Graphical Interface for Visual Exploration of Online Discussion Forums

Beomjin KIM
Department of Computer Science
Indiana University-Purdue University
Fort Wayne, IN 46805, U.S.A.

and

Philip JOHNSON
Department of Computer Science
Indiana University-Purdue University
Fort Wayne, IN 46805, U.S.A.

ABSTRACT

Studies have shown that visualization can be an effective methodology in analyzing a large amount of data rapidly by exploiting the user's perceptual cognition. Especially, where the information space is filled by considerable amount of noise. This paper introduces a visualization method that presents online message boards through an intuitive graphical illustration. The improved system achieves higher visual abstraction by applying color coding to symbolize multiple attributes – activities, time-related factors, and turbulence – of threads together. The magnitude of each thread is projected as the dimension of a graphical shape. In addition, the system provides the significance of threads with the positional attribute on the viewing window. Preliminary analysis reiterates the efficiency of the visual interface for search activities and also suggests further agendas for future study.

Keywords: Visualization, Information Retrieval, Information Filtering, Online Discussion Forums.

1. INTRODUCTION

The rapidly increasing amount of information in online spaces has brought a considerable interest in investigating methods to assist user's search activities. It is a time-consuming and frustrating task to explore huge information spaces without the assistance of any navigation tools. Visualization is a promising approach in analyzing a compilation of data through the user's perceptual cognition. Studies have shown that proper use of information visualization will improve the users' reviewing speeds and also increase their underastanding of the data [1, 2, 3, 4].

Various forms of online spaces such as newsgroups, bulletin boards, and chat rooms have been used for social discourse and discussion on a range of subjects. Text-based information is the most common medium for communication among users in these online spaces. Many applications employing visualization have been developed to provide the user richer and more effective communication environments. Chat Circles is a graphical interface for synchronous online conversation [5]. This system uses abstract shapes to convey identity and activity of the participants in online chatting. Meanwhile, the Conversation Map allows the user to access messages of an archive, and acquires social and semantic structures of the newsgroup through a graphical interface [6].

The online message board is an asynchronous social environment, which has been a valuable channel in sharing information and discussing many issues. Bulletin boards include huge numbers of threads composed of a group of text-based messages. The current presentation mechanism of message boards is ineffective in conveying the quality and correlations of threads to the users. It lists the threads in order of last posting date while just providing basic information such as subject, starter of the thread, and number of messages and references. Further, significant portions of the messages in the forum are either non-related to the user's interests or the contents are of poor quality. Information filtering and visualization have been explored to support the user's search in such noisy spaces. The GroupLens employs collaborative filtering in predicting the value of articles [7]. The accuracy is highly related with the number of evaluators, but the online forum generally has a much larger number of silent readers than active participants [8].

Xiong and Donath (1999) used a garden metaphor to visualize the history of users' interactions on a bulletin board. In the PeopleGarden, flowers represent authors and the collection of flowers shows the environment of a bulletin board. Other attributes, such as pedals, color, and height of flowers, represent the author's profile regarding the level of activity and posting history [9]. Other researchers have investigated the behavioral aspects of Usenet authors over time in predicting the quality of postings [10, 11, 12]. One study has found a closer relationship between the value of messages, and the longevity and the amount of contributions of the author to the newsgroups [12].

These studies have suggested ways of using social interaction to estimate the satisfaction of readers to the content of messages, but have not delivered other properties of the thread. Other characteristics including the activities, history, and thread topic are all important components for recognizing worthwhile threads from noise. Our previous study introduced a visualization method, which presented multiple attributes and those relations through intuitive, but informative visual abstraction [13]. The usability testing demonstrated the feasibility of the system as a supportive visual interface in exploring the bulletin board and also suggested a framework for future directions.

The main goal of this study is the creation of a visual interface that will support the user in identifying the major

discussions of an online message board while reducing the access to the underlying content of threads. By addressing the issues found in existing systems, the new system presents various aspects and correlations of threads more intuitively. The system called as Discussion Forum Visualizer (DifVis), achieves higher visual abstraction by employing additional attributes in presenting other information of the threads on the limited screen space. The following section of the paper introduces a detailed algorithm for projecting several aspects of threads together via a visual representation. Then the paper discusses new issues and future tasks related with the extreme variability of values that have been identified in our preliminary analysis.

2. VISUALIZATION

The DifVis system utilizes three different components to transform the content of threads, user activities bound to the thread, and the predicted value of threads into the visual notations. Each thread in the forum is mapped to a square shape drawing object whose dimension is defined relative to the magnitude of the associated thread. Popularity, activity, and the temporal aspect of the threads are visualized by color components in which the intensity delivers the significance of each characteristic. The last factor used in the visualization, positional information, delivers the relative importance of a thread within the entire message boards.

Visualizing Magnitude of Threads

After reading a number of messages in the forum, and finding a thread related to their interests, a user will post their opinion on the topic. Due to the underlying nature of this posting mechanism, threads which contain an interesting topic to more users will have a higher chance of getting replies. A previous study found a close correlation between the quality of contents and the quantity of posted messages in the bulletin board [12]. Providing the magnitude of each thread to users who are casually navigating the message boards is valuable information necessary as the workspace grows in size.

The magnitude of each thread, namely the total number of words in a thread, is projected as the dimension of the corresponding square object. The posted messages include stop-words that are used to establish the context of information. In order to define the magnitude of each thread based on the meaningful words, a stop-words removal process eliminates insignificant words from the threads [14].

Depending on the history and recognition of the discussion forum, the range of magnitude of threads in the forum varies unpredictably. This causes a significant fluctuation of output when an absolute mapping is adopted to visualize the erratic pattern of magnitude to the dimension of drawing object. In order to create a more balanced visual illustration of the threads, the magnitude of an entire discussion forum is mapped to the dimension of screen space. The dimension of each square is defined relative to the magnitude of the associated thread to the magnitude of the entire discussion forum. The dimension of the square object associated to the thread i, (T_i^D), is computed as

$$T_i^D = \sqrt{(T_i^W \cdot P_{tot}) / \sum_{j=0}^{n} T_j^W} \qquad \text{Eq.(1)}$$

where T_i^W is the total number of words in a thread i, P_{tot} is the total number of pixels forming the viewport of a window, and $\sum_{j=0}^{n} T_j^W$ is the total magnitude of a message board. By utilizing this comparative transformation, the DifVis can display a large number of threads, in which the magnitude distribution is erratic, on a limited screen space. Figure 1 is a sample view of DifVis where each square represents a thread in the discussion forum.

Visualizing References to Threads

The lifetime and popularity of a thread affects its amount of information. A thread having reasonable history generally includes many postings. Although the content-based evaluation is one of the most effective methods to examine the postings, it is impractical to find worthwhile information from a bulletin board without any assisting tools for the navigation. Collaborative filtering can be a practical mechanism for distinguishing noise from constructive information in the information space [7]. But there are many more silent readers than authors in the discussion forum. The user spends more time to read postings in a thread providing their interest than a thread including monotonous contents. The number of references to a thread represents the readers' activities to the messages in a thread. The higher the number of references suggests that the attractiveness of the thread among readers is higher and is a good measure to predict the quality of the thread's contents by proxy.

In the DifVis system, a green color component is used to represent the number of references to a thread. The higher number of references to a thread is mapped to a brighter intensity, which defines the level of green component of the associated square object. The frequency-to-intensity conversion for the number of references to the threads is computed

$$T_i^{G \in C} = (T_i^R \cdot I_{max}) / T_{max}^R \qquad \text{Eq.(2)}$$

where $T_i^{G \in C}$ is the intensity of green component associated with a thread i, T_i^R is the total number of references to the thread i, I_{max} is the brightest hardware intensity of the color component, and T_{max}^R is the highest number of references to a thread of a forum.

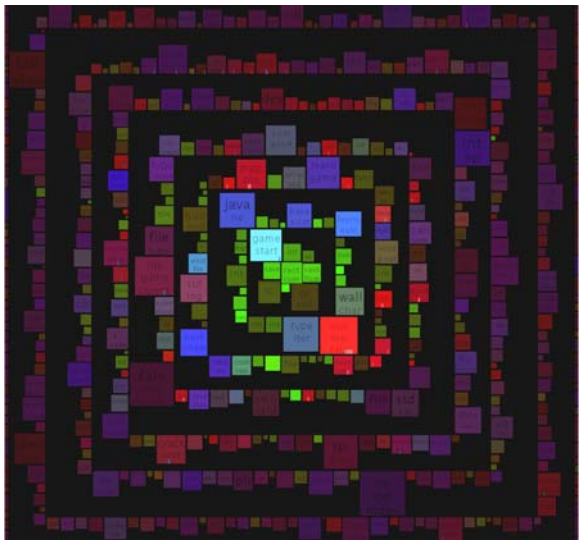

Figure 1. A visualization of a message board by DifVis system.

Visualizing Time Related Factors

Major topics of discussion vary over time, depending on the users' interests and events at that time. Some subjects, i.e. investment and travel tips, have been around the readers for a long time probably with recent postings. Meanwhile subjects like PL/1 programming might have been an interesting topic in past, but now has disappeared from the memory of modern programmers. The time related information of treads are valuable resources for understanding the history of the threads and also the activities of the threads over a temporal domain. These time related attributes also assist readers in understanding the topical transitions of discussions over time.

The DifVis uses a color component, red, for symbolizing two time related pieces of information, the lifetime of the thread and the posting date. Based on the temporal length of the thread and the last message posting date, the intensity of the red component is defined as follow

$$T_i^{R \in C} = ((T_i^L \cdot W_L) / T_{max}^L + (T_i^{P=Last} \cdot W_P) / F_{life}) \cdot I_{max} \quad \text{Eq.(3)}$$

where $T_i^{R \in C}$ is the intensity of red component of thread i, T_i^L is the lifetime of the thread i which is the distance from the first message posting to the last message posting date in number of days. The T_{max}^L is the longest lifetime of a thread in the discussion forum, F_{life} is the lifetime of the forum in number of days, and I_{max} is the brightest hardware intensity of the color component. The $T_i^{P=Last}$ is the last message posting date to the thread i, of which the posting date represents the distance in number of days from the forum initiating date which starts at 0. The W_L and W_P are the weighting factors that define the significance of each time related aspect for computing the intensity of red color. Although further studies are required, commonly the lifetime of a thread is more important than the last message posting date for predicting the overall quality of thread contents.

Visualizing Turbulence of Threads

The popularity of discussion subjects varies depending on the current issues among the users. Higher user access to a thread reflects higher interest to that thread at that time. A thread including a discussion about a long favorite subject among the readers will show relatively constant daily activities to the thread over time. Meanwhile, the users' daily activities to a thread about a periodic issue will considerably change in temporal domain. Especially the users' access pattern to a newly initiated thread regarding a current issue will show an extreme fluctuation of their daily activities.

In order to deliver the users' access pattern to threads, the DifVis system uses the last color component blue. The turbulence of the users' activities is symbolized by the standard deviation of the daily-based number of postings to a thread (T_i^{STDEV}). The T_i^{STDEV} of each thread linearly maps to the intensity of blue color component. The brightest intensity is assigned to the thread that shows the maximum turbulence level of activities

$$T_i^{B \in C} = (T_i^{STDEV} \cdot I_{max}) / T_{max}^{STDEV} \quad \text{Eq.(4)}$$

where $T_i^{B \in C}$ is the intensity of blue component of thread i, I_{max} is the brightest hardware intensity of the color component, and T_{max}^{STDEV} is the largest standard deviation of a thread regarding the daily-based number of message postings to the thread. The above explained three primary colors, red, green, and blue, are mixed together and painted on top of the corresponding square object. The blended color assists the readers to understand patterns of the number of reference, time related issues, and turbulence of threads intuitively.

The Ranking of Threads

Ranking is an effective mechanism that shows the major threads of a forum in the order of its importance to the readers. After analyzing the significance and correlations of the above factors, the rank of importance of threads can be assigned. This is a difficult task that requires analysis of the relationship among various factors and further considers an individual reader's preference and their point of view.

The positional factor is used in the DifVis system to represent the relative worth of a thread in the forum. The order of threads is defined in descending order of weighted sums of the above factors as follow

$$T_i^{Rank} = (Normal[T_i^D] + Normal[T_i^{G \in C}]) * \alpha + Normal[T_i^{R \in C}] * \beta + Normal[T_i^{B \in C}] * \gamma \quad \text{Eq.(5)}$$

where T_i^{Rank} is the rank value of a thread i, $Normal[Factor]$ is the normalized value of the above factors, and α, β, γ represent the weight for the factors. The weight for each factor has been defined through experimental investigations. The thread having the largest rank value is positioned at the center of viewing window. The following threads are positioned in a spiral layout radiating from the center of the screen outward in descending rank value of threads. Although the ranking algorithm needs further systematic investigation in the future, this positional information will assist the general readers in reaching worthwhile threads more rapidly. Figure 2 is a zoom-out view of visualized message board including 2750 different threads.

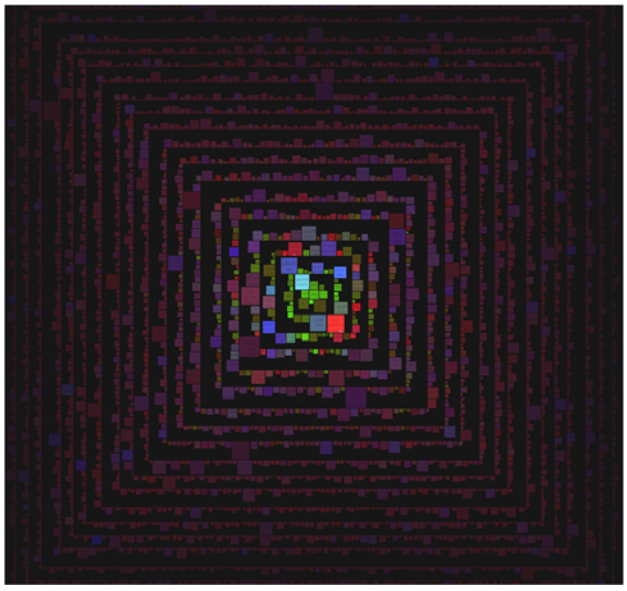

Figure 2. A zoom-out view of visualized message board.

Supporting Tools

The DifVis provides several tools to support the user in exploring the online bulletin board more effectively. To assist the readers in understanding the main subjects of threads without reading the content of messages, the highly occurred

keyword(s) in a thread are displayed on top of the associated square object. In displaying the major keywords of each thread, the DifVis system uses different font sizes to deliver the relative frequency of major keywords in the thread. This pre-analyzed content-based filtering accompanied with the visual representation of the thread will improve the readers' search speed while eliminating the access to the thread including non-relevant information.

The DifVis is equipped with a multi-scale viewing functionality that allows the readers to explore the message board in different levels of visual abstraction. The system provides further information about a thread or message in respond to the user action. For example, when the reader moves the mouse on a specific square object, a tool-tip provides supplementary information such as the title of thread, the author initiating the thread, the number of replies and references, and other suggestive words in the thread.

The user can utilize traditional search supporting tools to customize the output depending on their preferences in exploring the information space. Instead of using the rank provided by DifVis system, the reader can rearrange the visual abstractions based on the posting date, the number of replies and responses, and the lifetime of threads. The filtering functionality allows the reader to eliminate noisy threads such as having low number of messages and references from the viewport. Finally, the user can execute the traditional query-based search on the messages and search for authors to identify social activities of participants.

3. PRELIMINARY ANALYSIS

A pilot study was conducted on five data sets that were collected from a public Internet forum. In order to create a variability of values of attributes, the data sets were assembled from five different message boards. The number of threads in the data sets varies from 2509 to 7732. The range of lifetime of forums was from 30 days to 600 days. These variations of the lifespan and the number of threads from the forums allow us to have data sets that show significant fluctuations of the values of the thread's attributes. A graphical illustration, which is visualized by the proposed method and five different histograms are produced for each dataset. The histograms show the distributional properties of attributes of the threads. The different attributes are: magnitude, reference, mean posting time, lifetime, and turbulence of the threads. Five individual evaluators who have experience in visualization studies participated in the preliminary study. Each evaluator investigated the validity of the developed visual abstraction by comparing the graphical illustration with the individual histograms in turn.

After exploring the visualized discussion forum, evaluators participated in a post experimental interview. All of the evaluators expressed their positive support for the proposed visualization as a search supporting tool in exploring huge text-based message boards. They agreed that the DifVis system significantly improved their ability to recognize the major subjects of discussions in a noisy information space. The DifVis system also helped them to predict the value of the thread, activity within the thread, and popularity of the thread among readers before reading its contents.

4. DISCUSSION AND FUTURE RESEARCH

The number of attributes and the erratic pattern of the distribution make it difficult to develop a visualization algorithm that represents the various factors of message boards compactly and intuitively. The post experimental analysis suggests several tasks for which enhancement in the future will immediately contribute to the improvement of users' search efficiency.

Threads with longer lifespan commonly have more contributions from users than newly initiated threads because of the accumulative characteristics regarding message posting and references. Due to this underlying nature of bulletin boards the visualization can be distorted. The linear interpolation between the magnitude of threads and the dimension of drawing objects might mislead the users from a new thread containing interesting subject matter to somewhat older threads having a long lifetime. A thread having popularity among the readers for a long time generally show an exceptionally large magnitude with a high reference value compared to other threads in the forum. This unbalanced distributional quality of the attributes can generate an improper projection of its real magnitude. The DifVis system defines the level of intensity based on the relative significance of an attribute of a thread to the maximum value of the same attribute in the forum. Assume the brightest hardware color intensity is assigned to a thread with an exceptionally large number of references. Due to the possibly huge difference in the number of references, other threads, with a fair amount of references, will be represented with a lower intensity value than what one would want projected. These aliased visual abstractions can be corrected by either applying a non-linear function for redistributing the range of values or adjusting the magnitude of attributes relative to the lifespan of the associated thread. By achieving the proper balance in the distribution, a newly developed thread gaining rapid popularity will be visualized more notably than a long lasting thread, which has already lost its popularity with the readers.

The turbulence represents the level of variation in the posting of messages over the lifetime of the associated thread. This does not reflect the correlation of the turbulence with the temporal variation. The location of turbulence on the temporal coordinate is another factor to be considered for defining the value of thread. Applying an inverse weight function to the temporal location can be a simple approach to reflect the temporal position in calculating the meaning of turbulence of a thread.

Improper user actions can misrepresent the true importance of a thread. In order to bring the users' attention to a subject continually, sometime a user posts a message to a thread disappearing from the group's interest. This user's intentional action makes the thread current and exaggerates the lifetime of the thread. One way of preventing this falsification is the use of a daily reference counter with the number of days on which messages have been posted. Most of the existing online message boards don't provide the daily-based reference activities. By defining the lifetime of a thread based on the number of actual posting days supplemented by the referential activities to the thread, the issue regarding the falsified lifespan of the thread will be addressed. The daily-based referential activities will also contribute in the computation of the extent of turbulence within a thread.

The most challenging future task will be the implementation of a non-biased usability test. A large scale experiment is an essential task for evaluating the performance

of the proposed method as a search supporting interface. One of the major difficulties for executing a systematic experiment is the lack of a gold standard to be compared to for the evaluation of the system. As used in other studies [12], a survey-based usability test, after a large number of participants used the system for a period of time, should be a feasible approach to assess the effectiveness of the system objectively.

5. CONCLUSIONS

The enhanced system reiterates the feasibility of using a visual representation of threads for navigating a message board filled with numerous non-related or minor subjects. By presenting various attributes of the threads through the visual abstraction, the DifVis system assists the user to recognize more information about the threads than the conventional bulletin board. With the use of an enhanced visualization mechanism, the DifVis system achieves higher visual abstraction and better intuitive representation than our previous system. The preliminary experiment suggests several future tasks that should be addressed to make the DifVis system a practical application. In summary, the DifVis system will be an effective visual interface in exploring discussion forums with the user's perspective cognition. This compact, but informative graphical illustration of threads boosts the user's understanding of the threads without reading the contents, and will eventually contribute to improving the user's search efficiency.

6. REFERENCES

[1] Card, S.K., Mackinlay, J.D., Shneiderman, B., "Readings in information visualization using vision to think," Morgan Kaufmann (1999).

[2] Woodruff, A., Faulring, A., Rosenholtz, R., Morrsion, J., Pirolli, P., "Using thumbnails to search the Web," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.198–205, 2001.

[3] Kim, B., Johnson, P., Huarng, A., "Colored-sketch of Text Information," *Journal of Informing Science*, Vol. 5, No. 4, pp.163–173, 2002.

[4] Shneiderman, B., Feldman, D., Rose, A., Grau, X.F., "Visualizing digital library search results with categorical and hierarchical axes," *Proceedings of the Fifth ACM Conference on Digital Libraries*, pp.57–66, 2000.

[5] Vieas, F.B., Donath, J.S., "Chat circles," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.9–16, 1999.

[6] Sack, W., "Conversation map: a content-based Usenet newsgroup browser," *Proceedings of the fifth International Conference on Intelligent User Interfaces*, pp.233–240, 2000.

[7] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., "GroupLens: An open architecture for collaborative filtering of netnews," *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp.175–186, 1994.

[8] Allen, R.B., "User models: Theory, method, and practice," *International Journal of Man-Machine Studies*, Vol. 32, pp.511–543, 1990.

[9] Xiong, R., Donath, J., "PeopleGarden: Creating data portraits for users," *Proceedings of the Twelvesth Annual ACM Symposium on User Interface Software and Technology*, pp.37–44, 1999.

[10] Boyd, D., Lee, H-Y., Ramage, D., Donath, J., "Developing legible visualizations for online social spaces," *Proceedings of the Hawaii International Conference on System Sciences*, pp.115, 2002.

[11] Donath, J., "A semantic approach to visualizing online conversations," *Communication of the ACM*, Vol. 45, No. 4, pp.45–49, 2002.

[12] Fiore, A.T., Tiernan, S.L., Smith, M.A., "Observed behavior and perceived value of authors in usenet newsgroups: bridging the gap," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.323–330, 2002.

[13] Kim, B., "Astrograph for Exploring Discussion Forums," *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*, pp. 321–327, 2003.

[14] Paice, C.D., "An evaluation method for stemming algorithms," *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.42–50, 1994.