and to identify specific problems [Ivory, 2001a]. Capture, analysis and critique were common activities involved with their usability evaluation. Rosenholtz et al. claimed that management of clutter was an important factor in the design of user interfaces and information visualizations, allowing improved usability and aesthetics [Rosenholtz, 2005].

Usability can be quantified by measuring several usability metrics. Whitehead defined usability metrics as measures of a particular website or web page that had an impact on usability [Whitehead, 2006]. Our study evaluated usability of a sample scientific website by determining values for a subset of these usability metrics. Ivory, Sinha et al. evaluated web pages on the basis of attributes selected from the set of attributes used by Webby Awards [Ivory, 2001b] [Webby, 2000]. Webby organizers categorized the websites into different disciplines (e.g. news, finances and services). A panel of judges rated these websites on six primary criteria: content, structure and navigation, visual design, functionality, interactivity and overall experience. The metrics used by Ivory, Sinha et al. included word count, body text percentage and emphasized body text percentage, text positioning count, text cluster count, link count, page size, graphic percentage, graphics count, color count and font count.

Ivory et al. developed and analyzed over one hundred and fifty quantitative measures of page-level and site-level interface aspects (e.g. text count, number and types of links and consistency) [Ivory, 2005]. For some given sets of tasks, Brinck et al. measured the task completion rates of users, time taken by the task, average subjective ratings of individual

tasks and global subjective rating (including attractiveness, prestige, simplicity and so forth) [Brinck, 2003]. Improving scores for these metrics was used as an indication of the design and readability of the website. Hall et al. examined and presented the impact of text-color combinations on web page readability and the associated effect on behavioral intention of a user [Hall, 2004].

Joshi et al. proposed using web server logs to analyze and explore usage information for a website [Joshi, 1999]. Schaik et al. presented three important parameters for questionnaire design to evaluate readability of websites – namely response format, questionnaire layout and interaction mechanism [Schaik, 2007]. Schaik et al. measured four main aspects of quality of human-computer interaction – perceived ease of use, perceived usefulness, disorientation and flow. Swaak et al. examined the contribution of website characteristics (information usefulness, visual attractiveness, actual and perceived usability) to the success of the organization behind the website [Swaak, 2009].

Several researchers used various tools and techniques for evaluating websites. Ivory, Sinha et al. mentioned that the traditional quantitative methods for evaluating websites focused on statistical analysis of usage patterns in server logs, traffic-based analysis (e.g., pages-per-visitor or visitors-per-page) and time-based analysis (e.g., click paths, page-view durations) [Ivory, 2001b]. These methods had less reliability as web server logs often had only partial information about usage and timing estimates could be influenced by network latencies. In addition, these methods mainly concentrated on the operational usability of the websites and were not concerned with their aesthetic design. Ivory, Sinha et al. developed an automated tool to compute a subset of Webby web page metrics for about two thousand pages belonging to several Webby website categories [Webby, 2000]. The scores computed

by the tool were analyzed to evaluate if they could predict the Webby experts' judgments about web pages accurately. The study concluded that simple and superficial web page metrics measured using the automated tool were capable of predicting Webby experts' judgments with some degree of accuracy. The current study used a similar subset of web page metrics to evaluate the readability of scientific web pages.

Results from the above research formed the basis for the study conducted by Ivory et al. [Ivory, 2002]. The analytic tool was modified to include evaluation of page performance and consistency of page measure across a website. The results of this analysis were used to make suggestions about how to modify the site to comply with highly rated websites. Some of the recommendations made by Ivory et al. were used to verify the analysis results from our study.

A longitudinal study of web design patterns was carried out over a period of four years by Ivory et al. [Ivory, 2005]. The results from the study were used to compare designs of websites to the well designed ones in order to determine whether their designs exhibited similar properties and if not, to determine how their designs differed. The study also provided an evolution of website design over the selected time frame. The analysis and design recommendations from this study were useful in verifying the results from our study.

Some common evaluation techniques, such as formal user testing, were presented by Ivory et al. and could be applied in the early stages of design [Ivory, 2001a]. Ivory et al. suggested that each technique had its own requirements and discovered different usability issues. They presented taxonomy for the process of automating website evaluation. Description and procedural analysis of various website evaluation automation tools was also provided.

A Feature Congestion measure was proposed by Rosenholtz et al. for display clutter, based upon the saliency of elements in a display [Rosenholtz, 2005]. A set of maps was used and tested for two main features: color and luminance contrast. User surveys were used to collect observer rankings for measuring perceived clutter on the maps. Feature Congestion measure of clutter was made and compared to the observer rankings. Correlation between the two was very high, proving that the Feature Congestion measure of clutter had some reliability. The procedure, analysis and recommendations from this study contributed to the way our sample website was designed for survey.

Brinck et al. redesigned a school website based upon the results from a metrics-based user testing process [Brinck, 2003]. User performance on the two websites (original vs. redesigned) was compared to determine the improvement in the readability and usability of the website. This was done incrementally with continuous user-testing throughout the development of the redesigned website. In each round of testing, problems from previous rounds were considered and design changes were made and tested to address them. The idea was to make the website visually attractive and functional, but also to offer simple and successful user experiences. Recommendations from the experiment included use of consistency across the pages, use of more colors, use of a breadcrumb display to show the progress on a page, design and use of links and design of utility pages (e.g., page not found). Changes made to the website on the basis of these recommendations improved the overall score for the website and successfully improved the readability of the pages. Although, the analysis methods and web metrics used in our study were very different from the study conducted by Brinck et al., the basic procedure of collecting user data, analyzing it, making

recommendations and incorporating the recommendations in the design of the website was comparable to our approach for the current study.

Understanding how sighted users browsed web pages could provide important information to enhance website accessibility for visually impaired users [Michailidou, 2008]. For this study, Michailidou et al. conducted an eye tracking study for investigating the browsing behavior of sighted users and how it related to the pages' visual clutter. Results demonstrated that majority of the users tended to spend more time on the main content of a web page and fixated on the first three or four items on the menu lists. Gaze patterns were tracked to understand the most common way of reading web pages. Michailidou et al. recommended that the results could be used to develop guidelines for designing and modifying web pages for easier and faster access for visually impaired users. The study gave useful insight into how users perceived and interpreted the presentation of information and elements on a web page. It provided information on the relationship between visual presentation and users' browsing behavior. The results helped in eliminating some extraneous variables during the user survey of the sample scientific website.

Angeli et al. provided a comparison between two websites with the same content, but different interfaces (traditional menu-based vs. interactive animated), on the basis of heuristic assessment of aesthetics, questionnaire assessment of aesthetics, content, information quality, usability and engagement [Angeli, 2006]. This procedure was analogous to the comparison of the sample scientific website before and after the user survey in our study. Angeli et al. reported that initial research findings suggested a correlation between aesthetic quality of an interface and its perceived usability and overall user satisfaction. They presented a model of user experience building on their initial findings. Responses from the user survey of the

websites were categorized as per the cause of usability problem and then analyzed. Scores of usability and aesthetic factors were graphically represented and a correlation matrix was developed with the evaluation measures.

The complete process of evaluating, analyzing and improving the usability of a website was described by Erinaki et al. [Erinaki, 2003]. Erinaki et al. outlined the methods for the collection of website usage data, the modeling and categorization of the data, analysis of collected data and determination of actions performed for improving the readability of the website. Erinaki et al. performed user profiling on the basis of online surveys and questionnaires or navigational behavior of the users. The user profiles were then used to categorize the preferences, characteristics and activities of users. The results from user profiling were utilized for designing the survey questionnaire for the current study. Erinaki et al. described various methods to uniquely identify visitors to a web page and discussed several procedures and tools for data mining techniques like clustering. Such information provided useful inputs for the data mining analysis techniques used in this study.

A research model was proposed by Hall et al. based on the contention that contrast factors (e.g., dark background with light foreground text) influenced readability and retention and preference influenced aesthetic perception and behavioral intention [Hall, 2004]. Findings of the study proved that for the selected sample, pages with higher color contrast were perceived to be more readable. Color or content did not have a significant effect on the retention ability. Different color combinations highly influenced the aesthetic perception of the pages by users. There was a high correlation between the positive perception of a web page by a user and the amount of interest the user had in that particular page content (e.g., if the user desired to purchase a product displayed on a page). The experimental results from

the study were examined to design and analyze the sample scientific website and eliminate some of the extraneous variables due to the perception of a web page by a user.

Joshi et al. extracted structure from a dataset containing users' behavior accessing a website [Joshi, 1999]. The web server log information was preprocessed to be analyzed further. Unwanted entries were filtered out of the log information (e.g., access to image files embedded in web pages whose hit had already been recorded). The pre-processed log files were then analyzed using data mining techniques like session generation, clustering and association rules. Our study analyzed scientific website usage information and readability of such websites using similar data mining procedures.

Test Environment Automation (TEA), a flexible tool to support user tests by automating repetitive tasks and collecting data of user inputs and actions, was evaluated by Obendorf et al. [Obendorf, 2004]. TEA controlled test procedures, managed the interaction with users, provided survey questionnaires and recorded responses. It automated random display of pages in the browser and traced navigational actions of users. TEA traced user events and captured data for further analysis.

Design recommendations provided by Schaik et al. for questionnaires were used for the current study [Schaik, 2007]. Also, the procedure for evaluating the quality of human-computer interface elaborated by Schaik et al. was useful in the procedural setup for the study of scientific web pages.

Swaak et al. proposed a research model that hypothesized a relationship between website characteristics and people's trust in the organization with the website [Swaak, 2009]. Participants browsed through the website under evaluation and then recorded their opinions about the website characteristics. Regression analysis was conducted on the collected data to

verify the hypothesis. Results from the study confirmed that users' trust and behavioral intentions were affected by the visual attractiveness of a website and that perceived usability strongly related to actual usability. The observations made in this study were very recent and partly explained the attitude of users towards the perception of readability of web pages for our study.

1.3 Research Questions

Several factors affect the readability of a scientific website. Our research considers a subset of the aesthetic design factors and measures their impact on the readability of a web page. The research questions focus on the subset of the factors considered for the study. We address the following questions:

1. How do different fonts (face-type, size, color) affect readability of web pages?
2. Does having to scroll vertically on a web page affect its readability?
3. Does having to scroll horizontally on a web page affect its readability?
4. How does formatting (page justification) on a web page affect its readability?
5. Does readability of a web page depend on whether colored or grayscale images are used in it?
6. Is readability impacted by the presence of "ALT" descriptions for images on a page?
7. Does presence of a background image or color impact the readability of a web page?
8. Does the formatting and presentation of a graph or chart affect readability of a web page?
9. Does the formatting, size and presentation of tabular information affect the readability of a web page?

10. Does the formatting and presentation of mathematical formulae affect the readability of a web page?
11. Does the format or type of documents used on a web page affect its readability?
12. Does the arrangement of information, logical positioning of content and general format and display of data affect the readability of a web page?

1.4 Research Hypotheses

Based on the research questions addressed above, we propose the following hypotheses:

1. Fonts:

- Common fonts like Times New Roman, Arial, Verdana, Trebuchet and the likes are most preferred by users for better readability of web pages.
- Web pages with high readability scores have font sizes between 9-point and 14-point.
- Common font colors like red, blue and black give better readability for web pages.
- Content with bold, italicized and regular fonts improves the readability of web pages.

2. Page scroll:

- Web pages without vertical or horizontal scrolling are more readable.
- Web pages with vertical scrolling are preferred over horizontal scrolling.

3. Page formatting:

- Left justification of content on web pages gives better readability.

4. Image properties:

- Grayscale images give better contrast to web pages, thus improving the readability.
- Images with alternate descriptions improve the readability of web pages.

5. Background:

- Blank backgrounds or light colored backgrounds improve the readability of web pages.

6. Graph properties:

- Well-formatted graphs, with legible sizes and relevant colors, contribute towards better readability of web pages.

7. Table properties:

- Well-formatted tables with legible font sizes, headings, captions, descriptions and boundary lines result in better readability of web pages.

8. Mathematical data properties:

- Well-formatted mathematical formulae with legible font sizes are better for readability of web pages. Presenting mathematical formulae as images may be preferred by users, since they can be clicked and enlarged to get a better view.

9. Web document types:

- Web documents, articles, white papers in formats like PDF, Postscript and HTML are preferred and most readable on websites.

10. General content and presentation:

- Relevant and logical arrangement of content on web pages improves the readability of web pages. Users can find information more efficiently on a web page with data organized according to the topic of interest.

1.5 Significance of Current Research

The literature review suggests that most readability evaluation studies for websites have been conducted for educational, financial, commercial and service sectors. Although, scientific websites can be categorized under academic websites, there has not been much

research on evaluating and improving the readability of scientific web pages. Our study aims at understanding the factors that impact the readability of websites belonging to the varied disciplines of science (e.g., Mathematics, Astronomy and Computer Science), suggest and implement recommendations for improving the readability and usability of such websites.

Most previous research studies have focused on the traditional operational usability of websites including download speeds, bandwidth requirements and server log analysis. Fewer studies have concentrated on the evaluation of the aesthetic designs of websites. Norman suggested that aesthetic design of a website can have a significant influence on user perception of usability and readability of the website [Norman, 2004]. The current research study focuses on the aesthetic features of a website such as fonts, backgrounds and formatting.

The majority of previous studies have used professional resources like Webby awards and earlier literature to review and report the factors impacting readability of web pages. Webby awards have a panel of experts who evaluate and rate websites on certain pre-defined criteria. The statistics reported by much of the previous research have been extracted from such literature. Our study obtains opinions from actual users through survey questionnaires. A significant and representative sample of participants should reflect the general attitude of users towards the readability and usability of a website.

Unlike previous work, our study utilizes data mining techniques to analyze the collected survey data, in addition to statistical analysis. Data mining techniques would allow us to discover the trends of preference for various aesthetic characteristics and also to represent them visually. SAS Enterprise Miner software is used for the analysis and representation.

Further, none of the previous research studies have attempted to reconstruct or reformat the web pages to improve their readability. Our study is ultimately aimed at using the analyzed data to develop an algorithm to redesign a web page to obtain better readability.

1.6 Thesis Organization

Further chapters are organized as follows:

- Chapter 2: Theoretical Background:

  This chapter elaborates some important concepts and methodologies used for conducting this research study.

- Chapter 3: Methodology and Research Design:

  This chapter details the research design and methodology for conducting the study. It addresses the data collection methods, analyses techniques and expected outcomes.

- Chapter 4: Analyses and Results:

  This chapter describes the data collected through user surveys, clustering and data mining methods employed for analyzing the data. It also details the analysis results from testing the hypotheses stated in the above chapters. It presents the recommendations for improving the readability of a scientific web page.

- Chapter 5: Summary, Conclusion and Future Work:

  This chapter presents a summary of the study and elaborates the conclusions from the research study. It also lists some of the possible future enhancements.

CHAPTER 2: THEORETICAL BACKGROUND

This chapter discusses the significant concepts and techniques used for conducting the research study.

2.1 Scientific websites

Scientific websites can offer current information about the latest scientific discoveries and explanations of scientific principles. These websites provide information in various formats – charts, figures, mathematical equations and embedded documents (PDF and PostScript). The content can belong to any of the various fields of science – Computer Science, Math, Chemistry, Physics, Astronomy and Biology [ACM, 2009; IEEE, 2009]. The users of such websites tend to be experts in these disciplines and have at least basic experience and knowledge of using scientific websites. Organizations like ACM and IEEE provide rules and standards for technical papers, which can be applied to the design of scientific websites.

2.2 Readability

Websites can be designed in a number of ways. Although there is no right or wrong way of creating a website, certain combinations of properties can lead to websites that are relatively easier to view, read and understand. Readability is an indicator of such websites. It shows how efficiently users can achieve what they set out to do with the website and how satisfied they are with the process of finding the required information. Readability is a complex and user-centric concept [Whitehead, 2006]. Readability is an important factor in the design of user interfaces and aesthetic features of websites.

There are various factors that can impact the readability of a website. Surveys can help to gather first hand information on user preferences about the values of these readability factors [Schaik, 2007]. Some of the aesthetic design related readability factors include font type, font color and background [Webby, 2000].

2.3 Data Mining and SAS Enterprise Miner

Data mining is a technique for searching, analyzing and sifting through large amounts of data to find relationships, patterns or any significant statistical correlations. SAS or Statistical Analysis System, is a collection of software products that are grouped and offered by the SAS Institute. SAS Enterprise Miner streamlines the data mining process to create highly accurate predictive and descriptive models based on analysis of vast amounts of data [SAS, 2009]. The data mining process can be summarized below:

- Prepare appropriate data by creating one or more data tables. The sample should be large enough to contain the significant information, yet small enough to process.
- Explore the data by searching for anticipated relationships, unanticipated trends and anomalies in order to gain understanding and ideas.
- Modify the data by creating, selecting and transforming the variables to focus the model selection process.
- Model the data by using the analytical tools to search for a combination of data that reliably predicts a desired outcome.
- Assess the data by evaluating the usefulness and reliability of the findings from the data mining process.

All of the above steps may not be included in the analysis process and it might be necessary to repeat one or more of the steps several times before satisfactory results are achieved.

SAS Enterprise Miner contains a collection of sophisticated analysis tools that have a common user-friendly interface that one can use to create and compare multiple models. Analytical tools include clustering, association and sequence discovery, market basket analysis, path analysis, Kohonen self-organizing maps, variable selection, decision trees and gradient boosting, linear and logistic regression, two stage modeling, partial least squares, support vector machines and neural networking. Data preparation tools include outlier detection, variable transformations, variable clustering, interactive inning, principal components, rule building and induction, data imputation, random sampling and the partitioning of datasets (into train, test and validate datasets). Advanced visualization tools enable you to quickly and easily examine large amounts of data in multidimensional histograms and to graphically compare modeling results.

Our study utilizes the techniques provided by SAS Enterprise Miner for analysis of the data collected through user surveys. Datasets can be exported into process flow diagrams. The exported datasets can be partitioned into training, test and validation sets. The training dataset can be used for preliminary modeling. Validation and test datasets can be used for estimation and assessment of the model. Filters can be created and applied to each of the datasets to exclude certain observations like errant data or extreme outliers. Metadata about the input data can be found in the input data source node. Enterprise Miner provides association rules to identify association relationships within the data. Association algorithms help to discover sequences in the data that are based on certain patterns.

Clustering techniques segment the data so that data observations that are similar in some way can be identified. When displayed in a plot, observations that are similar tend to be in the same cluster and observations that are different tend to be in different clusters. A

cluster identifier for each observation can be used as a group variable to construct separate models for each group. The Graph Explore node provides an advanced visualization tool that can be used to explore large volumes of data graphically to uncover patterns and trends and reveal extreme values in the datasets. The graph plots are fully interactive and can be rotated or moved to get different angles or perspectives on the data. The Enterprise Miner provides a path analysis tool to analyze web log data and to determine paths that visitors take as they navigate through a website. This tool can also be used for sequence analysis. The StatExplore tool can be used to compute the distribution statistics and correlation statistics for the data. Our study uses a combination of some of the above tools provided by SAS Enterprise Miner to analyze the collected survey data. This analysis aims at discovering and modeling user preferences for the selected subset of factors supposedly impacting the readability of scientific web pages.

2.4 Clustering

Clustering techniques segment the data so that data observations that are similar in some way can be identified. The clustering algorithm decides on some initial cluster seeds depending on the desired number of clusters. Each observation is assigned to strictly one cluster, that cluster and the neighboring clusters are updated. Clusters are represented as circles on a 2-dimensional (2-D) plane, with the radius representing the number of observations assigned to the cluster and also the distribution of the observations within the cluster. Distribution of the observations within the circle is not even. The radius indicates the maximum distance of any observation from the cluster seed. The circles can overlap, but each observation is assigned to only one cluster. The size of a cluster, also called the frequency of the cluster, indicates the number of observations belonging to that cluster and

can suggest the general preference for the variable being evaluated. A cluster represented by just the cluster seed contains only a single observation and could be an outlier [Tan, 2006].

For the current study, the clustering algorithm divides observations into groups depending upon the average scores given across the sample website by each observation. The size and positioning of the clusters on the 2-D plane determines the general preferences of users about readability factors at a high level. Also, the variable analysis within the clustering indicates the important variables that divide the data up into clusters. This is used to find the readability factors that affect readability the most.

CHAPTER 3: METHODOLOGY AND RESEARCH DESIGN

3.1 Introduction

This chapter describes the general methodology used for conducting the research study. It details the kind of user participation used for the surveys. We include the method to operationalize the readability factors on the sample scientific website. The chapter also discusses the survey designed to collect preference data from the users.

3.2 General Method and Participation

The current study evaluates the readability of a typical scientific website based on aesthetic design heuristics, like font and background, by analyzing the survey responses obtained from real users. For this purpose, we include participants from various disciplines of science, e.g., Computer Science, Mathematics, Physics, Chemistry and Biology, at Appalachian State University. Participants have a basic understanding and minimal experience with conventional scientific websites that provide white papers and general information about different topics related to science. However, the participants do not necessarily have familiarity with designing or programming websites. The survey has subjective questions related to the aesthetic design and quality and appearance of the sample website. Participants are not required to answer the website's architecture related questions. The survey participants are undergraduates, graduate students or faculty. Since the survey questions are based only on usability, knowledge level of the participant does not play a significant role in the type of responses. We anticipate that the responses to the questionnaire are influenced by personal aesthetic choices of the individuals and not related to their level of education.

The survey is conducted with the knowledge and approval of the Institutional Review Board (Study # 10 – 0032) at Appalachian State University. We collect survey responses from ninety participants. This target sample size is reasonable enough to draw inferences about the general preferences of users about the factors affecting readability of scientific websites and their ideal values. The proportion of undergraduates, graduate students and faculty participants is not significant. Also, there is a random combination of participants from different departments.

The study and survey were advertised in the selected science-related departments through the respective department chairs. The faculty and students are informed and briefed about their potential involvement in the evaluation survey. We address the questions and concerns in-person and through emails. The first page on the survey is the Informed Consent. All the participants voluntarily agree to participate in the survey by accepting the Informed Consent and indicating their approval for the evaluation procedure. For the Computer Science department specifically, the department mailing lists are used for undergraduates, graduate students and faculty.

The website evaluation and survey are administered online. Some participants access the website and survey from the comfort of their homes or offices. This introduces some error in the responses due to the different sizes of the computer monitors, type of browsers, and software configuration. Most participants use computers on the Appalachian campus in the CAP Science building laboratories for consistent survey procedure. Each participant browses through the sample scientific website and answers twelve survey questions addressing different aesthetic factors affecting readability of the website. The complete process takes about twenty-five minutes including briefing, website navigation and

responding to the survey questionnaire. Each participant takes more time if needed and the submitted responses cannot be changed. The submitted responses are analyzed to confirm the proposed hypotheses.

3.3 Operationalization of Variables

The sample website for readability evaluation could belong to any field of science. We have chosen Astronomy to create a scientific website for the user survey. The sample website about Astronomy includes all the elements that we want to examine to assess the readability of a scientific website. The website contains basic information on various subjects within Astronomy such as the solar system, constellations, meteors and comets. Each of these subjects forms a separate page on the website. On each web page, the content is presented through text, images, tables, graphs and links to other informative resources (e.g., PDF documents, HTML pages). The formatting of the text, tables and graphs impacts readability of the web pages. These can be considered as the readability factors that are evaluated in the current study. Users browse through the website and answer a set of survey questions on each of the selected readability factors.

The selected readability factors have been specified in the following tables, along with the methods used to operationalize the values for those factors. Table 3.1 elaborates the font-related factors and how they have been operationalized. Table 3.2 describes the page scroll related factors and the operationalization methods. Table 3.3 specifies the page justification, background factors and the general presentation of content. Table 3.4 describes the image, graph, mathematical data and web document properties being evaluated.

Table 1 - Font-related Factors

Factor	Operationalization
Font type	Use a combination of various fonts for different pages in the website. Use fonts from serif (e.g. Times New Roman) and sans-serif (e.g. Arial) families for headings, sub-headings and body text.
Font size	Use several font sizes (between 9-point and 14-point) across different pages and across a single page in the website. Demarcate headings, subheadings and content on some pages. Do not make a distinction between the sections of a page for the rest of the pages.
Font color	Use a combination of font colors across a single page on the website. Use different font colors for the different pages on the website. Use conventional browser-safe colors like red, blue, black and purple. Use default colors for hyperlinks.

Table 2 - Page Scroll-related Factors

Factor	Operationalization
Vertical scroll	Have some pages with just one page of information. Some navigation links can have more than one page of data so as to increase the length of the page beyond one screen. Users have to scroll vertically to access all the information on the page.
Horizontal scroll	Have a few pages wider than the maximum width that can fit on a given screen resolution, so that a user will have to scroll horizontally to access all the information on the page.

Table 3 - Page justification, Background and Content

Factor	Operationalization
Page justification	Use a combination of page justifications for the different pages (left, centered, right, justified).
Background	Includes background images for some pages or background colors for few pages and leave the rest of the pages without any background. The information provided by that page should be over the background (in the foreground).
Content	Includes the general format and display of data. Arrangement of labels or captions. Relevance of content and logical positioning of the data (text and images).

Table 4 - Image, Graph, Table, Mathematical Data and Web Document Properties

Factor	Operationalization
Image properties	Use ALT and/ or TITLE to describe images. Use colored images and grayscale images.
Graph properties	Include graphical figures and charts in various formats on the pages. E.g. bar graphs and line graphs. Use different color combinations for representing the data (e.g. yellow to show daytime, black to indicate night time) or colors irrelevant to the data being presented.
Table properties	Include information in tabular format. Use tables with only horizontal lines, only vertical lines or both. Use table headings or notes to explain the table for some of the tables. Use different sizes and fonts for the information represented in the table.
Mathematical formulae	Include mathematical formulae with numbers, symbols, subscripts, superscripts and mathematical operators. Use some formulae as images and type some using HTML tags. Use different fonts and font sizes for these formulae.
File types	Include links to white papers, journal articles and online material. Make sure these resources open in common web document formats like PDF, HTML and Postscript.

The above quantification of readability factors has been extracted from a literature review of similar studies conducted in the past. Ivory, Hearst et al. and Ivory, Megraw et al. made several recommendations regarding the use of font types, colors and sizes, following their evaluation studies of website usability [Ivory, 2002], [Ivory, 2005]. Some of the recommendations included font size between 9-point and 14-point, minimum color usage, at least one sparsely used accent color for navigation bars, high-contrast color combinations, default hyperlink colors and browser-safe colors like red, blue and purple.

ACM and IEEE are well-known organizations and online repositories for journal articles, technical papers, conference proceedings in science and technology. We have tried to base our operationalization of readability factors as per the template specifications provided by each of these organizations for submission of abstracts, papers and articles.

ACM and IEEE specifications included the use of serif fonts like Times New Roman for headings and sans-serif fonts like Verdana for body content, font sizes between 9-point and 14-point and justified page formatting [ACM, 2009], [IEEE, 2009].

Image formats like jpeg, png, bmp, gif and svg may have an impact on a web page. However, from a user's perspective, the different formats may only affect download speeds or bandwidth usage, but not the appearance or readability. A grayscale image can provide better contrast than a colored image, which can contribute towards better readability of a page. However, a colored image may present information more clearly and distinctly than a grayscale image (e.g. the composition of the sun). Also, the use of the ALT and/ or TITLE attributes to describe images can improve their readability and understandability.

Popular web document formats include PDF, HTML and Postscript. Traditional formats for documents on the internet include Envoy, Common Digital Paper, Farallon and Replica. PDF is the de facto standard for printing on the web currently. Consequently, PDF, HTML and Postscript are the formats that we evaluated for readability.

3.4 Instrumentation

The survey captured the opinions and preferences of scientific website users in general with regard to the potential factors that impact the readability of such websites. We used a consistent format of rating for the responses to all the survey questions so that the analysis could yield significant trends of preference data. Below is a sample survey questionnaire form that requires the user to first navigate through the entire website. We limited the number of questions to one per factor, so that the survey would not take much time and effort for a user to complete. In general, users were asked to rate each page on the website with respect to each selected readability factor.

- Font type

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of legibility and appearance of the font types used. Users can choose Not Applicable (NA) if this factor is not present on a page.

- Font size

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of use and arrangement of the font sizes. Users can choose Not Applicable (NA) if this factor is not present on a page.

- Font color

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of attractiveness and appearance of the font colors used. Users can choose Not Applicable (NA) if this factor is not present on a page.

- Scrolling

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of your preference for the amount of information on each page (e.g. vertical scrolling or horizontal scrolling required to access all the content). Users can choose Not Applicable (NA) if this factor is not present on a page.

- Page justification

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of your preferred page formatting (e.g. left, right, center or justified page content). Users can choose Not Applicable (NA) if this factor is not present on a page. Users can choose Not Applicable (NA) if this factor is not present on a page.

- Image properties

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of image formatting (e.g. relevancy of images, grayscale vs. colored images, captions on images). Users can choose Not Applicable (NA) if this factor is not present on a page.

- Background

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of your preference for the page background (e.g. blank, background image, background color, darker or lighter background). Users can choose Not Applicable (NA) if this factor is not present on a page.

- Graph properties

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of appearance, formatting and presentation of graphical data. Users can choose Not Applicable (NA) if this factor is not present on a page.

- Table properties

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of appearance, formatting and presentation of tabular data. Users can choose Not Applicable (NA) if this factor is not present on a page.

- Mathematical formulae

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of appearance, formatting, legibility and presentation of mathematical data. Users can choose Not Applicable (NA) if this factor is not present on a page.

- Article formats

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of your preference and legibility of the formatting used for the web documents on each page. Users can choose Not Applicable (NA) if this factor is not present on a page.

- Content

Rate each page on a scale of 1 to 5 (5 being the highest) on the basis of general arrangement and presentation data, relevancy and logical positioning of the information. Users can choose Not Applicable (NA) if this factor is not present on a page.

Survey Monkey was used to create and publish custom surveys [Survey Monkey, 2009]. Survey Monkey offers a wide range of questionnaire templates, along with an option to personalize the question patterns. The survey was administered online. Participants navigate through the sample website, then visit the Survey Monkey website, login and respond to the survey questionnaire. Users have two parallel screens so that they can have the website available as they respond to the survey questions. Figures 1 and 2 provide screenshots of the survey for the current study from Survey Monkey:

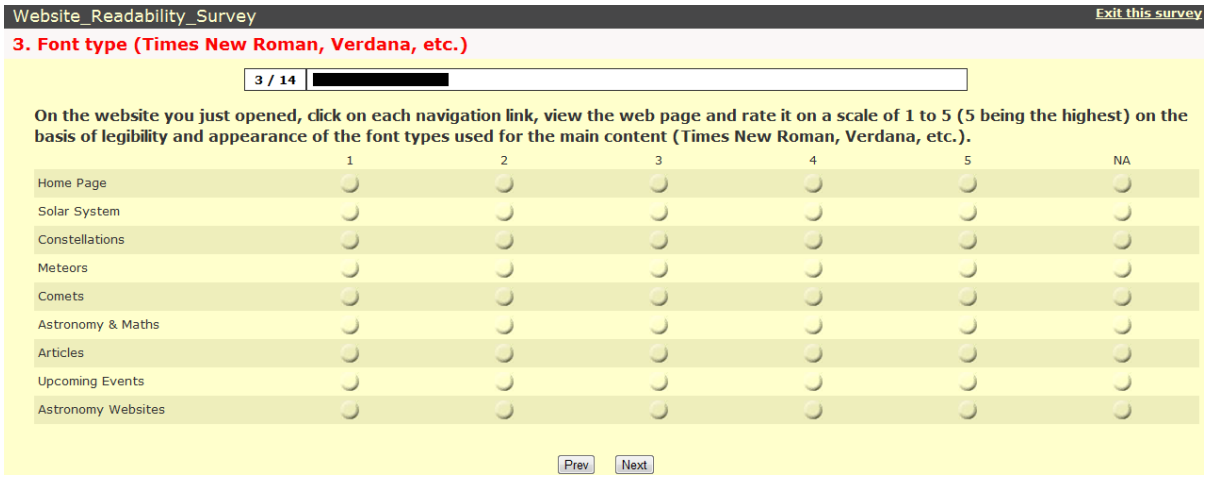


Figure 1 - Survey Question on Font Type

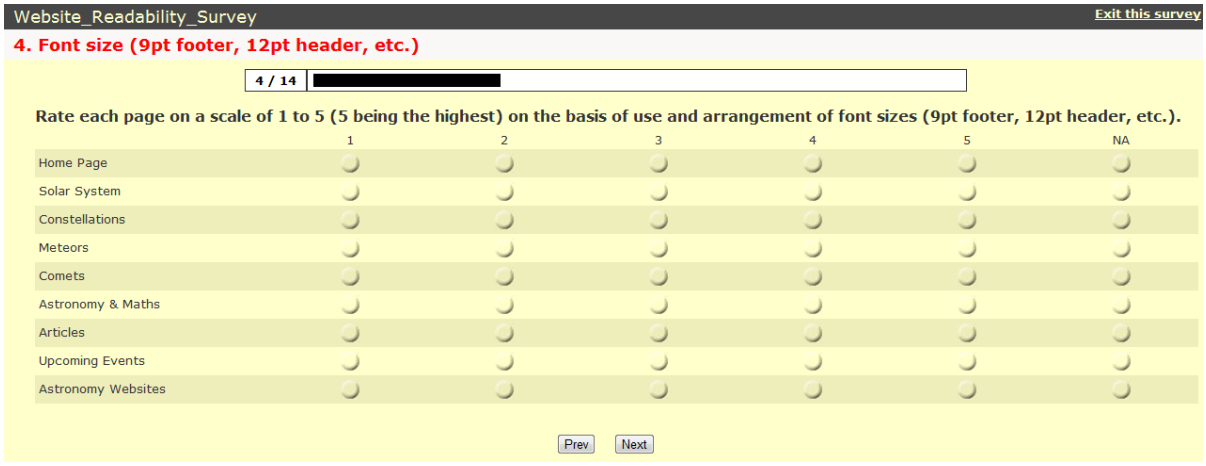


Figure 2 - Survey Question on Font Size

Responses collected by Survey Monkey can be provided to the surveyor in multiple document formats (e.g., PDF, MS Excel). These documents are further analyzed using other intelligent analysis tools. Survey Monkey also presents the survey results in real-time and as graphs and charts for better visualization. The reports are filtered for certain information and can be shared with future research investigators.

3.5 Demographics of Survey Participants

Figure 3 shows the relative number of freshmen, sophomores, juniors, seniors, graduate students and faculty that took the survey. Figure 4 indicates the proportion of participants from various departments.

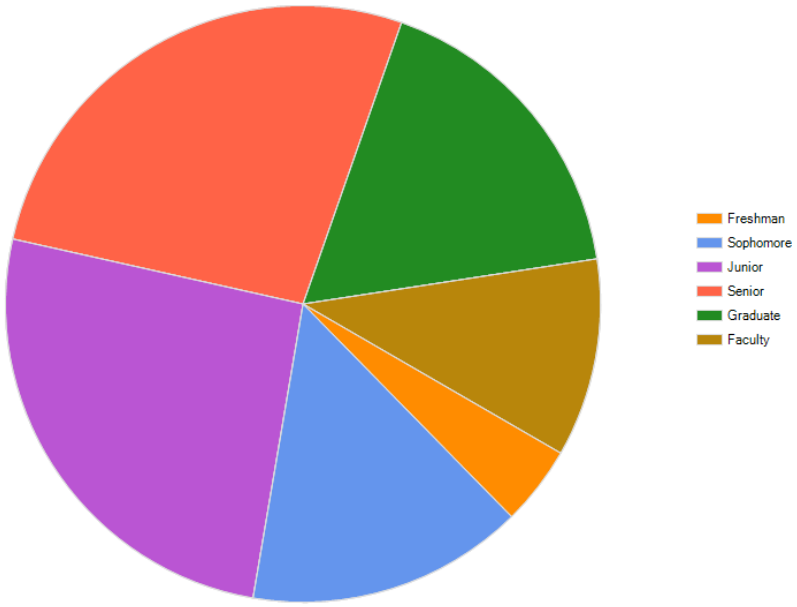


Figure 3 - Relative Number of Participants Based on Experience

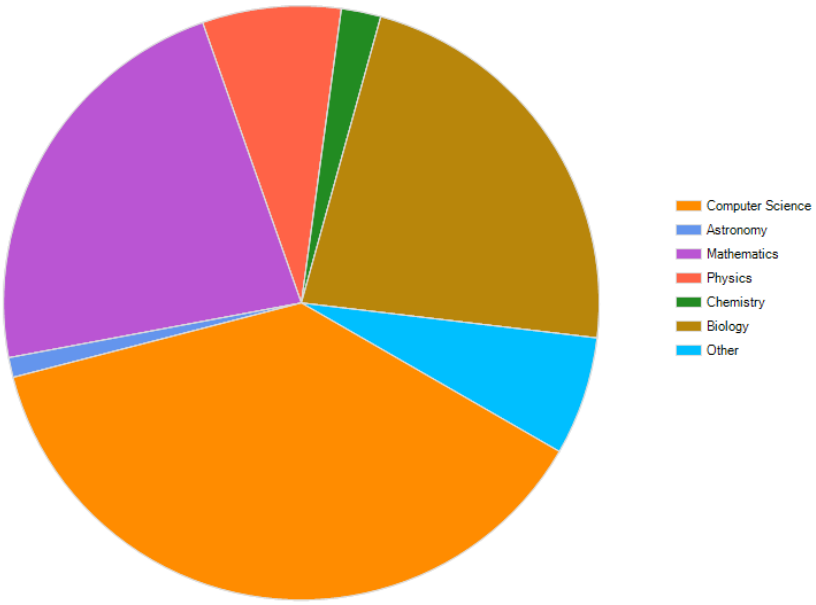


Figure 4 - Proportion of Participants Based on Department

3.6 Overall Procedure

The overall procedure is summarized in Figure 5:

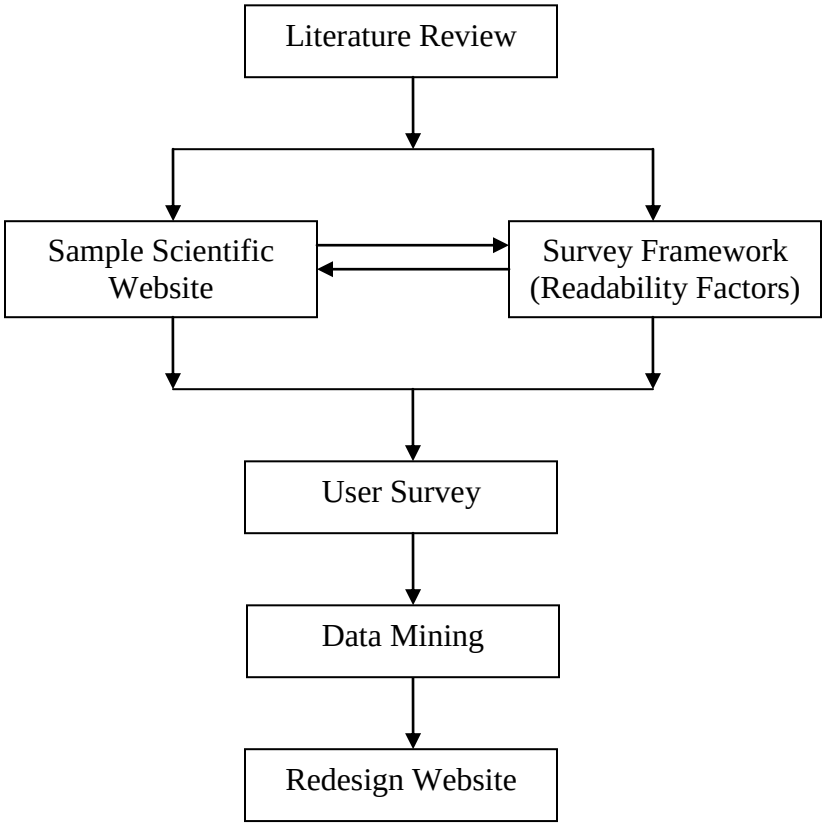


Figure 5 - Overall Research Design

CHAPTER 4: ANALYSES AND RESULTS

This chapter describes the procedures to test the hypotheses and the results of the analysis.

4.1 Data

For this study, we chose to download detailed responses from Survey Monkey in excel format. Excel spreadsheet columns contain the questions and relevant web pages and rows contain values entered by users. The number of rows indicates the number of survey participants. The downloaded reports are imported into SAS Enterprise Miner in the form of SAS datasets. We have created one dataset for each readability factor being evaluated. Each dataset contains a table with rows indicating the web pages being evaluated and columns containing the values entered by the participants for each web page. These datasets have been saved in the SASUSER library within SAS Enterprise Miner. Table 4.1 is the sample dataset for font type readability factor:

Table 5 - SAS Dataset for Font Type Factor

Home Page	Solar System	Constellations	Meteors	Comets	Astronomy & Math	Articles	Upcoming Events	Astronomy Websites
5	5	5	5	5	5	5	5	5
4	3	3	3	3	2	3	3	3
4	4	3	4	4	5	5	5	1
4	2	4	2	2	4	4	4	4

4.2 Analysis Method

Ninety user responses were collected and analyzed to get some preferred values of the readability factors using Survey Monkey and SAS Enterprise Miner. Enterprise Miner uses the Self Organizing Maps clustering technique that directly considers relationships between clusters during the clustering process [SAS, 2009]. Assignment of a point to a cluster affects the definition of that cluster and those of the neighboring ones. The cluster proximities graph attempts to find a set of centroids that best approximate the data subject to topographic constraints among the centroids. The algorithm finds clusters that minimize the sum of squared distance of each point from its closest cluster centroid. The points can be standardized before assigning them to the clusters, but since the same rating scale has been used for all the readability factors for this study, no standardization is required.

4.3 Analytical Settings

A data flow diagram in SAS Enterprise Miner indicates the tools or nodes used by our analyses with SAS Enterprise Miner, e.g., Clustering, Distribution Explorer, Multiplot and Reporter. Clustering is the main tool used for the analysis of the preferred readability factors. For each factor, the clustering node uses the web pages as variables. The level of measurement for each variable is "ordinal" and the model role is "input." The clustering method used is "centroid" with a clustering cubic criterion cutoff of 3. The minimum number of clusters is set to 2 and the maximum number of clusters is 5. The maximum number of clusters is low since we have only 90 responses.

Below are some of the nodes used and generated for each factor analysis in the current study [SAS, 2009]:

This node represents the selection of the dataset on which the analysis is performed. The data imported into Enterprise Miner is saved in the form of a SAS dataset.

This node represents exclusion of unacceptable values and outlier values from consideration for the analysis.

This node indicates the basic criteria for clustering of the observations in the included dataset.

This node represents an advanced visualization tool that enables exploring large volumes of data graphically. The tool can be used to uncover patterns and trends to reveal extreme values in the database.

Multiplot node provides functionality to observe data distributions and examine the relationship between variables.

Reporter node represents results from an Enterprise Miner process flow into an HTML report that can be viewed with a web browser.

4.4 Preferred Font Types

Figure 6 shows the dataflow diagram for preferred font type analysis indicating the nodes from SAS Enterprise Miner that are used for this analysis. SASUSER.FONTTYPENEW is the input dataset that has the user ratings for font type analysis. Filter Outliers node represents exclusion of unacceptable values from the input dataset. Clustering node indicates grouping of the acceptable values from the input dataset. Figure 7 shows the inappropriate values that have been filtered out from consideration for clustering. In this case, each web page in the website is being evaluated for font type. So, if a web page has a score of 6 for font type, then that is excluded through this node. For every web page, the acceptable values for font type are 1 through 5.

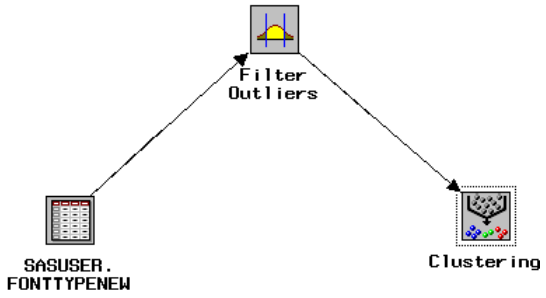


Figure 6 - Data Flow Diagram for Font Type Analysis

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	6
SOLAR_SYSTEM	Solar System	0	6
CONSTELLATIONS	Constellations	0	6
METEORS	Meteors	0	6
COMETS	Comets	0	6
ASTRONOMY___MATHS	Astronomy & Maths	0	6
ARTICLES	Articles	0	6
UPCOMING_EVENTS	Upcoming Events	0	6
ASTRONOMY_WEBSITES	Astronomy Websites	0	6

Figure 7 - Filtered Values for Clustering

Figure 8 shows a 3-D pie chart indicating the 5 clusters, into which the observations are grouped. The color of each slice indicates how close the observations are to the respective cluster seed. The green color indicates the frequency of each slice, i.e. the number of observations belonging to each cluster. From Figure 8, clusters 2 and 5 have the maximum frequency. Each cluster in the pie chart can be selected to view the normalized values of observations in that cluster compared to an average across all the other clusters in the input grid plot shown in Figure 9. The input grid plot profiles each cluster by identifying the standard input variables that are significantly different from overall mean. These input variables best characterize the corresponding cluster. Figure 9 compares normalized values of cluster 5 with those of all the remaining clusters. Observations belonging to cluster 5 have higher normalized values as compared to rest of the clusters. Distance plot in Figure 10 indicates the positioning of cluster 5 relative to the other clusters and its size in a 2-dimensional space.

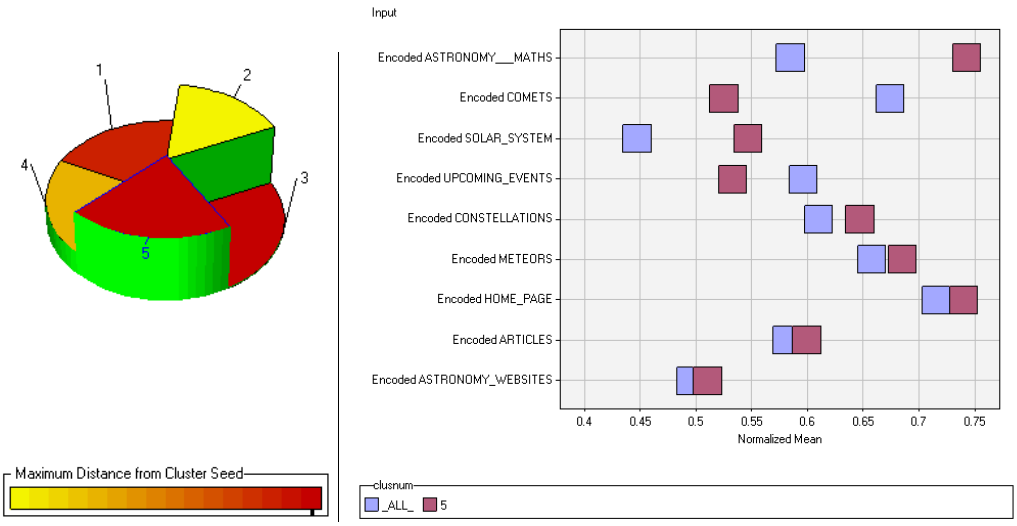


Figure 8 - Cluster Pie Chart (L)

Figure 9 - Input Means Plot (R)

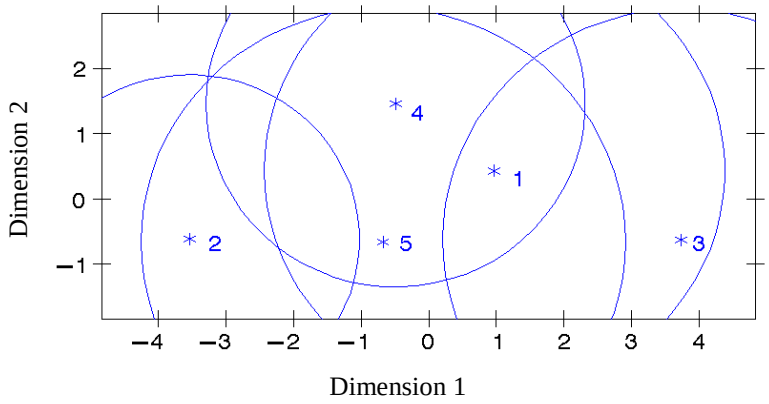


Figure 10 - Distance Plot for Clusters

From the statistics plot in Figure 11, cluster 2 has a frequency of 24 and cluster 5 has a frequency of 25.

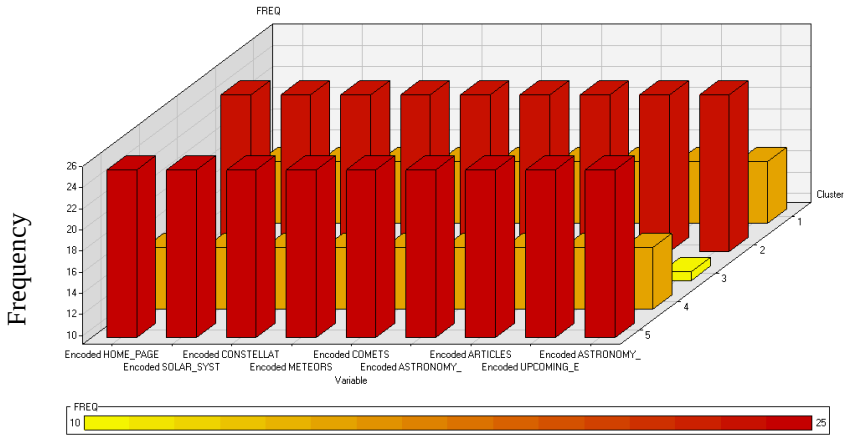


Figure 11 - Statistics Plot for Clusters

Since cluster 5 has the maximum frequency of observations and also overlaps with most other clusters, we consider the average ratings for pages in cluster 5 for determining the preferences of most users with respect to font types. Table 6 displays these average scores. From the input grid plot, we already know that the normalized values for observations in cluster 5 are mostly higher than other clusters. Astronomy & Math, Home Page and Articles were the top three most popular web pages on the site with respect to font types. Consequently, Helvetica was the most preferred font type for headings/ sub headings. Georgia, Arial and Verdana were the preferred font types for content.

Table 6 - Average Score for Web Pages from Cluster 5

Web Page	Average Score
Home Page	3.84
Solar System	3.04
Constellations	3.48
Meteors	3.80
Comets	3.16
Astronomy & Math	4.12
Articles	3.84
Upcoming Events	3.32
Astronomy Websites	3.44

4.5 Preferred Font Sizes

Figure 12 shows the dataflow diagram for the preferred font size analysis. Figure 13 shows the cluster pie chart. Cluster 2 has the maximum frequency. The input means plot in Figure 14 shows that the observations in cluster 2 have higher normalized values as compared to rest of the clusters. Figure 15 shows the distance plot in 2-dimensional space. From the statistics plot in Figure 16, cluster 2 has a frequency of 26. The average ratings for pages in cluster 2 were used to determine the user preferences with respect to font sizes; Table 7 displays these average scores. Home Page, Articles and Comets pages had the top three scores with respect to font sizes. The most preferred font size for headings was 13-point, 11-12point for sub-headings and 10-11point for content.

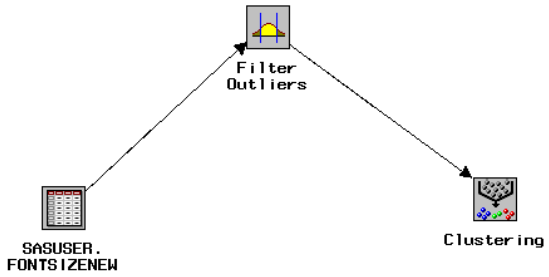


Figure 12 - Data Flow Diagram for Font Size Analysis

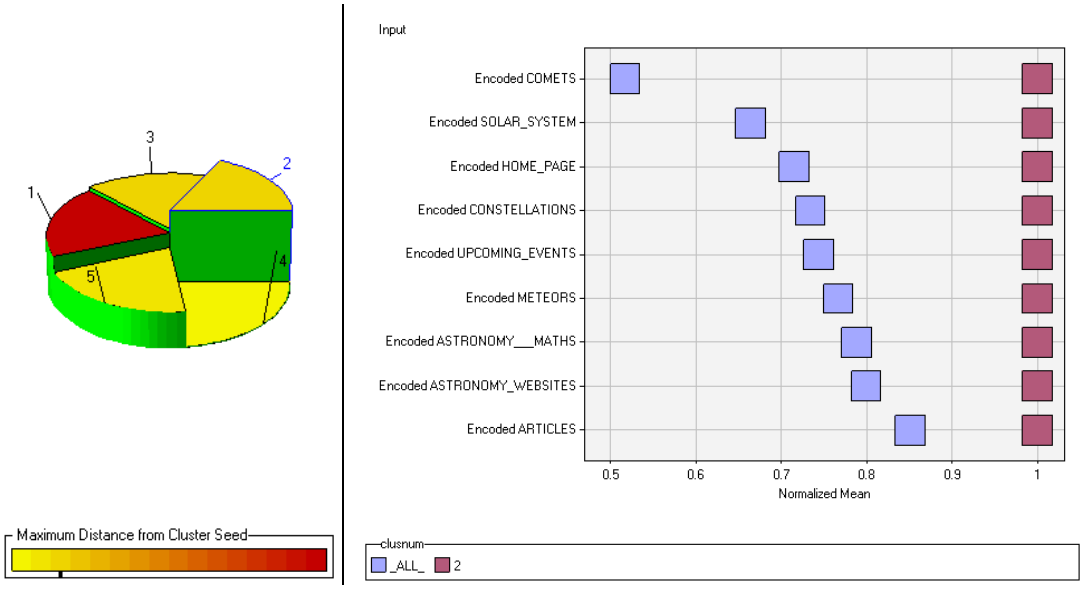


Figure 13 - Cluster Pie Chart (L)

Figure 14 - Input Means Plot (R)

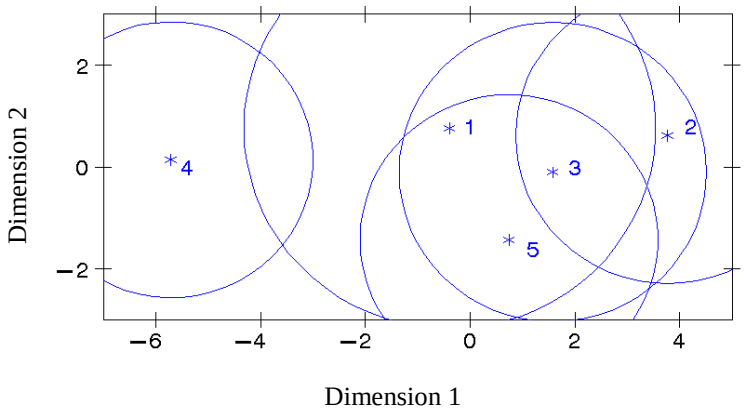


Figure 15 - Distance Plot for Clusters

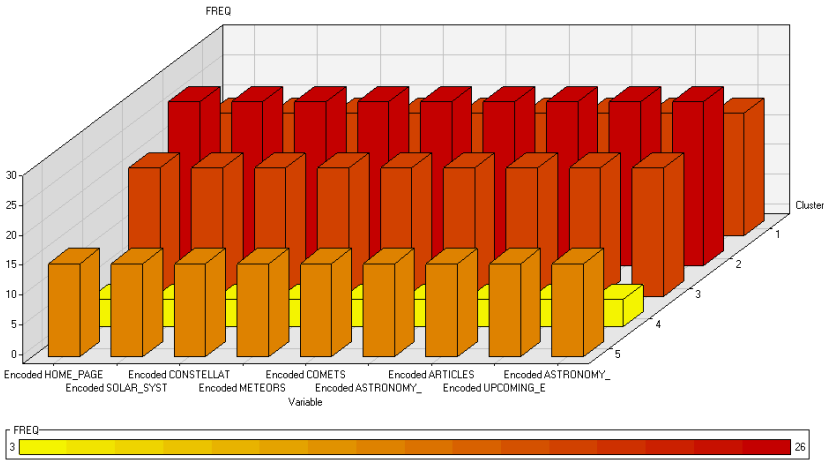


Figure 16 - Statistics Plot for Clusters

Table 7 - Average Score for Web Pages from Cluster 2

Web Page	**Average Score**
Home Page	4.62
Solar System	4.46
Constellations	4.35
Meteors	4.31
Comets	4.58
Astronomy & Math	4.50
Articles	4.62
Upcoming Events	4.38
Astronomy Websites	4.35

4.6 Preferred Font Colors

Figure 17 shows the dataflow diagram for preferred font color analysis. Figure 18 shows the filtered out values. Figure 19 shows a cluster pie chart. Cluster 4 has the maximum frequency. Input means plot in Figure 20 indicates that the normalized values of observations in cluster 4 are lower than the normalized values for the rest of the clusters. This means that the majority of the users have rated the pages low. Figure 21 shows the distance plot in a 2-dimensional space. From the statistics plot in Figure 22, cluster 4 has a frequency of 22. We consider the average ratings for pages in cluster 4 for determining user preferences with respect to font colors. Table 8 displays these average scores. Articles, Home Page and Comets had the top three scores with respect to font colors. Most preferred font color combinations included blue-gray, red-blue-purple and blue-black.

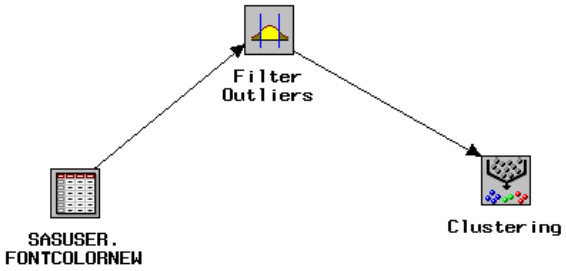


Figure 17 - Data Flow Diagram for Font Color Analysis

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	
SOLAR_SYSTEM	Solar System	0	
CONSTELLATIONS	Constellations	0	
METEORS	Meteors	0	
COMETS	Comets	0	
ASTRONOMY___MATHS	Astronomy & Maths	0	
ARTICLES	Articles	0	
UPCOMING_EVENTS	Upcoming Events	0	6
ASTRONOMY_WEBSITES	Astronomy Websites	0	

Figure 18 - Filtered Values for Clustering

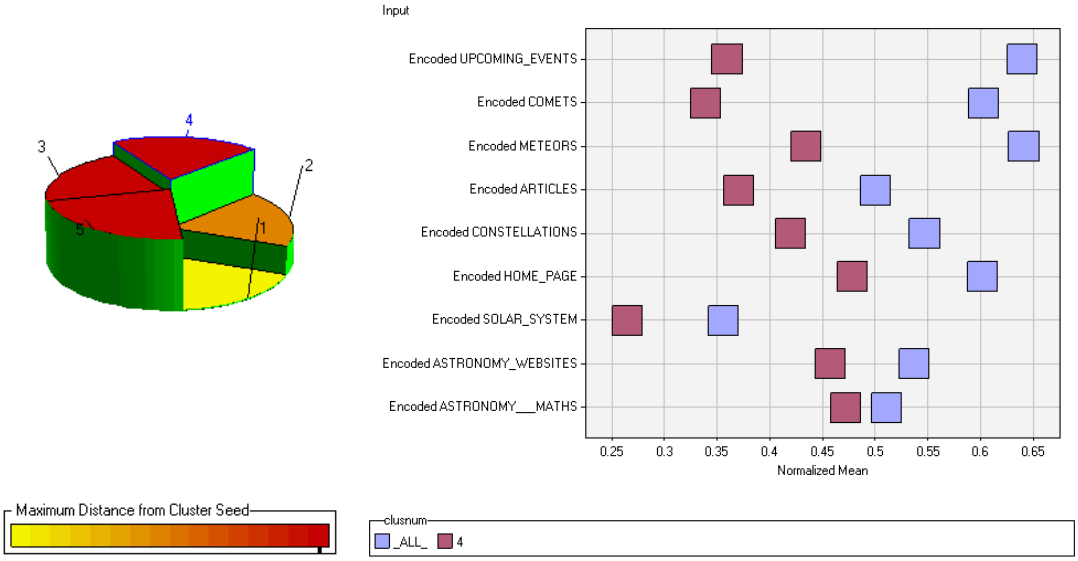


Figure 19 - Cluster Pie Chart (L)

Figure 20 - Input Means Plot (R)

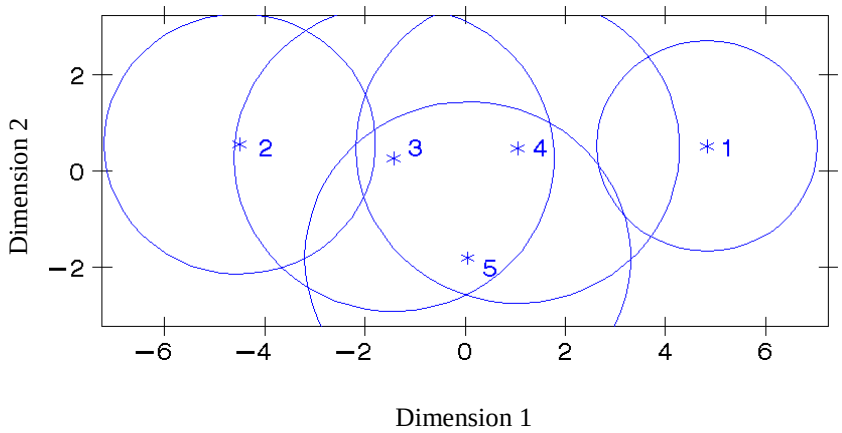


Figure 21 - Distance Plot for Clusters

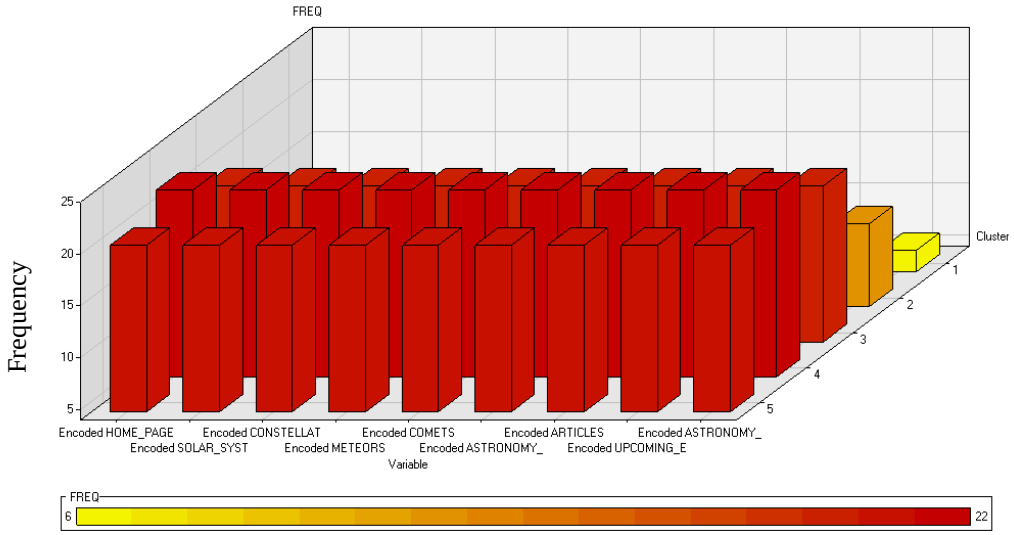


Figure 22 - Statistics Plot for Clusters

Table 8 - Average Score for Web Pages from Cluster 4

Web Page	Average score
Home Page	3.05
Solar System	2.00
Constellations	2.59
Meteors	2.59
Comets	2.82
Astronomy & Math	2.95
Articles	3.09
Upcoming Events	2.55
Astronomy Websites	2.50

4.7 Preferred Page Scrolling

We analyzed the impact of page scroll on the web page readability similar to the analysis performed above for the font factors. Figure 23 shows the dataflow diagram for preferred page scroll analysis. Figure 24 shows the inappropriate values that have been filtered out from consideration for clustering. Figure 25 shows a 3-D pie chart indicating the 5 clusters, into which the observations are grouped. Cluster 5 has the maximum frequency.

Figure 26 compares values of cluster 5 with all the remaining clusters. Figure 27 shows the distance plot and Figure 27 shows the statistics plot for the clusters.

We considered the average values for cluster 5, which has the maximum frequency of 22. Table 9 shows these average values. Articles, Home Page and Astronomy Websites pages had the top three scores with respect to page scroll. None of these pages needed to be scrolled to see the complete content. This indicated that users did not wish to scroll vertically or horizontally to view a complete web page. The Comets page had the worst score indicating horizontal scrolling was the least preferred.

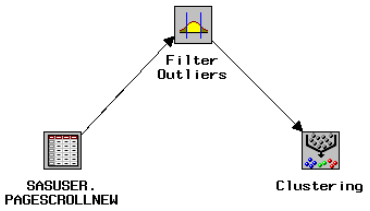
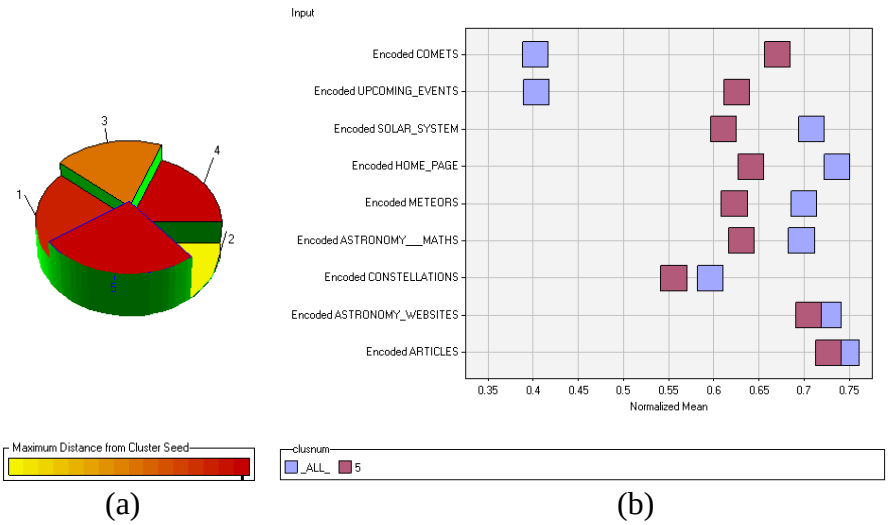


Figure 23 - Data Flow Diagram for Page Scroll Analysis

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	6
SOLAR_SYSTEM	Solar System	0	
CONSTELLATIONS	Constellations	0	
METEORS	Meteors	0	
COMETS	Comets	0	
ASTRONOMY___MATHS	Astronomy & Maths	0	
ARTICLES	Articles	0	
UPCOMING_EVENTS	Upcoming Events	0	6
ASTRONOMY_WEBSITES	Astronomy Websites	0	

Figure 24 - Filtered Values for Clustering



(a) (b)

Figure 25 - Cluster Pie Chart (L)

Figure 26 - Input Means Plot (R)

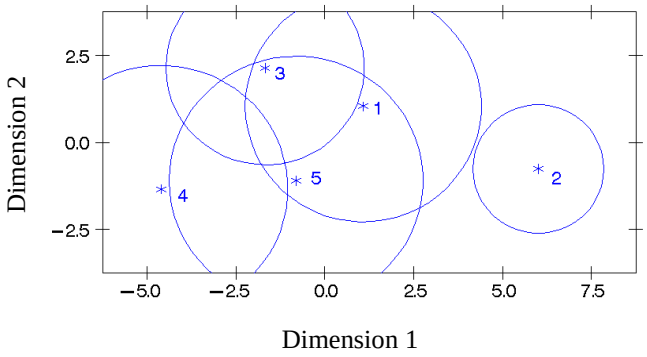


Figure 27 - Distance Plot for Clusters

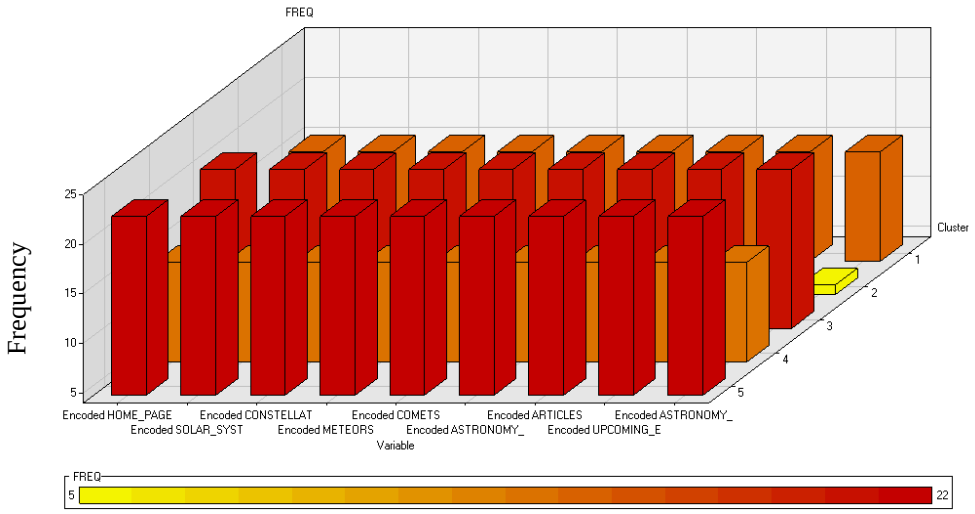


Figure 28 - Statistics Plot for Clusters

Table 9 - Average Score for Web Pages from Cluster 5

Web Page	Average Score
Home Page	3.73
Solar System	3.59
Constellations	3.18
Meteors	3.59
Comets	2.86
Astronomy & Math	3.45
Articles	3.91
Upcoming Events	3.23
Astronomy Websites	3.82

4.8 Preferred Page Justification

Figure 29 shows the dataflow diagram for preferred page justification analysis. Figure 30 shows the inappropriate values that have been filtered out from consideration for clustering. Figure 31 shows a 3-D pie chart indicating the 5 clusters, into which the observations are grouped. Cluster 2 has the maximum frequency. Figure 32 compares values

of cluster 2 with all the remaining clusters. Figure 33 shows the distance plot for the clusters and the statistics plot in Figure 34 indicates cluster 2 has a frequency of 27.

We considered cluster 2 with a frequency of 27 for the analysis of preferred page justification readability factor. Table 10 shows the average values for cluster 2 for determining the user preference for page justification. Articles, Astronomy Websites and Home Page were the pages with top average scores with respect to page justification. This indicated that users prefer left justification most. Constellations page, which was right-justified had the minimum score.

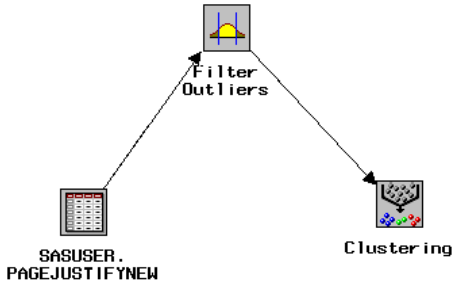


Figure 29 - Data Flow Diagram for Page Justification Analysis

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	
SOLAR_SYSTEM	Solar System	0	
CONSTELLATIONS	Constellations	0	
METEORS	Meteors	0	
COMETS	Comets	0	6
ASTRONOMY___MATHS	Astronomy & Maths	0	
ARTICLES	Articles	0	
UPCOMING_EVENTS	Upcoming Events	0	
ASTRONOMY_WEBSITES	Astronomy Websites	0	

Figure 30 - Filtered Values for Clustering

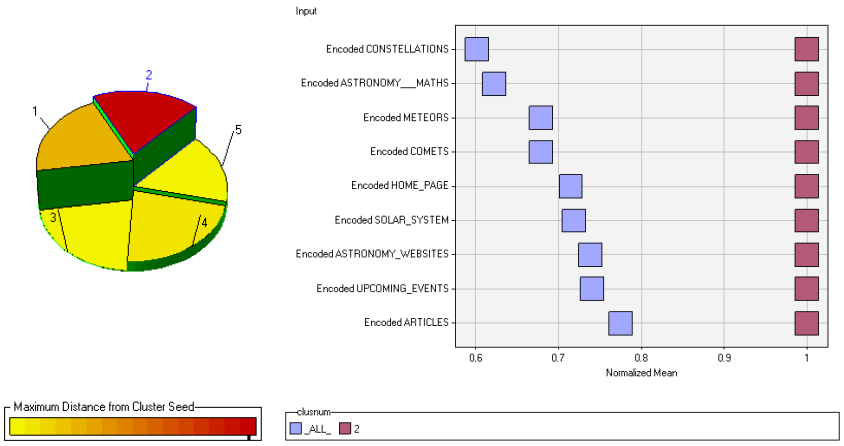


Figure 31 - Cluster Pie Chart (L)

Figure 32 - Input Means Plot (R)

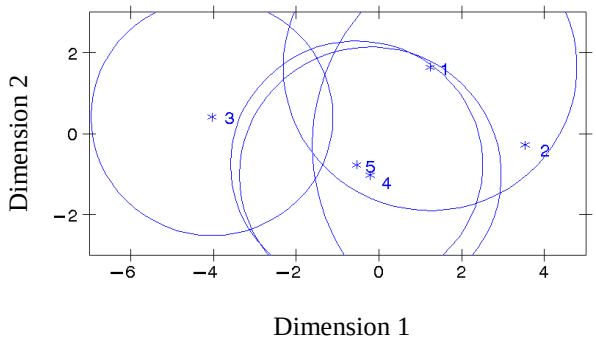


Figure 33 - Distance Plot for Clusters

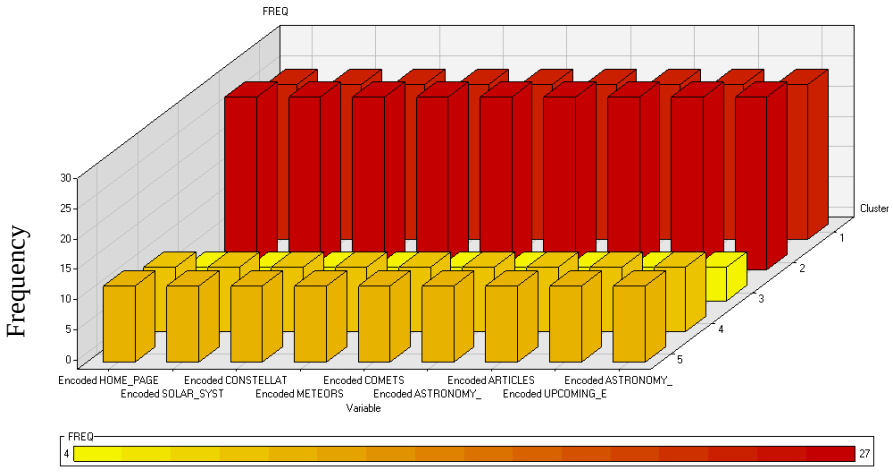


Figure 34 - Statistics Plot for Clusters

Table 10 - Average Score for Web Pages from Cluster 2

Web Page	**Average Score**
Home Page	4.63
Solar System	4.56
Constellations	3.41
Meteors	4.56
Comets	4.26
Astronomy & Math	4.04
Articles	4.67
Upcoming Events	4.41
Astronomy Websites	4.67

4.9 Preferred Image Properties

The pages on the website with images were considered for the analysis of readability with respect to image properties. Figure 35 shows the dataflow diagram for preferred image properties analysis. Figure 36 shows the inappropriate values that have been filtered out from consideration for clustering. Figure 37 shows the pages that are being evaluated for image properties. The status column in the table indicates the web pages on that are being used in this analysis. Figure 38 shows a 3-D pie chart indicating the 5 clusters, into which the

observations are grouped. Cluster 2 has the maximum frequency. Figure 39 compares values of cluster 2 with all the remaining clusters. Figure 40 shows the distance plot for the clusters and Figure 41 shows the statistics plot.

Table 11 shows the average values from cluster 2, which has the maximum frequency of 23. Astronomy & Math, Home Page and Solar System had the highest scores on the site with respect to image properties. This indicated that users preferred grayscale images over colored images. Also, scores indicated that users liked to see captions with source and other information about images.

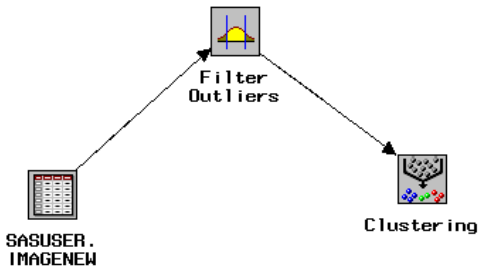


Figure 35 - Data Flow Diagram for Analysis of Image Properties

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	
SOLAR_SYSTEM	Solar System	0	
CONSTELLATIONS	Constellations	0	
METEORS	Meteors	0	7
COMETS	Comets	0	
ASTRONOMY___MATHS	Astronomy & Maths	0	7
ARTICLES	Articles	0	
UPCOMING_EVENTS	Upcoming Events	0	
ASTRONOMY_WEBSITES	Astronomy Websites	0	

Figure 36 - Filtered Values for Clustering

Name	Status	Model Role	Measurement	Type	Format	Label
HOME_PAGE	use	input	ordinal	num	BEST12.	Home Page
SOLAR_SYSTEM	use	input	ordinal	num	BEST12.	Solar System
CONSTELLATIONS	use	input	ordinal	num	BEST12.	Constellations
METEORS	don't use	input	ordinal	num	BEST12.	Meteors
COMETS	use	input	ordinal	num	BEST12.	Comets
ASTRONOMY___MATHS	use	input	ordinal	num	BEST12.	Astronomy & Maths
ARTICLES	don't use	input	ordinal	num	BEST12.	Articles
UPCOMING_EVENTS	don't use	input	ordinal	num	BEST12.	Upcoming Events
ASTRONOMY_WEBSITES	don't use	input	ordinal	num	BEST12.	Astronomy Websites

Figure 37 - Web Pages being Evaluated for Image Properties

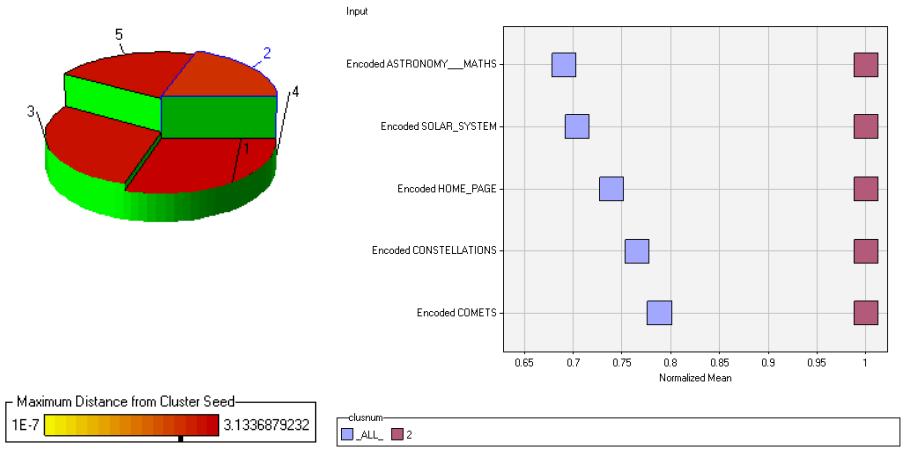


Figure 38 - Cluster Pie Chart (L)

Figure 39 - Input Means Plot (R)

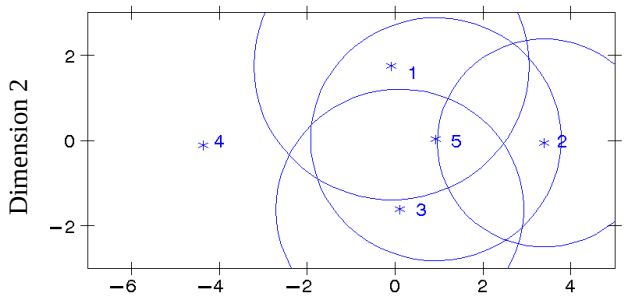


Figure 40 - Distance Plot for Clusters

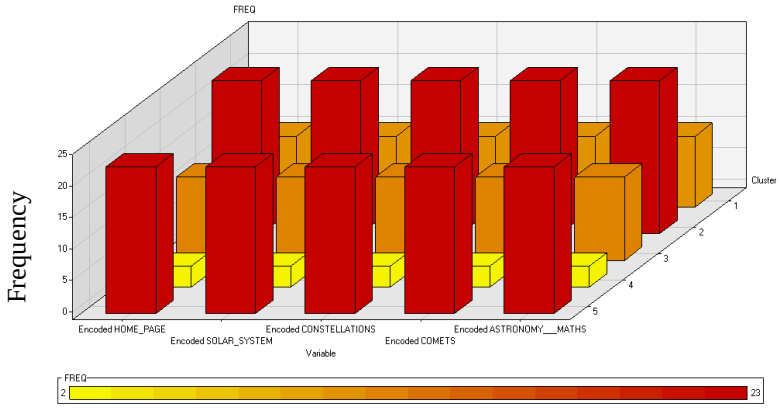


Figure 41 - Statistics Plot for Clusters

Table 11 - Average Score for Web Pages from Cluster 2

Web Page	Average Score
Home Page	4.57
Solar System	4.48
Constellations	4.35
Comets	4.17
Astronomy & Math	4.74

4.10 Preferred Background

We consider the average values for cluster 2 for determining the user preferences for background properties. Figures 42 and 43, respectively, illustrate the dataflow diagram for preferred background analysis and the inappropriate values that have not been considered in clustering. Figure 44 shows a 3-D pie chart indicating the 5 clusters, into which the observations are grouped. Cluster 2 has the maximum frequency. Figure 45 compares values of cluster 2 with all the remaining clusters. Figure 46 shows the distance plot for the clusters and Figure 47 shows the statistics plot.

Cluster 2 has the maximum frequency of 27. Table 12 shows the average scores for the web pages with respect to background properties. Comets, Upcoming Events and Home

Page were rated highest with respect to background. All of these pages had a plain white background. Light colored backgrounds and backgrounds with simple images were the next most preferred.

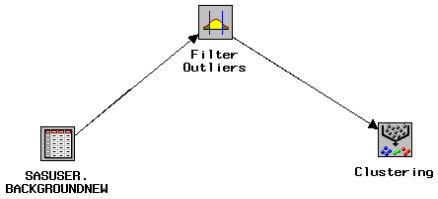
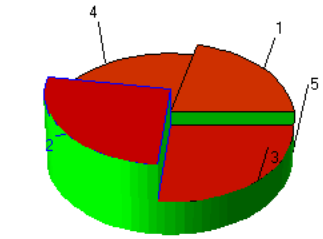


Figure 42 - Data Flow Diagram for Background Analysis

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	6
SOLAR_SYSTEM	Solar System	0	
CONSTELLATIONS	Constellations	0	
METEORS	Meteors	0	
COMETS	Comets	0	6
ASTRONOMY___MATHS	Astronomy & Maths	0	
ARTICLES	Articles	0	
UPCOMING_EVENTS	Upcoming Events	0	6
ASTRONOMY_WEBSITES	Astronomy Websites	0	

Figure 43 - Filtered Values for Clustering



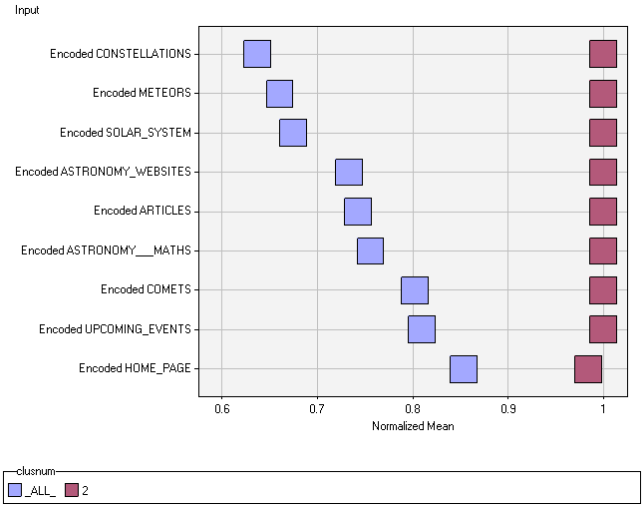


Figure 44 - Cluster Pie Chart (L)

Figure 45 - Input Means Plot (R)

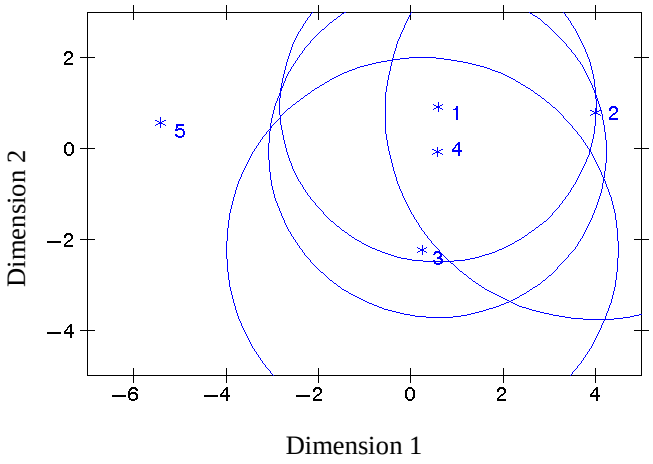


Figure 46 - Distance Plot for Clusters

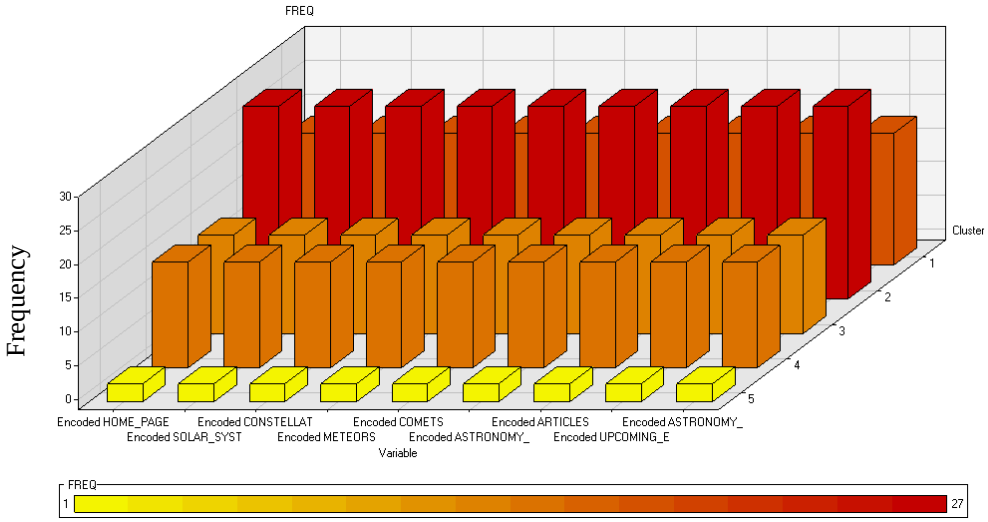


Figure 47 - Statistics Plot for Clusters

Table 12 - Average Score for Web Pages from Cluster 2

Web Page	Average Score
Home Page	4.48
Solar System	3.19
Constellations	3.81
Meteors	3.96
Comets	4.63
Astronomy & Math	4.15
Articles	4.33
Upcoming Events	4.63
Astronomy Websites	3.78

4.11 Preferred Graph Properties

Considering the average scores from cluster 1 with a maximum frequency of 19, we determine the user preferences for graph properties. Figure 48 shows the dataflow diagram for preferred graph properties analysis. Figure 49 shows the inappropriate values that have been filtered out from consideration for clustering. Figure 50 shows the pages that are being evaluated for graph properties. The status column in the table indicates the web pages on that are being used in this analysis. Figure 51 shows a 3-D pie chart indicating the 5 clusters, into which the observations are grouped. Cluster 1 has the maximum frequency. Figure 52 compares values of cluster 1 with all the remaining clusters. Figure 53 gives the distance plot for the clusters and Figure 54 shows the statistics plot.

Table 13 shows the average values from cluster 1. Meteors and Comets were the top two most popular web pages on the site with respect to graph properties. This indicated that users preferred simple bar or line graphs. Use of colors to indicate different values helped improve the readability of a graph. Complicated combination graphs and pie charts were not very popular as per the survey analysis.

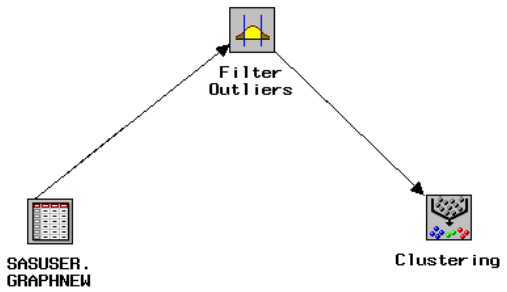


Figure 48 - Data Flow Diagram for Analysis of Graph Properties

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	6
SOLAR_SYSTEM	Solar System	0	6
CONSTELLATIONS	Constellations	0	
METEORS	Meteors	0	
COMETS	Comets	0	6
ASTRONOMY___MATHS	Astronomy & Maths	0	
ARTICLES	Articles	0	
UPCOMING_EVENTS	Upcoming Events	0	
ASTRONOMY_WEBSITES	Astronomy Websites	0	

Figure 49 - Filtered Values for Clustering

Name	Status	Model Role	Measurement	Type	Format	Label
HOME_PAGE	use	input	ordinal	num	BEST12.	Home Page
SOLAR_SYSTEM	use	input	ordinal	num	BEST12.	Solar System
CONSTELLATIONS	don't use	input	ordinal	num	BEST12.	Constellations
METEORS	use	input	ordinal	num	BEST12.	Meteors
COMETS	use	input	ordinal	num	BEST12.	Comets
ASTRONOMY___MATHS	don't use	input	ordinal	num	BEST12.	Astronomy & Maths
ARTICLES	don't use	input	ordinal	num	BEST12.	Articles
UPCOMING_EVENTS	don't use	input	ordinal	num	BEST12.	Upcoming Events
ASTRONOMY_WEBSITES	don't use	input	ordinal	num	BEST12.	Astronomy Websites

Figure 50 - Web Pages being Evaluated for Graph Properties

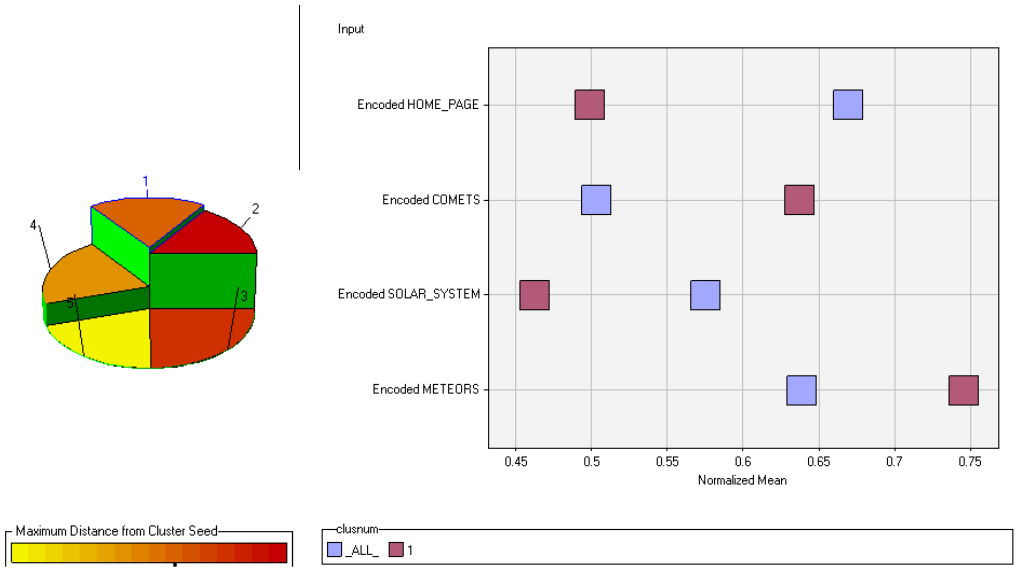


Figure 51 - Cluster Pie Chart (L)

Figure 52 - Input Means Plot (R)

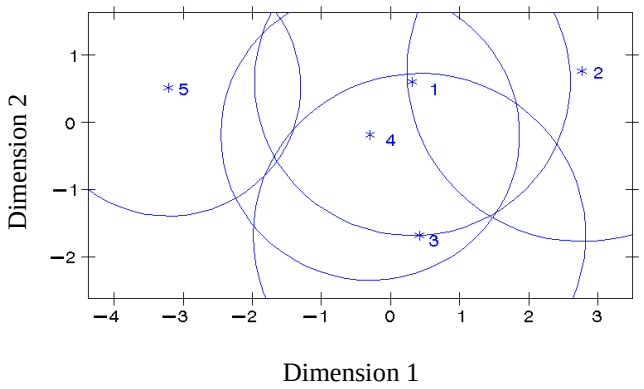


Figure 53 - Distance Plot for Clusters

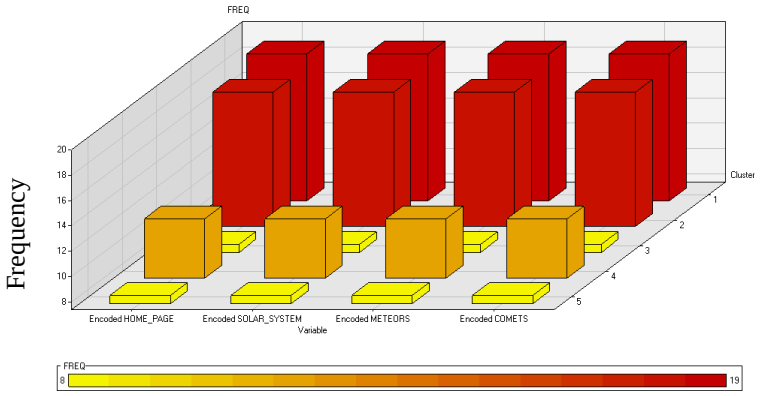


Figure 54 - Statistics Plot for Clusters

Table 13 - Average Score for Web Pages from Cluster 1

Web Page	Average Score
Home Page	2.89
Solar System	2.79
Meteors	3.89
Comets	3.53

4.12 Preferred Table Properties

Cluster 2 with maximum frequency of 24 is used to evaluate the user preference for table properties. Figure 55 shows the dataflow diagram for preferred table properties analysis. Figure 56 shows the inappropriate values that have been filtered out from consideration for clustering. Figure 57 shows the pages that are being evaluated for table properties. The status column in the table indicates the web pages on that are being used in this analysis. Figure 58 shows a 3-D pie chart indicating the 5 clusters, into which the observations are grouped. Cluster 2 has the maximum frequency. Figure 59 compares values of cluster 2 with all the remaining clusters. Figure 60 shows the distance plot for the clusters and Figure 61 gives the statistics plot.

Table 14 shows the average values for the web pages from cluster 2 to determine the user preferences for table properties. Constellations and Astronomy & Math were the pages with the top two scores with respect to table properties. This indicated that users preferred tables with different records separated by horizontal rules or tables where each cell was completely bordered. Tables with just outer boundary or those with only vertical rules were least preferred.

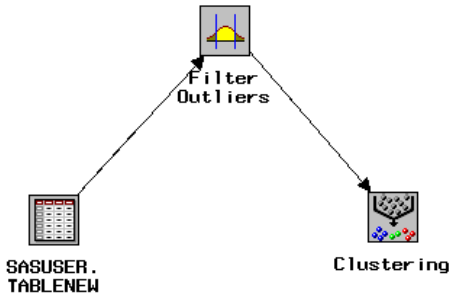


Figure 55 - Data Flow Diagram for Analysis of Table Properties

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	
SOLAR_SYSTEM	Solar System	0	6
CONSTELLATIONS	Constellations	0	6
METEORS	Meteors	0	
COMETS	Comets	0	6
ASTRONOMY___MATHS	Astronomy & Maths	0	6
ARTICLES	Articles	0	
UPCOMING_EVENTS	Upcoming Events	0	
ASTRONOMY_WEBSITES	Astronomy Websites	0	

Figure 56 - Filtered Values for Clustering

Name	Status	Model Role	Measurement	Type	Format	Label
HOME_PAGE	don't use	input	ordinal	num	BEST12.	Home Page
SOLAR_SYSTEM	use	input	ordinal	num	BEST12.	Solar System
CONSTELLATIONS	use	input	ordinal	num	BEST12.	Constellations
METEORS	don't use	input	ordinal	num	BEST12.	Meteors
COMETS	use	input	ordinal	num	BEST12.	Comets
ASTRONOMY___MATHS	use	input	ordinal	num	BEST12.	Astronomy & Maths
ARTICLES	don't use	input	ordinal	num	BEST12.	Articles
UPCOMING_EVENTS	don't use	input	ordinal	num	BEST12.	Upcoming Events
ASTRONOMY_WEBSITES	don't use	input	ordinal	num	BEST12.	Astronomy Websites

Figure 57 - Web Pages being Evaluated for Table Properties

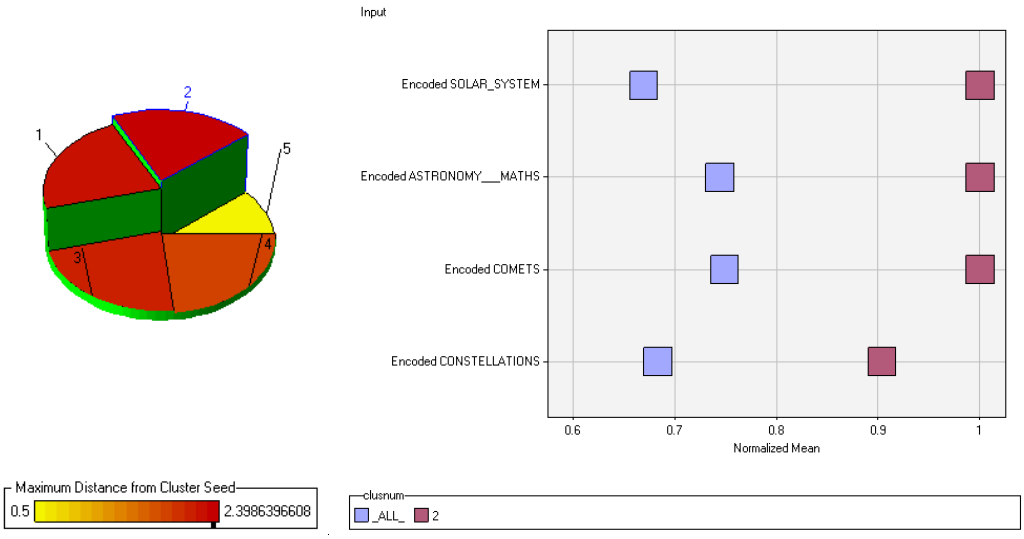


Figure 58 - Cluster Pie Chart (L)

Figure 59 - Input Means Plot (R)

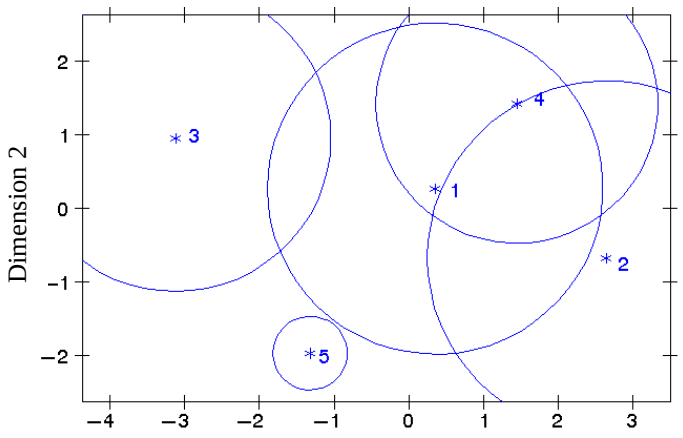


Figure 60 - Distance Plot for Clusters

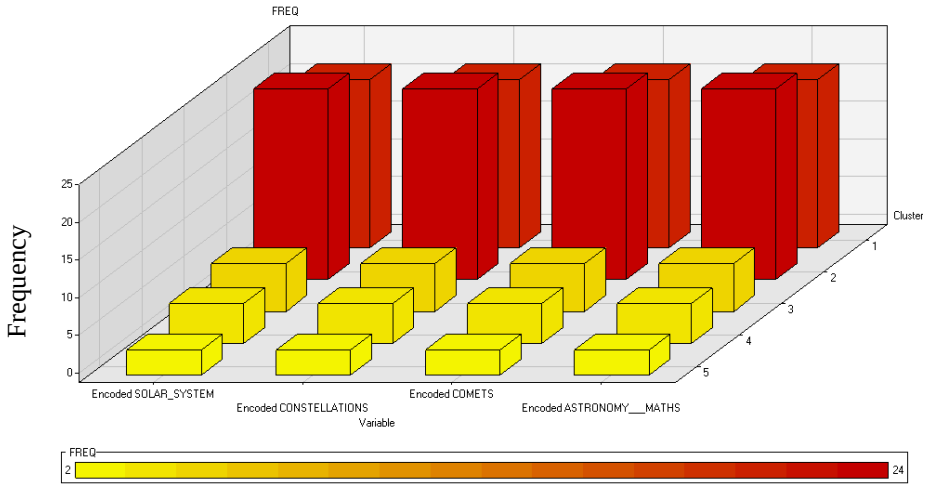


Figure 61 - Statistics Plot for Clusters

Table 14 - Average Score for Web Pages from Cluster 2

Web Page	Average Score
Solar System	4.33
Constellations	4.63
Comets	4.29
Astronomy & Math	4.50

4.13 Preferred Mathematical Data Properties

We use cluster 3 with maximum frequency of 24 for determining the user preferences for mathematical data properties. Figure 62 shows the dataflow diagram for preferred mathematical data properties analysis. Figure 63 shows the inappropriate values that have been filtered out from consideration for clustering. Figure 64 shows the pages that are being evaluated for mathematical properties. The status column in the table indicates the web pages on that are being used in this analysis. Figure 65 shows a 3-D pie chart indicating the 5 clusters, into which the observations are grouped. Cluster 3 has the maximum frequency.

Figure 66 compares values of cluster 3 with all the remaining clusters. Figure 67 shows the distance plot for the clusters and Figure 68 gives the statistics plot.

Table 15 shows the average values for cluster 3. Astronomy & Math and Constellations had the top two scores with respect to mathematical properties. This showed that users preferred mathematical data in images which could be clicked and enlarged for better readability. Mathematical formulae written using equation editor were not very popular with the users.

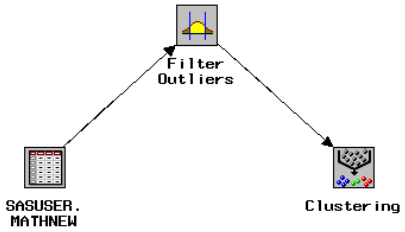


Figure 62 - Data Flow Diagram for Analysis of Mathematical Data Properties

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	
SOLAR_SYSTEM	Solar System	0	6
CONSTELLATIONS	Constellations	0	6
METEORS	Meteors	0	
COMETS	Comets	0	
ASTRONOMY___MATHS	Astronomy & Maths	0	6
ARTICLES	Articles	0	
UPCOMING_EVENTS	Upcoming Events	0	
ASTRONOMY_WEBSITES	Astronomy Websites	0	

Figure 63 - Filtered Values for Clustering

Name	Status	Model Role	Measurement	Type	Format	Label
HOME_PAGE	don't use	input	ordinal	num	BEST12.	Home Page
SOLAR_SYSTEM	use	input	ordinal	num	BEST12.	Solar System
CONSTELLATIONS	use	input	ordinal	num	BEST12.	Constellations
METEORS	don't use	input	ordinal	num	BEST12.	Meteors
COMETS	don't use	input	ordinal	num	BEST12.	Comets
ASTRONOMY___MATHS	use	input	ordinal	num	BEST12.	Astronomy & Maths
ARTICLES	don't use	input	ordinal	num	BEST12.	Articles
UPCOMING_EVENTS	don't use	input	ordinal	num	BEST12.	Upcoming Events
ASTRONOMY_WEBSITES	don't use	input	ordinal	num	BEST12.	Astronomy Websites

Figure 64 - Web Pages being Evaluated for Analysis of Mathematical Data Properties

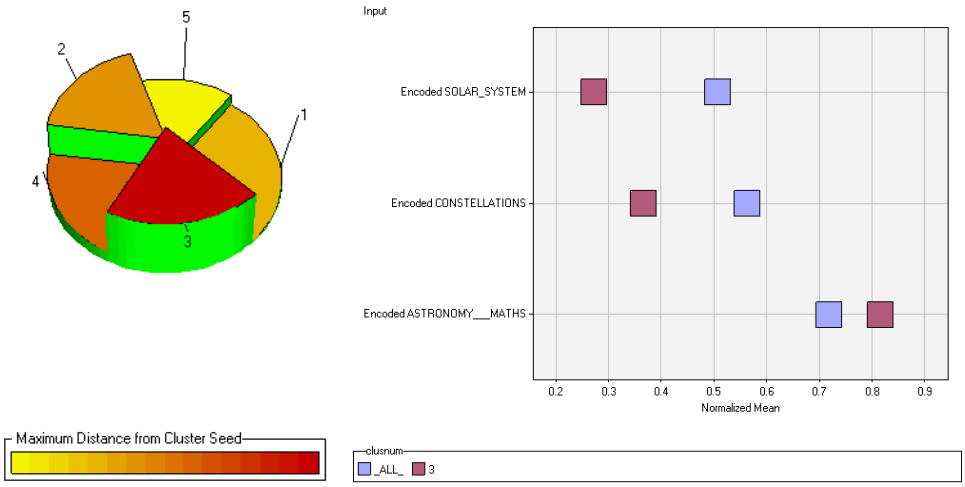


Figure 65 - Cluster Pie Chart (L)

Figure 66 - Input Means Plot (R)

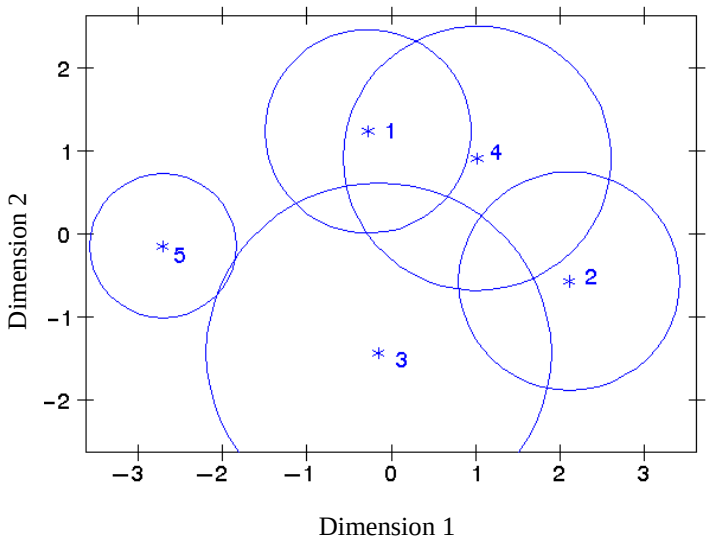


Figure 67 - Distance Plot for Clusters

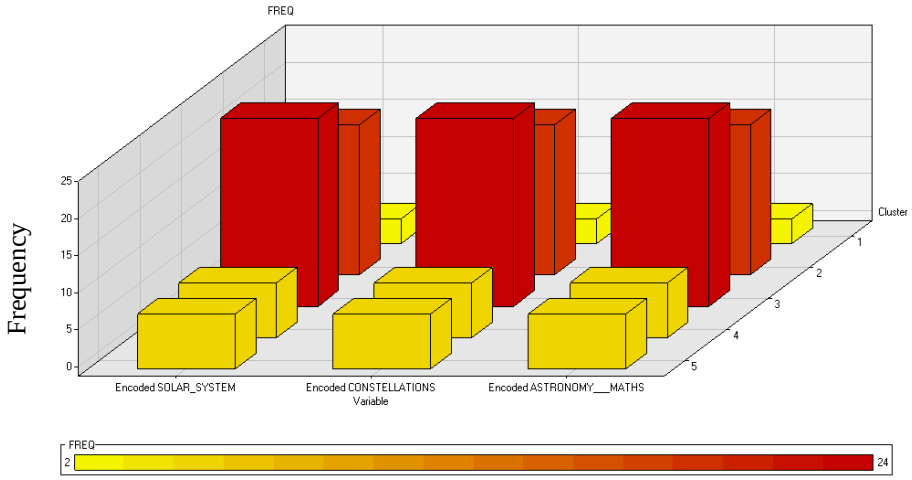


Figure 68 - Statistics Plot for Clusters

Table 15 - Average score for Web Pages from Cluster 3 for Mathematical Data Properties

Web Page	Average Score
Solar System	2.29
Constellations	2.58
Astronomy & Math	4.13

4.14 Preferred Article Formats

We consider the average values from cluster 3 with maximum frequency of 21 for determining the user preferences for article formats. Figure 69 shows the dataflow diagram for preferred article formats analysis. Figure 70 shows the inappropriate values that have been filtered out from consideration for clustering. Figure 71 shows the pages that are being evaluated for article formats. The status column in the table indicates the web pages on that are being used in this analysis. Figure 72 shows a 3-D pie chart indicating the 5 clusters, into which the observations are grouped. Cluster 3 has the maximum frequency. Figure 73 compares values of cluster 3 with all the remaining clusters. Figure 74 shows the distance plot for the clusters and Figure 75 shows the statistics plot.

Table 16 shows the average values from cluster 3. Astronomy Websites had the highest score with respect to article formats, indicating that users preferred web documents formatted as HTML web pages, than PDF or PostScript documents.

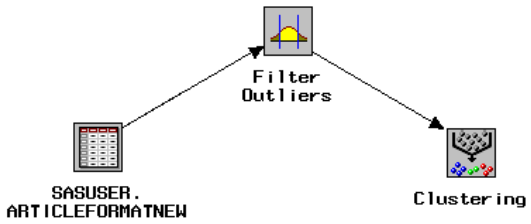


Figure 69 - Data Flow Diagram for Analysis of Web Document Formats

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	
SOLAR_SYSTEM	Solar System	0	
CONSTELLATIONS	Constellations	0	
METEORS	Meteors	0	
COMETS	Comets	0	
ASTRONOMY___MATHS	Astronomy & Maths	0	
ARTICLES	Articles	0	6
UPCOMING_EVENTS	Upcoming Events	0	
ASTRONOMY_WEBSITES	Astronomy Websites	0	6

Figure 70 - Filtered Values for Clustering

Name	Status	Model Role	Measurement	Type	Format	Label
HOME_PAGE	don't use	input	ordinal	num	BEST12.	Home Page
SOLAR_SYSTEM	don't use	input	ordinal	num	BEST12.	Solar System
CONSTELLATIONS	don't use	input	ordinal	num	BEST12.	Constellations
METEORS	don't use	input	ordinal	num	BEST12.	Meteors
COMETS	don't use	input	ordinal	num	BEST12.	Comets
ASTRONOMY___MATHS	don't use	input	ordinal	num	BEST12.	Astronomy & Maths
ARTICLES	use	input	ordinal	num	BEST12.	Articles
UPCOMING_EVENTS	don't use	input	ordinal	num	BEST12.	Upcoming Events
ASTRONOMY_WEBSITES	use	input	ordinal	num	BEST12.	Astronomy Websites

Figure 71 - Web Pages being Evaluated for Analysis of Web Document Formats

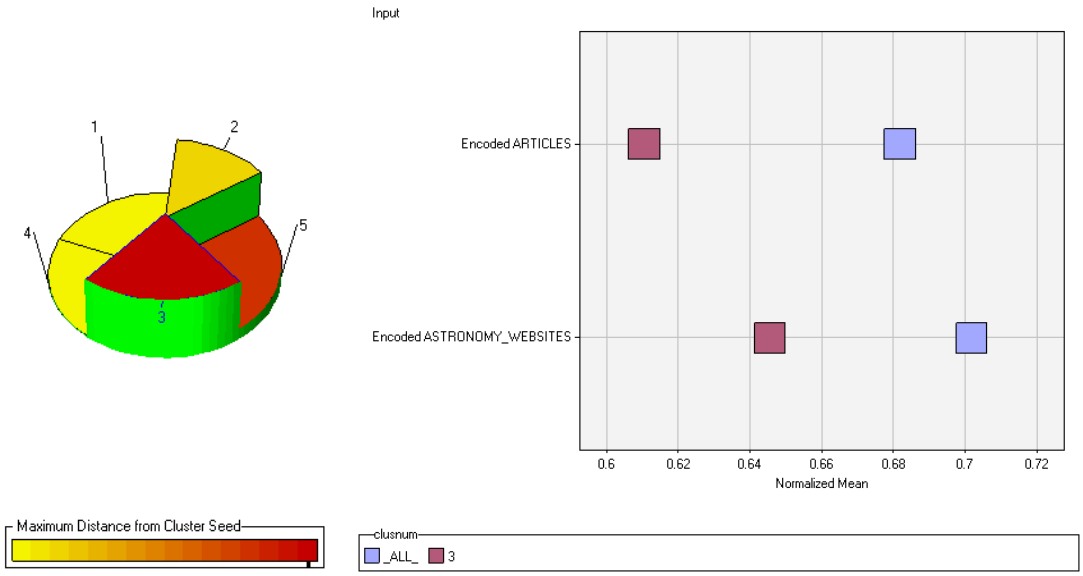


Figure 72 - Cluster Pie Chart (L)

Figure 73 - Input Means Plot (R)

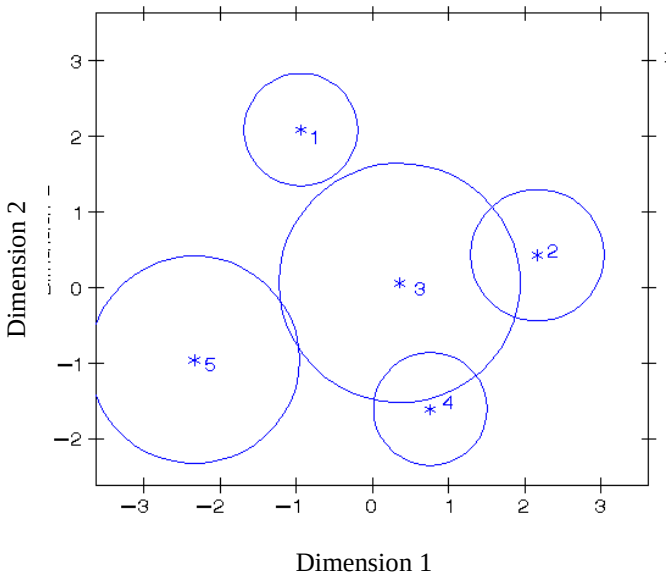


Figure 74 - Distance Plot for Clusters

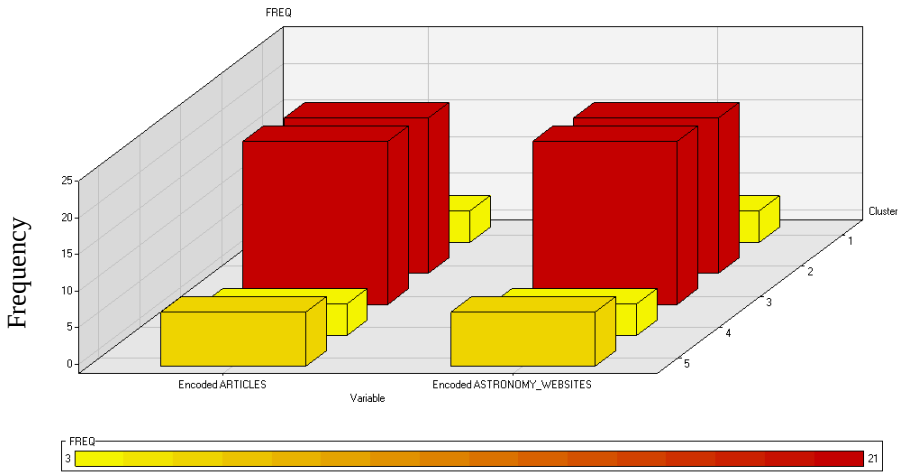


Figure 75 - Statistics Plot for Clusters

Table 16 - Average Score for Web Pages from Cluster 3

Web Page	Average Score
Articles	3.48
Astronomy Websites	3.57

4.15 Preferred Content Presentation

Cluster 2 with the maximum frequency of 22 is used to determine the most preferred format for content presentation. Figure 76 shows the dataflow diagram for preferred content presentation analysis. Figure 77 shows the inappropriate values that have been filtered out from consideration for clustering. Figure 78 shows a 3-D pie chart indicating the 5 clusters, into which the observations are grouped. Clusters 1 and 2 have the maximum frequency. We consider cluster 2 here. Figure 79 compares values of cluster 2 with all the remaining clusters. Figure 80 shows the distance plot for the clusters and Figure 81 shows the statistics plot.

Table 17 shows the average values from cluster 2, which are used to determine the user preferences for the entire content presentation. Articles, Home Page, Meteors, Astronomy & Math and Astronomy Websites had the top scores with respect to presentation of content. Users preferred simple pages, with basic arrangement of information in plain text, tables or lists. Simple bar graphs that presented information in an organized manner also helped in improving the readability. Plain, light or contrast backgrounds with different font sizes for headings, sub-headings and content contributed towards better visual presentation of the pages. Also, more readable pages did not require scrolling to view the complete content. In all, most preferred web pages contained a good combination of the most preferred factors evaluated in the earlier sections.

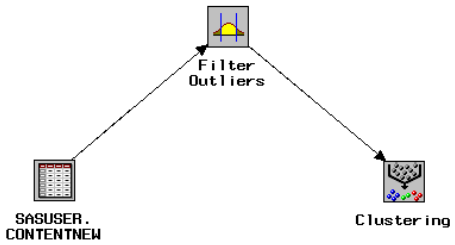


Figure 76 - Data Flow Diagram for Content Presentation Analysis

Name	Label	Min Freq	Values to exclude
HOME_PAGE	Home Page	0	6
SOLAR_SYSTEM	Solar System	0	6
CONSTELLATIONS	Constellations	0	6
METEORS	Meteors	0	6
COMETS	Comets	0	6
ASTRONOMY___MATHS	Astronomy & Maths	0	6
ARTICLES	Articles	0	6
UPCOMING_EVENTS	Upcoming Events	0	6
ASTRONOMY_WEBSITES	Astronomy Websites	0	6

Figure 77 - Filtered Values for Clustering

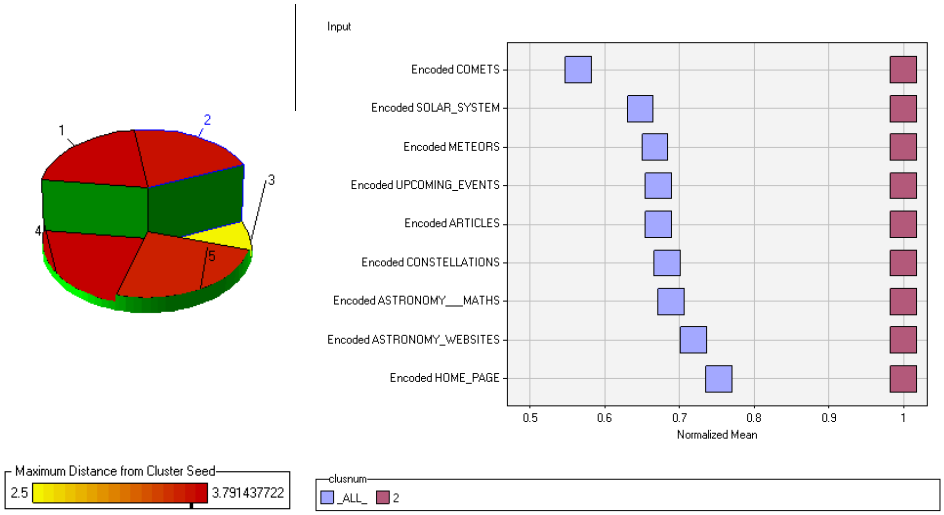


Figure 78 - Cluster Pie Chart (L)

Figure 79 - Input Means Plot (R)

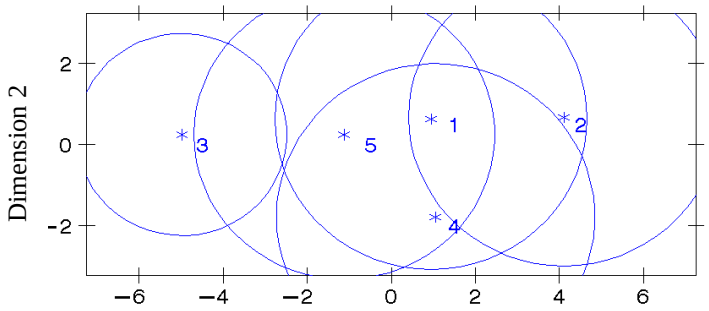


Figure 80 - Distance Plot for Clusters

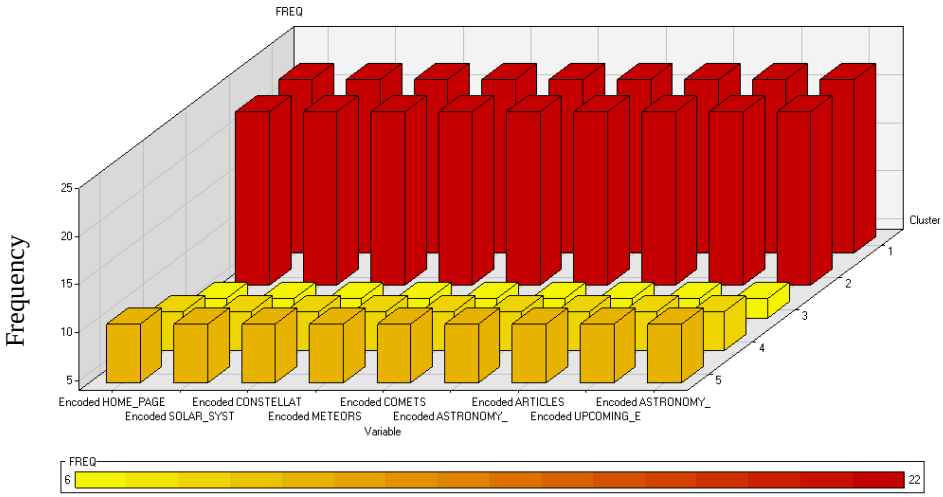


Figure 81 - Statistics Plot for Clusters

Table 17 - Average Score for Web Pages from Cluster 2

Web Page	**Average Score**
Home Page	4.32
Solar System	3.86
Constellations	4.05
Meteors	4.23
Comets	3.95
Astronomy & Math	4.23
Articles	4.36
Upcoming Events	4.14
Astronomy Websites	4.23

4.16 Association between Readability Factors

Cluster analyses can be performed via plots of the clustering history referred to as tree diagrams or dendrograms. Dendrograms graphically present the information concerning which observations are grouped together at various levels of (dis)similarity. At the bottom of the dendrogram, each observation is considered its own cluster. Vertical lines extend up for each observation and at various (dis)similarity values, these lines are connected to the lines

from other observations with a horizontal line. The observations continue to combine until, at the top of the dendrogram, all observations are grouped together. The height of the vertical lines and the range of the (dis)similarity axis give visual clues about the strength of the clustering. Long vertical lines indicate more distinct separation between the groups. Long vertical lines at the top of the dendrogram indicate that the groups represented by those lines are well separated from one another. Shorter lines indicate groups that are not as distinct.

SAS provides a procedure to create such plots called PROC TREE. This procedure uses the output dataset from PROC CLUSTER. PROC TREE has options to enhance the plot by altering its shape and labeling. The association between all the readability factors can be analyzed using the cluster analysis on each web page. Clustering is performed with all readability factors and dendrograms are extracted from the cluster output. Analysis of the dendrograms reveals the association between readability factors. For factors clustered together, a change in one factor impacts the user rating for that factor and the other factors in the cluster. As the factors get more separated in the dendrogram, their impact on each other reduces. Dendrograms can thus be used to understand the association between readability factors.

The simple and default code for the PROC TREE procedure is:

PROC TREE data=<cluster output dataset>;

We use the PROC VARCLUS to create clusters of the readability factors affecting each web page and then use the tree procedure to get the corresponding dendrograms.

For example, the code for creating the dendrogram for home page is written as:

PROC VARCLUS data=SASUSER.HOMEDENDROGRAM

outtree=SASUSER.HOMETREE **centroid maxclusters**=5 **noprint**;

run;

PROC TREE data=SASUSER.HOMETREE;

PROC VARCLUS above creates clusters for the home page on the website with all the readability factors as variables. The centroid clustering algorithm is used and maximum number of clusters is set to 5. The output dataset from the VARCLUS procedure is passed as an input dataset to the TREE procedure, which represents the clusters in the form of a hierarchical tree.

The dendrogram from the above TREE procedure is shown in Figure 82.

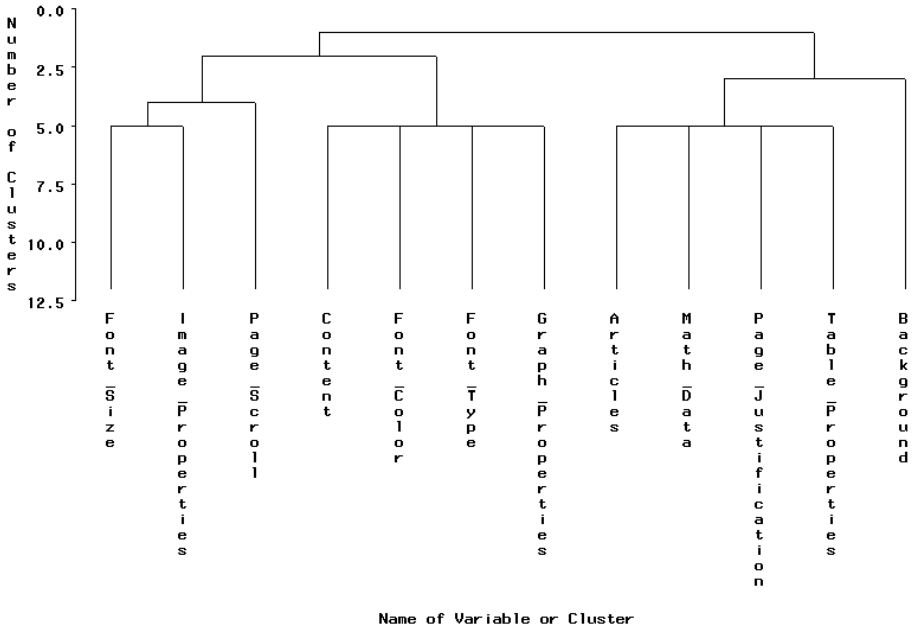


Figure 82 - Dendrogram for the Home Page

Figure 82 indicates two main clusters for variables (i.e. readability factors) in the home page. Table 18 indicates the factors that belong to these two main clusters for the home page. Table 18 indicates that the factors Font Color, Font Size, Font Type, Content Presentation, Image Properties, Graph Properties and Page Scroll are more associated with each other while factors Article Formats, Math Data Properties, Page Justification, Table Properties and Background are more associated with one another. For example, if Font Size is changed such that it gets higher user ratings, there is more possibility that Font Type, which belongs in the same cluster, will also get a higher user rating. However, change in the user rating for Background in this case would be comparatively less. So, the impact of Font Size is more on factors that belong to the same cluster than another cluster.

We analyze the impact of readability factors on each of the web pages in the website and the association between them using the method described above for the Home Page.

Table 18 - Main Clusters for the Home Page

Cluster 1	**Cluster 2**
Font Color	Article Formats
Font Size	Math Data Properties
Font Type	Page Justification
Content Presentation	Table Properties
Image Properties	Background
Graph Properties	
Page Scroll	

Solar System Page

Figure 83 shows the dendrogram created by the TREE procedure for Solar System web page.

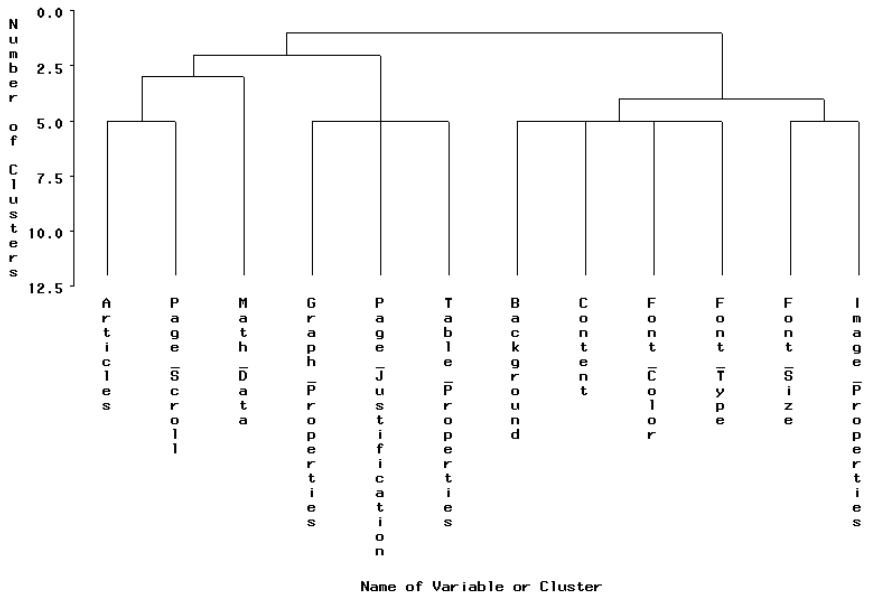


Figure 83 - Dendrogram for Solar System Page

Figure 83 indicates two main clusters for variables (i.e. readability factors) in the Solar System page at the top level. Table 19 indicates the factors that belong to these two main clusters for the Solar System page.

Table 19 - Main Clusters for the Solar System Page

Cluster 1	**Cluster 2**
Font Color	Article Formats
Font Size	Math Data Properties
Font Type	Page Justification
Background	Table Properties
Image Properties	Page Scroll
Content	Graph Properties

Constellations Page

Figure 84 shows the dendrogram created by the TREE procedure for Constellations web page.

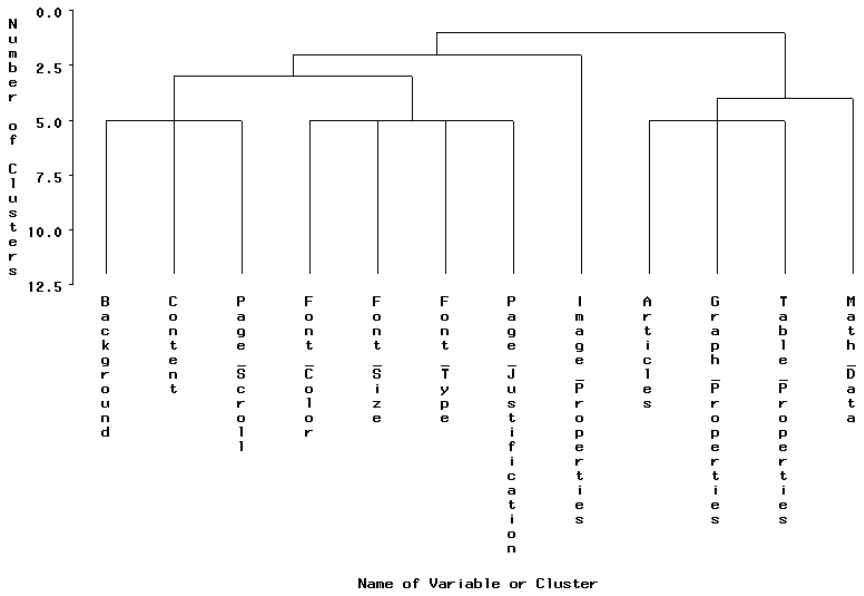


Figure 84 - Dendrogram for the Constellations Page

Figure 84 indicates two main clusters for variables (i.e., readability factors) in the Constellations page at the top level. Table 20 indicates the factors that belong to these two main clusters for the Constellations page.

Table 20 - Main Clusters for the Constellations Page

Cluster 1	Cluster 2
Font Color	Article Formats
Font Size	Graph Properties
Font Type	Table Properties
Background	Math Data Properties
Image Properties	
Content	
Page Scroll	
Page Justification	

Meteors Page

Figure 85 shows the dendrogram created by the TREE procedure for Meteors web page.

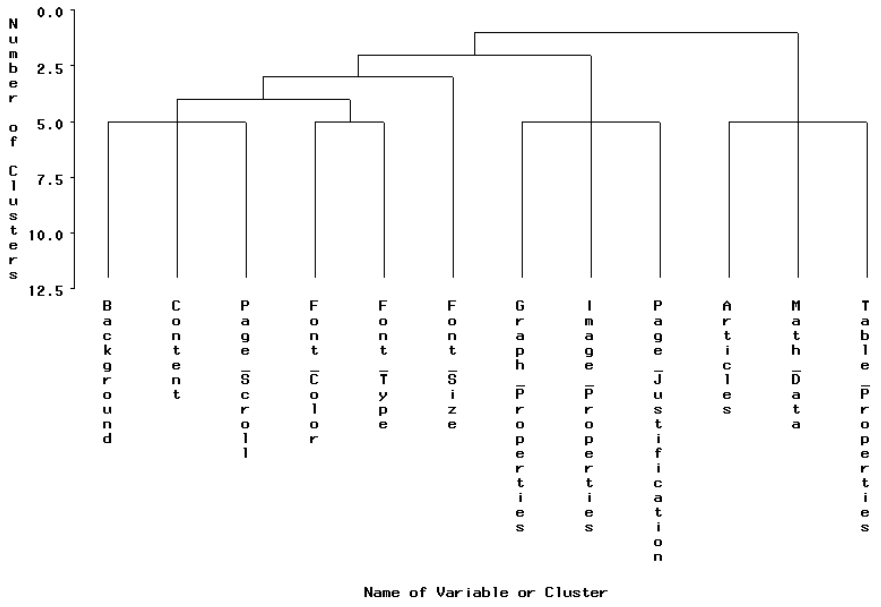


Figure 85 - Dendrogram for the Meteors Page

Figure 85 indicates two main clusters for variables (i.e. readability factors) in the Meteors page at the top level. Table 21 indicates the factors that belong to these two main clusters for the Meteors page.

Table 21 - Main Clusters for the Meteors Page

Cluster 1	Cluster 2
Font Color	Article Formats
Font Size	Table Properties
Font Type	Math Data Properties
Background	
Image Properties	
Content	
Page Scroll	
Page Justification	
Graph Properties	

Comets Page

Figure 86 shows the dendrogram created by the TREE procedure for the Comets web page.

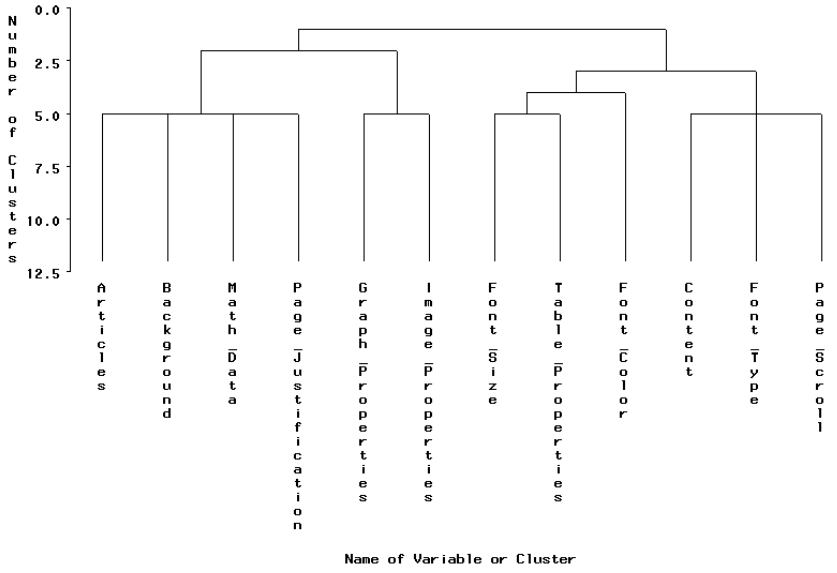


Figure 86 - Dendrogram for the Comets Page

Figure 86 indicates two main clusters for variables (i.e., readability factors) in the Comets page at the top level. Table 22 indicates the factors that belong to these two main clusters for the Comets page.

Table 22 - Main Clusters for the Comets Page

Cluster 1	Cluster 2
Font Color	Article Formats
Font Size	Background
Font Type	Math Data Properties
Content	Page Justification
Table Properties	Graph Properties
Page Scroll	Image Properties

Astronomy & Math Page

Figure 87 shows the dendrogram created by the TREE procedure for the Astronomy & Math web page.

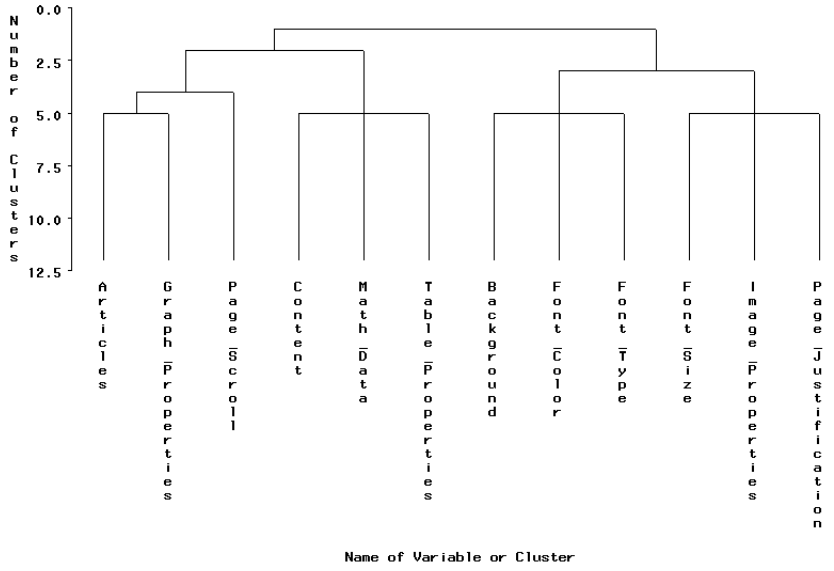


Figure 87 - Dendrogram for the Astronomy & Math Page

Figure 87 indicates two main clusters for variables (i.e., readability factors) in the Astronomy & Math page at the top level. Table 23 indicates the factors that belong to these two main clusters for the Astronomy & Math page.

Table 23 - Main Clusters for the Astronomy & Math Page

Cluster 1	Cluster 2
Font Color	Article Formats
Font Size	Graph Properties
Font Type	Page Scroll
Background	Content
Image Properties	Math Data Properties
Page Justification	Table Properties

Articles Page

Figure 88 shows the dendrogram created by the TREE procedure for the Articles web page.

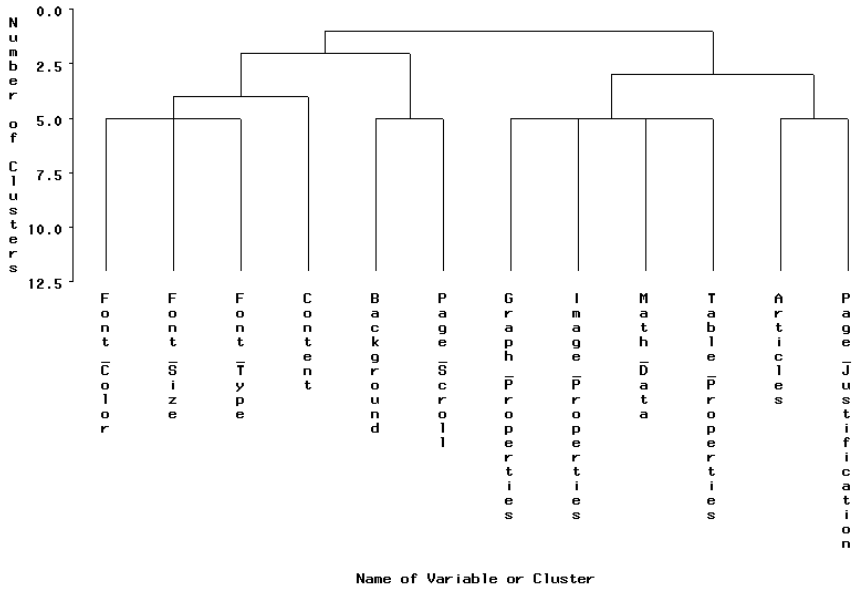


Figure 88 - Dendrogram for the Articles Page

Figure 88 indicates two main clusters for variables (i.e., readability factors) in the Articles page at the top level. Table 24 indicates the factors that belong to these two main clusters for the Articles page.

Table 24 - Main Clusters for the Articles Page

Cluster 1	Cluster 2
Font Color	Article Formats
Font Size	Graph Properties
Font Type	Page Justification
Background	Image Properties
Content	Math Data Properties
Page Scroll	Table Properties

Astronomy Websites Page

Figure 89 shows the dendrogram created by the TREE procedure for the Astronomy Websites page.

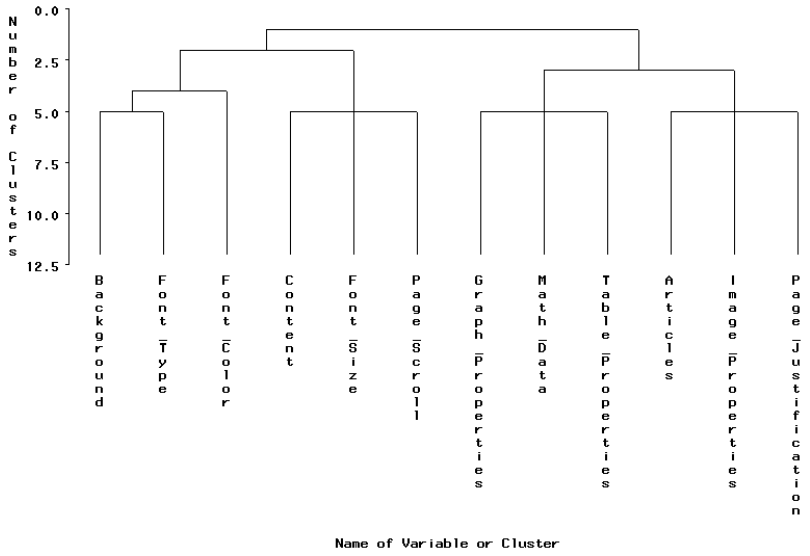


Figure 89 - Dendrogram for the Astronomy Websites Page

Figure 89 indicates two main clusters for variables (i.e., readability factors) in the Astronomy Websites page at the top level. Table 25 indicates the factors that belong to these two main clusters for the Astronomy Websites page.

Table 25 - Main Clusters for the Astronomy Websites Page

Cluster 1	Cluster 2
Font Color	Article Formats
Font Size	Graph Properties
Font Type	Page Justification
Background	Image Properties
Content	Math Data Properties
Page Scroll	Table Properties

Upcoming Events Page

Figure 90 shows the dendrogram created by the TREE procedure for the Upcoming Events web page.

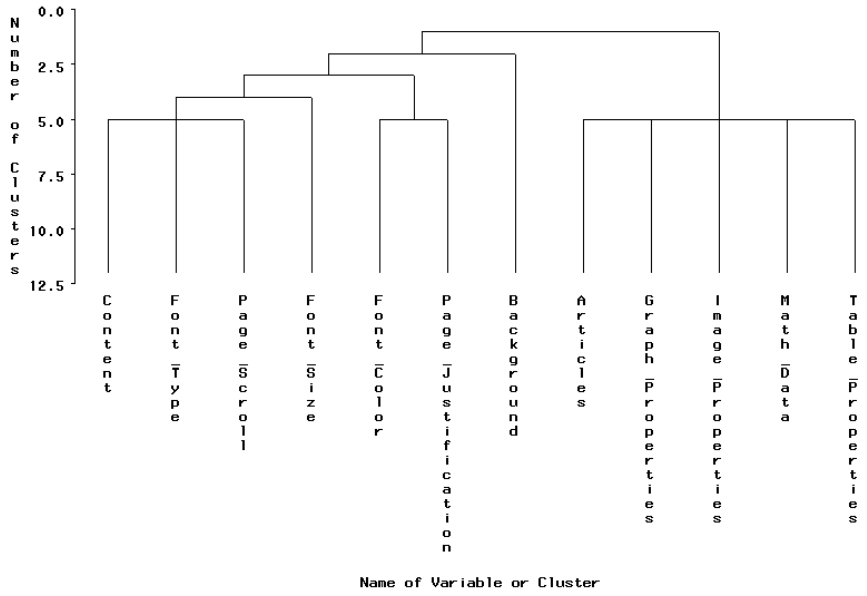


Figure 90 - Dendrogram for the Upcoming Events Page

Figure 90 indicates two main clusters for variables (i.e., readability factors) in the Upcoming Events page at the top level. Table 26 indicates the factors that belong to these two main clusters for the Upcoming Events page.

Table 26 - Main Clusters for the Upcoming Events Page

Cluster 1	**Cluster 2**
Font Color	Article Formats
Font Size	Graph Properties
Font Type	Image Properties
Background	Math Data Properties
Content	Table Properties
Page Scroll	
Page Justification	

CHAPTER 5: SUMMARY, CONCLUSION AND FUTURE WORK

5.1 Introduction

This study was designed to investigate the factors affecting readability of scientific web pages by collecting the preferences and opinions from the users of these web pages. Users responded to survey questions about a subset of factors that could impact readability of scientific websites. The analysis and data extraction of these survey responses revealed how the selected factors affected readability and also provided quantitative measures for those factors. In the future, this analysis may help to develop an algorithm to redesign web pages to improve their readability and usability.

5.2 Summary of Results

5.2.1 Page Preference Analysis

From the cluster analysis of the readability factors for each web page on the website, we found the web pages that were given the highest scores by users with respect to each readability factor. Table 27 summarizes these preference results.

Table 27 - Preferred Web Pages

Readability Factor	Preferred Web Pages
Font types	Astronomy & Math Home Page Articles
Font sizes	Home Page Articles Comets
Font colors	Articles Home Page Comets
Page Scroll	Articles Home Page Astronomy Websites
Page Justification	Articles Astronomy Websites Home Page
Image Properties	Astronomy & Math Home Page Solar System
Background	Comets Upcoming Events Home Page
Graph Properties	Meteors Comets
Table Properties	Constellations Astronomy & Math
Mathematical Data Properties	Astronomy & Math Constellations
Article Formats	Astronomy Websites
Content Presentation	Articles Home Page Meteors Astronomy & Math Astronomy Websites

5.2.2 Associated Readability Factors

Each web page on the website was analyzed using cluster analysis and dendrograms or tree diagrams were extracted from the clusters to evaluate the association between the readability factors. For most of the web pages, font color, font size, font type, content

presentation and background clustered together on one top level branch of the dendrogram. Article formats, graph properties, table properties and mathematical data properties usually grouped together in one cluster. Page scroll, page justification and image properties clustered with either of the two groups of readability factors.

5.3 Conclusion

From the summary of results section above, we determined the preferred web pages with respect to each readability factor. Considering the values for the readability factors for these preferred web pages, we concluded that the survey presented the most preferred font types, font sizes, font colors, page scroll, page justification, image properties, background, graph properties, table properties, mathematical data properties, article formats and content presentation. Table 28 specifies the same.

Table 28 - Preferred Values for Readability Factors

Readability Factor	Preferred Web Pages	Preferred Values for Readability Factors
Font types	Astronomy & Math Home Page Articles	Helvetica – headings/ sub-headings Georgia, Arial, Verdana – content
Font sizes	Home Page Articles Comets	13pt – headings 12-11pt – sub-headings 11-10pt – content
Font colors	Articles Home Page Comets	Blue-Gray Red-Blue-Purple Blue-Black
Page Scroll	Articles Home Page Astronomy Websites	No page scroll most preferred Vertical scroll over horizontal scroll
Page Justification	Articles Astronomy Websites Home Page	Left-justified Center and justified preferred over right-justified
Image Properties	Astronomy & Math Home Page Solar System	Gray-scale images over colored images Titles and captions to describe images
Background	Comets Upcoming Events Home Page	Plain white Light colored and background with simple images over dark backgrounds
Graph Properties	Meteors Comets	Simple bar or line graph Colors to enhance different values
Table Properties	Constellations Astronomy & Math	Completely bordered tables with separate cells for each value
Mathematical Data Properties	Astronomy & Math Constellations	Mathematical data as clickable images
Article Formats	Astronomy Websites	Web documents in the form of HTML web pages over PDF or PostScript
Content Presentation	Articles Home Page Meteors Astronomy & Math Astronomy Websites	Simple pages with basic arrangement of information in plain text, tables or lists Simple bar or line graphs Plain and light backgrounds Headings and content distinguished by font properties Pages with no scrolling Good combination of the above preferred values for all the readability factors

From the dendrograms extracted from clustering the readability factors for each web page on the website, we understood the association between these factors. Since font color, font size, font type, content and background mostly clustered together, they had a higher degree of association between them. Similarly, article formats, graph properties, table properties and mathematical data properties were more associated with each other. The degree of association between these two groups of factors could be minimal. This pattern may be because the factors belonging to one cluster received similar ratings from the users. A change in the value of one factor, could lead to change in the user rating for that factor and the factors that belonged to the same cluster. Factors that belonged to separate clusters did not impact each other as much. Improving a subset of factors in a cluster could improve the overall perceived readability of that cluster of factors.

5.4 Current Limitations and Future Work

Our study focuses on scientific websites. This could limit the generalizability of findings to other websites. The readability factors selected for the evaluation study is a subset of many such factors that could impact the readability of a scientific website. The values chosen for readability factors for the sample scientific website are restricted by literature review and standards provided by technical organizations. This could limit the reliability of the analysis results. The sample size used for the study is ninety. The findings would be more reliable given a large sample size. The number of questions included in the survey is limited so that it does not consume too much time or effort. This restricts the amount of preference information obtained for each readability factor. Since the study involves browsing a website, changes in the monitor dimensions, configuration and screen resolution could impact the appearance and aesthetic quality of the website. Consequently, the user responses may be

skewed. In the future, this may be avoided by requiring the participants to take the surveys at designated laboratories with consistent system settings. The study is limited to static web pages and does not evaluate readability of interactive or dynamic pages.

BIBLIOGRAPHY

[ACM, 2009] Association for Computing Machinery (2009). http://www.acm.org/sigs/publications/proceedings-templates

[Angeli, 2006] Angeli, A. D., Sutcliffe, A., & Hartmann, J. (2006, June). *Interaction, usability and aesthetics: What influences users' preferences?* Paper presented at the sixth conference on designing interactive systems, University Park, PA. doi:10.1145/1142405.1142446

[Brinck, 2003] Brinck, T., Ha, S., S., Pritula, N., Lock, K., Speredelozzi, A., & Monan, M. (2003, June). *Making an impact: Redesigning a business school web site around performance metrics.* Paper presented at the 2003 conference on designing for user experiences, San Francisco, CA. doi:10.1145/997078.997084

[Erinaki, 2003] Erinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT), 3,* 1-27. doi:10.1145/643477.643478

[Hall, 2004] Hall, R., H., & Hanna, P. (2004). The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioral intention. *Behavior & Information Technology, 23,* 183-195. doi:10.1080/01449290410001669932

[IEEE, 2009] Institute of Electrical and Electronics Engineers (2009). http://www.sec09.com/content/Paper_Submission

[Ivory, 2001a] Ivory, M., Y., & Hearst, M., A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR), 33,* 470-516. doi:10.1145/503112.503114

[Ivory, 2001b] Ivory, M., Y., Sinha, R., R., & Hearst, M., A. (2001, March). *Empirically validated web page design metrics.* Paper presented at the SIGCHI conference on human factors in computing systems, Seattle, WA. doi:10.1145/365024.365035

[Ivory, 2002] Ivory, M., Y., & Hearst, M., A. (2002, April). *Statistical profiles of highly-rated web sites.* Paper presented at the SIGCHI conference on human factors in computing systems: Changing our world, changing ourselves, Minneapolis, MN. doi:10.1145/503376.503442

[Ivory, 2005] Ivory, M., Y., & Megraw, R. (2005). Evolution of web site design patterns. *ACM Transactions on Information Systems (TOIS), 23,* 463-497. doi:10.1145/1095872.1095876

[Joshi, 1999] Joshi, K. P., Joshi, A., Yesha, Y., & Krishnapuram, R. (1999, November). *Warehousing and mining web logs.* Paper presented at the second international workshop on web information and data management, Kansas City, MO. doi:10.1145/319759.319792

[Michailidou, 2008] Michailidou, E., Harper, S., Bechhofer, S. (2008, September). *Visual complexity and aesthetic perception of web pages.* Paper presented at the twenty-sixth annual ACM international conference on design of communication, Lisbon, Portugal. doi:10.1145/1456536.1456581

[Norman, 2004] Norman, D. A. (2004). *Emotional Design: Why We Love (or Hate) Everyday Things.* New York, NY: Basic Books.

[Obendorf, 2004] Obendorf, H., Weinreich, H., & Hass, T. (2004, April). *Automatic support for web user studies with SCONE and TEA.* Paper presented at the 2004 conference on human factors in computing systems, Vienna, Austria. doi:10.1145/985921.986007

[Rosenholtz, 2005] Rosenholtz, R., Li, Y., Mansfield, J., & Jin, Z. (2005, April). *Feature Congestion: A measure of display clutter.* Paper presented at the 2005 conference on human factors in computing systems, Portland, OR. doi:10.1145/1054972.1055078

[SAS, 2009] Statistical Analysis System (2009). http://support.sas.com/documentation/onlinedoc/miner/getstarted53.pdf

[Schaik, 2007] Schaik, P.V., & Ling, J. (2007). Design parameters of rating scales for web sites. *ACM Transactions on Computer-Human Interaction (TOCHI), 14.* doi:10.1145/1229855.1229859

[Survey Monkey, 2009] Survey Monkey (2009). http://www.surveymonkey.com/

[Swaak, 2009] Swaak, M., Jong, M, D., & Vries, P., D. (2009, July). *Effects of information usefulness, visual attractiveness and usability on web visitors' trust and behavioral intentions.* Paper presented at the 2009 IEEE international professional communication conference, Waikiki, HI. doi:10.1109/IPCC.2009.5208719

[Tan, 2006] Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Boston, MA: Pearson Education Inc.

[Webby, 2000] Webby Awards (2000). http://www.webbyawards.com/entries/criteria.php#websites

[Whitehead, 2006] Whitehead, C., C. (2006, March). *Evaluating web page and web site usability.* Paper presented at the forty-fourth annual southeast regional conference, Melbourne, FL. doi:10.1145/1185448.1185637

APPENDIX A: IRB APPROVAL

To: Seena Menon
Computer Science
CAMPUS MAIL

From: _____
Julie Taubman, Institutional Review Board

Date: 9/22/2009

RE: Notice of IRB Exemption

Study #: 10-0032
Study Title: Evaluating The Readability of Scientific Web Pages Using Intelligent Analysis Tools
Exemption Category: 4: Existing data/specimens, publicly available, unlinkable to individuals

This submission has been reviewed by the IRB Office and was determined to be exempt from further review according to the regulatory category cited above under 45 CFR 46.101(b). Should you change any aspect of the proposal, you must contact the IRB before implementing the changes to make sure the exempt status continues to apply. Otherwise, you do not need to request an annual renewal of IRB approval. Please notify the IRB Office when you have completed the study.

APPENDIX B: INFORMED CONSENT

APPALACHIAN STATE UNIVERSITY

Informed Consent for Participants in Research Projects Involving Human Subjects

Institutional Review Board
Study #: 10-0032

Title of Project: Evaluating the Readability of Scientific Web Pages Using Intelligent Analysis Tools

Investigator(s): Seena S. Menon

I. Purpose of this Research/Project
To collect the readability information of scientific web pages using intelligent analysis tools and use the collected data to reformat/ reconstruct the web pages for better readability.

II. Procedures
The subjects will have to navigate through some created web pages, and the intelligent analysis tools will be used to report the clicks, time between pages, readability of figures, formulae, etc. Once the analyzed information has been used to reconstruct the same web pages, subjects will have to once again navigate through the reconstructed pages. The statistics will be reported and then the values will be compared to conclude the study. The hand movements/ clicks and the user experience will be recorded for further analysis. The investigator will make sure that the video recording does not include the subject' faces, and the audio recording does not include any personal/ contact information.

III. Risks
There are no risks to the subjects in this research.

IV. Benefits
Technical papers have a pool of information for the purpose of research and further study. In general, scientific web pages can be difficult to read because of the figures and mathematical calculations. Through this study, we would be able to analyze the factors that affect the readability of such web pages, and reformat them such that their readability and usability can be improved.

V. Extent of Anonymity and Confidentiality
No personal or contact information is requested from the subjects, and therefore their anonymity is maintained.

VI. Compensation
There is no compensation for participation in this study.

VII. Freedom to Withdraw
Subjects are free to withdraw from participation at any point during the testing.

VIII. Approval of Research
This research project has been approved, as required, by the Institutional Review Board of Appalachian State University and _____ (if others, i.e., school or school system, hospital, daycare center, multi-institutional project etc.).

_09/22/2009_____ __Till completion of project_____
IRB Approval Date Approval Expiration Date

IX. Subject's Responsibilities
I voluntarily agree to participate in this study. I have the following responsibilities:

Visit and navigate through the web pages created by the investigator. The recording software will capture my hand movements and clicks. I have to narrate my experience with the web pages so that it can also be recorded for further analysis by the investigator. I am aware that the investigator will be using some web analytic tools to extract the readability information of the web pages as I use them.

X. Subject's Permission (May be modified in the case of minors or members of other vulnerable populations.)
I have read and understand the Informed Consent and conditions of this project. I have had all my questions answered. I hereby acknowledge the above and give my voluntary consent:

_____Date_____
Subject signature

_____ Date _____
Witness (Optional except for certain classes of subjects)

Should I have any questions about this research or its conduct, I may contact:

Seena S. Menon____ 517 755 7887 menonss@appstate.edu
Investigator(s) Telephone/e-mail

Dr. Rahman Tashakkori 828-262-7009 rt@cs.appstate.edu ____
Faculty Advisor (if applicable) Telephone/e-mail

Ms. Julie Taubman 828-262-7981 irb@appstate.edu
IRB Administrator Telephone/e-mail

Graduate School and Research and Sponsored Programs
Appalachian State University
Boone, NC 28608
irb@appstate.edu

Subjects must be given a complete copy (or duplicate original) of the signed Informed Consent.

EXPLANATION OF TERMS

I. Purpose of this Research/Project Subjects should be informed in clear, concise language about the nature of the study and the purpose for conducting the research. The total number of subjects involved and a brief description of the subject pool (age range, health status, etc...) should be given.

II. Procedures The research procedures that involve human subjects should be explained in sufficient detail so that the subjects will be fully informed about their role, what activities or functions they will be expected to perform, for how long, the number of times they are expected to appear and over what period of time. They must be told where the research will take place, what instrumentation is to be used, if any, and conditions involved. At the end of this section, the subjects must have a clear understanding of what will be expected of them.

III. Risks Any risks or discomforts to the research subject must be fully disclosed. Risks may range from physical danger such as muscle injury from strenuous exercise to emotional distress caused by remembering unpleasant experiences. Safeguards that are to be employed to reduce or minimize the risks must be described.

IV. Benefits The tangible or intangible benefits, if any, to the subjects who participate must be described. If no benefits accrue to the subjects, what are the larger societal benefits for conducting the research? An analysis of the risks to benefits must clearly be on the benefits side. A statement must be included to the effect that -- no promise or guarantee of benefits have been made to encourage you to participate. At the option of the investigator, subjects may be informed that they may contact the researcher at a later time for a summary of the research results. If subjects are children, the parent/guardian must make the request.

V. Extent of Anonymity and Confidentiality
The extent to which subjects will be identifiable must be explained. If anonymity is promised (individuals cannot be identified), you need to explain how that will be accomplished. If confidentiality is promised (individuals can be identified, but the researchers promise not to divulge that information), you must explain how that will be accomplished. Social security numbers should not be used as identifiers in lieu of names. You may also say, "At no time will the researchers release the results of the study to anyone other than individuals working on the project without your written consent". If taping (video or audio) is to occur, the subjects must be informed. You must state how the tapes will be secured and stored, under whose supervision, who will score or transcribe, who will have access and when they will be destroyed. In some situations, it may be necessary for an investigator to break confidentiality. If child abuse is known or strongly suspected, investigators are required to notify the appropriate authorities. If a subject is believed to be a threat to herself/himself or others, the investigator should notify the appropriate authorities. The conditions under which the investigator may break confidentiality must be described in the Informed Consent.

VI. Compensation
There is no requirement that subjects are compensated, but if they are, they must be fully informed. If no compensation is to be earned, subjects must be so informed. Money or redeemable coupons or other currency may be given. Subjects must be informed about how much, when it will be paid, any bonuses for completing all the tasks, etc. If extra credit in a course is the compensation, the subject must be informed as to how much credit is to be earned and the impact of that extra credit on their course grade. If extra credit is a form of compensation for participation in research involving human subjects, there must be alternate and equitable ways to earn the equivalent credit in the same course without participating as a subject in research. The subjects must be so informed. The course syllabus must describe the alternate ways to earn extra credit. If as a result of a research project, the investigator determines that the subject should seek counseling or medical treatment, a list of local services should be provided. Also, a statement should be included indicting that no funds have been set aside for any injury or illness resulting from this project.

VII. Freedom to Withdraw
Subjects are free to withdraw from a study at any time without penalty. If they choose to withdraw, they will be compensated for the portion of the time of the study (if financial compensation is involved). If they choose to withdraw, they will not be penalized by reduction in points or grade in a course (if course credit is involved). Subjects are free not to answer any questions or respond to experimental situations that they choose without penalty. There may be circumstances under which the investigator may determine that a subject should not continue as a subject. The subject must be compensated for the portion of the project completed.

VIII. Approval of Research

This research project has been approved, as required, by the Institutional Review Board of Appalachian State University and _____ (if others, i.e., school or school system, hospital, daycare center, multi-institutional project etc.).

VITA

Seena Sukumaran Menon was born in Mumbai, India in 1981. She completed her Bachelor of Engineering in Computer Engineering from University of Mumbai in December 2003. She worked as a Senior Software Engineer with an IT company in Mumbai for almost four years before moving to USA in 2007. In August of 2008, she began her graduate studies in the Department of Computer Science at Appalachian State University. Ms. Menon has maintained a GPA of 4.0. She was offered a job opportunity with a leading retail company at their corporate headquarters in North Carolina in April of 2010 as a Programmer Analyst. Ms. Menon received the Master of Science in Computer Science degree in December 2010 and continues to work for her current employer.