Visualized Index-Based Search for Digital Libraries

Jon Scott, Beomjin Kim, and Sanyogita Chhabada

Department of Computer Science
Indiana University-Purdue University
Fort Wayne, IN, U.S.A.
scotjc01@ipfw.edu, kimb@ipfw.edu

Abstract. As the amount of available information continues to expand, traditional text-based searches for digital libraries and similar systems become increasingly cumbersome to the user. Selection of the best result calls upon the user to compare and contrast top results; this can involve investigative reading of each, to determine what quality and amount of the desired topic is present in each. This paper presents an alternative search strategy, utilizing visualization to relate detailed content information obtained through indexes. By providing such information in a visual manner, the aim is to reduce the burden of investigation placed upon in present systems.

1 Introduction

The proliferation of computing technology and the advance of the Internet has greatly changed the way in which people obtain information. In these days, a person can access to the information which is more than we can handle without the limitation of time and location. The amount of available data grows ever rapidly, but the ability for a user to find their desired information has progressed with less vigor. This is particularly true of highly concentrated sources of a broad array of information, such as documents on the Internet and Digital Libraries.

The current library system provides several attributes associated with books as a response to the users' inquiry. The search results include book title, author, publication year, ISBN number, thickness, etc. However, there is often a large expectation left on the user's ability to read through the results. Furthermore, the text-based approach is non-intuitive and inefficient for finding suitable information through comparison of many possible search results [1]. While ranked search results may assist the user in this endeavor, there is still a reliance on the user investigating the top results individually [2, 3]. This will only become more problematic as the information domain they are applied to continues to grow larger and more complex.

Information visualization is an effective tool that can present a large amount of data compactly, but intuitively for easy comparison. By exploiting users' perceptual cognition, studies have shown that the graphical illustration of data has contributed in improving the users' understanding and reviewing speeds [2, 4, 5]. Borner and Chen explained that there are three common usage requirements for visual interfaces to Digital Libraries: First, to support the identification of the composition of retrieval result; second, to understand the interrelation of retrieved documents to one another,

G. Bebis et al. (Eds.): ISVC 2009, Part I, LNCS 5875, pp. 125–134, 2009.

and last, to refine a search, to gain an overview of the coverage of a Digital Library (DL) and to facilitate browsing and to visualize user interaction data in relation to available documents in order to evaluate and improve DL usage [6].

When using physical books, people tend to view multiple at once; to better compare and review information across multiple sources, and to have a better overall understanding of the domain. In their study, Good et al, identify this to be a major weakness in current DL displays [1]. To address related issues, researchers have conducted studies applying visualization techniques for book searches and presenting various forms of search results [7, 8].

The Graphical Interface for Digital Libraries (GRIDL) is a system that displays a hierarchical cluster of the relevant data to a query on two-dimensional display [9]. This system uses a two-dimensional coordinate, the axes of which are selectable from a variety of different attributes. Results were displayed within each cell as a collection of different size icons, color coded by document type. Marks and his colleagues present a similar approach, based on scatter plots, known as ActiveGraph [10]. Because this approach results in much more node clustering and overlap, a logarithmic transformation is provided, along with the ability to filter out user specified documents. ActiveGraph also provides the ability to specify shape, color, and size of nodes representing documents. By allowing users to manipulate the manner in which data is displayed, these visualizations provided a strong ability to reveal patterns within the data that may not typically be apparent.

These studies mainly focused on aiding the user in comparing the search results effectively by presenting book properties through various visual attributes; but they don't express in detail the amount of content related to user interest. Lin proposed a graphical table of contents (GTOC) that showed the dimension of items in the table of contents based on Kohonen's self organizing feature map algorithm [11]. The paper introduces how documents can be organized and then visualized to allow the user easy access of underlying contents. The GTOC prototype describes various interactive tools to assist the user exploring document contents and analyzing relationships among terms in the table of contents.

The main goal of the research presented in this paper is the development of visualization techniques that will make the user's book search effective by exploiting the book index.

2 Methods

The Visualized Index-Based Search (VIBS) system utilizes an Overview + Detail approach for presenting book search results. This is a visualization technique that uses multiple images to display the entire data space, as well as show an up-close, detailed view of the data [12]. Similar to traditional library searches, the overview will present outline of the book search results through graphical illustration. The user interactively selects a subset of visualized icons that will allow them to execute content level exploration. When a user provides search terms of interests, the Detail view presents a rich visualization of the assets of the given query in a book index with other related information. The resulting visualization supports the user for intuitive comparing and contrasting the selections in greater detail.

2.1 Overview Visualization

The Overview allows the user to perform a general search on the data space, similar to traditional library tools. The current prototype utilizes a title based search, although a more robust implementation would make use of additional categorization provided by the environment. To address the inefficiency of space utilization of text-based book search application, VIBS presents the results using a grid based layout. The X and Y axis represent individual book attributes, including author, publication year, number of pages, and review details. These are freely changed via drop-down selection boxes; providing greater control over the result display and assisting with user understanding [9]. Book nodes are located accordingly.

Fig. 1. Overview display (left). Close-up of Overview with tool-tip (right).

Total page count defines the radius of the corresponding circular icon. The books are classified into *N* categories depending on the number of pages, where each group is mapped to a predefined radius. Within the same category, the radius is linearly interpolated according to the page count of the item. By properly utilizing non-liner and linear transformation, the VIBS can display variable sizes of book volume on the limited screen space while still delivering its magnitude. Through the visual illustration of book dimension, the users can intuitively estimate the amount of content available.

VIBS uses the RGB color model to deliver other attributes to the user. As part of this, the Overview employs collaborative filtering to relate the perceived user value of each result. Collaborative filtering is a type of content-based filtering which utilizes the opinions of users who have already evaluated the quality of information [13]. This is done by collecting user reviews and their fidelity, typically found on most book merchant web-sites. The user's satisfaction with a book corresponds to the green color component of the circular node. The average review score a book receives is mapped to the green intensity, as shown by equation 1, where C_i^G is the green intensity associated with a book *i*, E_i^G is the average user rating of book *i*, I_{max} is the brightest hardware intensity of green, and E_{max}^G is the highest available user rating.

$$C_i^G = (E_i^G \cdot I_{max}) / E_{max}^G \quad (1)$$

The accuracy of collaborative filtering is highly dependent on the number of evaluators. Because of this, the user rating of a book can be unreliable if the number of individual reviews is relatively low [14]. To deliver the fidelity of user ratings of the book, the Overview utilizes the blue color component to represent the number of unique evaluators who rated the book. The red color components relates the publication year, as more recent works are likely to hold more up to date information. These two colors are treated similarly to equation 1. The three color components are blended together to form a color Ω_i which is the final color of the circular node. The end result is that books with high ratings, a large reviewer base, and more recent publication will have a prominent color closer to white. To assist the user, mouse-over tooltips are employed to relate the color data individually along with the title and author (figure 1).

2.2 Detailed Visualization

The Overview interface assists the user to compare attributes associated with books. Meanwhile the Detail visualization focuses on showing the amount of searched contents of selected books through a graphical illustration. The VIBS system utilizes the book index to present the amounts of related contents, search terms distribution, and associated sub-terms to search terms in the index pages.

When a user enters a search term of which information they are looking for, the Detail view displays the corresponding data using a radial tree structure. In the display, terms are represented as circles with the center representing the given search term and sub-terms are displayed as nodes branched off of the center circle (Figure 2). With the similar way used in Overview visualization, the VIBS system makes use of multiple attributes in relaying a range of information that will assists the user's finding a book of their interest.

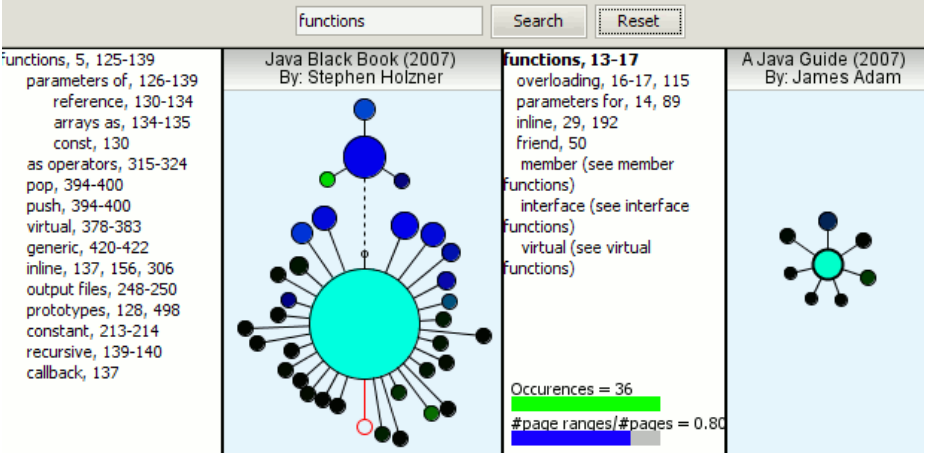

Fig. 2. Detail view showing (left) diagram with expanded sub-term and an overflow node, and (right) context highlighting of text list and bar display

The total number of pages associated with a term in the index could be considered to have strong correlation to the amount of information provided by the work on that subject. In the Detail view, this important measure is naturally mapped to the radius of the term icons. This is a cumulative page total of the node's associated term, and all of the term's subordinates within the index. This graphical illustration allows the user to gain a general understanding of how different book results compare, without having to compare each level of sub-terms in detail.

The VIBS system applies color-coding to the terms as a visual abstraction occurrences and concentration associated with that item. This is intended to provide the user with an understanding of the comparative value of the different entries being displayed. The green intensity represents the total number of occurrence of the term throughout the index. A larger number of occurrences could be indicative of the term having broader, or perhaps more complex coverage within the book. The system applies normalization to the occurrence of all terms in the index that results in the node with the highest occurrence having the brightest green intensity, and the node with the least having the darkest.

Blue represents the concentration of sub-terms appearance within the book. A term which appears in two books with the same number of pages allocated by both indexes may not necessarily indicate equal coverage on the topic. If one index were to have the term listed as a single contiguous section of pages while the other had each page listed independently, the former would be considered to likely have a more meaningful coverage of that term. The ratio of page continuity to all pages containing a sub-term determines the blue intensity of color-code by equation 2, where C_i^b is the blue intensity for sub-term i, $\sum_{j=0}^{n} P_j^C$ is the number of individual page ranges for i, T_i is the total number of pages for the term i, and I_{max} is the brightest intensity of blue possible on the machine. This gives terms with more concentrated information a stronger representation in the visualization.

$$C_i^b = (1 - (\sum_{j=0}^{n} P_j^C \ / \ T_i)) * I_{max} \tag{2}$$

Displaying the relationship between search term and its sub-terms is valuable information for content search. A sub-term which has page allocation close to the search term will likely have a stronger correlation than one which is on the opposite end of the book. This relationship is presented by the distance between the parent and sub nodes within the visualization. The magnitude of the distance is average distance between the search term and child page occurrences, normalized across all terms displayed. This is computed with equation 3 where D_j is a normalized distance between the root node and sub-node j, P_j^k is the page number contains sub-term j, $\overline{Root_i}$ is the mean page number having the root word i, and NORM and ABS are a normalize and absolute function respectively.

$$D_j = \mathrm{ABS}(\mathrm{NORM}(\sum_{k=0}^{n} (P_j^k - \overline{Root_i}) \ / \ \mathrm{N})) \tag{3}$$

Sub-terms with a larger number of associated pages are deemed more important, and are subsequently given display priority. These children nodes are arranged around the parent top to bottom in decreasing value, ensuring the user can readily identify which items are potentially more valuable for the particular work being displayed. In the

event that the space around the parent node can't accommodate any additional terms, the remaining un-drawn entities are collected in a single node, indicated via a red outline. As higher priority terms are drawn first, only lower value items will be present within this 'overflow' node. A user can still investigate these items by expanding the node.

Space is also preserved by displaying only the first level of sub-terms around the root as these are considered higher value than subsequent child nodes. Terms with hidden children are indicated by a jagged, dashed outline. When the users want to explore deeper into the index hierarchy, they can freely expand these elements by clicking on them; clicking the root will expand all such nodes. When a node is expanded, it is drawn further away from the parent in order to avoid overlapping. To maintain the information originally relayed by the branch length, a marker is drawn to indicate the original length.

The Detail view interface supports additional tools to assist the user. This is in the form of an information panel that updates based upon the user mouse activity with the display. By hovering over a node, an information panel will appear, highlight the term in a text list. A histogram reflecting the value of the number of occurrences, and page range ratio is also shown. This will support the user in keeping track of what different topics are available in the different books, while also relating more specific information for the different attributes.

3 Experiment

To investigate the effectiveness of the VIBS system as a search supporting tool, we conducted survey-based usability tests that collected users' opinion to the VIBS over traditional text-based search systems. For the experiments, two testing applications were implemented using Java. A text-based application was constructed simulating the conventional library book search system. The second was implemented as explained in Section 2.

Seventeen students majoring in computer science participated in the experimental sessions. Participants had no problems in color perception, were comfortable using a mouse-driven GUI application, and had no previous exposure to VIBS. Sessions were held in groups of three to five, and started with a brief orientation. This orientation presented the functionality provided by both systems along with an explanation of the visualization. Each participant was then given time to investigate the systems until they were comfortable using it.

Next, the testers were asked to search books by using the text and visualization based approaches independently. Both applications show the overall view first either text-based or graphically forms respectively. From the overall view each student selected desirable books by comparing their different properties while utilizing the interactive functionality of the system. The selection of books from the overall view leads the user to the detail view, where participants are able to perform searches on the indexes. From this view, participants can analyze the index-level information related to their queries which will assist them to understand better about the underlying contents; eventually the better chance to find books has more useful or related information.

Participants were then asked to fill out a post experimental survey. The survey was designed to determine user satisfaction about the proposed visualization over a traditional text-based approach. Comments and suggestions on how the visualization can be practically deployed to fill such a role in the future were also collected.

3.1 Survey Results

The questionnaire has 14 questions which utilize a 5 point Likert scale; 5 indicating the highest level of satisfaction, and 1 the lowest. Table 1 below shows the survey results that summarizes the user's feedback to VIBS over a text-based library search system. 16 out of 17 participants stated that they had previous experiences using text-based book search system. Overall users were satisfied with both the overall and detailed view of the proposed system. Although it was expressed in their comments indicating unfamiliarity of visualization system over the current text-based interface, 53% of participants stated that they are more satisfied using VIBS than a text-based approach, 35% of the participants had a neutral stance, 12% of participants still prefer to use text-based book search interface. 82% of participants responded that the overview visualization improved their ability to identify desirable book. For questions specifically associated with the overall view, determining the overall concentration of information for individual books, as well as comparing multiple books to decide which is most appropriate, was considered a strong benefit provided by the prototype. Participants strongly agreed that presenting a visual abstract of book attributes was meaningful for their search activities. They also expressed that the grid-based layout assisted them in comparing multiple books and the selectable axes were very helpful and assisted them in identifying desirable results.

Table 1. Post experiment survey results as percentages

Questions	Pos	Neut	Neg
VIBS overall was preferable to a text-based search	53	35	12
The Overview improved identification of desirable books	82	6	12
The choice of visualized attributes provided a meaningful search environment.	76	18	6
Use of the selectable axis facilitated a better understanding of a set of books.	71	24	5
The Overview helped in selecting a subset of books.	65	29	6
It was easy to discern book attributes based on node color	53	18	29
The Detailed view was preferable to a text approach	65	24	11
The radial layout of the Detail view was intuitive	65	18	17
The Detail view made concentrations of information easy to identify	76	24	0
It was easy to discern term attributes based on node color	47	24	29
The detail view made relevant book selection easier	88	12	0
The relation between a term and it's child was clear	65	35	0
The Detail view helped identify terms related to the search	94	6	0

The participants also replied positively to utilizing visualization for presenting index information of books. 65% of participants satisfied using the visual interface for presenting book contents, meanwhile 11% of users prefer to explore book indexes in text. Users responded optimistically to questions asking the effectiveness of radial tree visualization that displays hierarchies of index terms and its magnitude. Especially participants strongly agreed on two facts that the detailed visualization helped them to choose books with more relevant information and to understand other term related with the search term to 88% and 94% respectively. Meanwhile in both views, the primary recommendation provided by users was to make the color blending of the visualization more intuitive. The post-experiment interview showed that this stems primarily from the fact that some users were unfamiliar with the RBG color blending scheme, having had little contact with it in their daily lives.

4 Conclusion and Future Work

Overall, feedback was positive toward the VIBS system. The testers found it to be an interesting and robust alternative to more traditional search methodologies. While there were users that still held preference toward a traditional text-based approach, several brought up the topic of familiarity. Although these types of searches have the weaknesses outlined earlier, they benefit in wide-spread usage and familiarity among users. Even though the testers had limited exposure to the VIBS system prior to the experiment; that more than half would prefer it over a traditional text-based search is very promising.

Taking a closer look, the Overview was well liked by most users. Similar comments were made regarding familiarity about this display; the chart-like layout was also readily understood, and was an aspect of the view that users felt improved their understanding of the data. The interaction provided through the selectable axes assisted in this, as users appreciated the ability to tailor the display closer to the characteristics more relevant to his or her interest.

Although the Detail view lacked the familiarity present in the Overview, participants still responded favorably. The ability to compare multiple sources at once in detail was an aspect of the system users were very appreciative of. One important point in particular is the high percentage of users that found this view greatly enhanced the identification of other important terms related to the original query. This exploratory aspect of the system is important, in that it can lead users to more appropriate terminology to refine the search. Additionally, it could expose users to other concepts contained within each work that, while not to original goal, could be something which would sway the user to placing more value in one result over another.

The primary issue highlighted by the experiment is user difficulty with interpreting the color codes of both views. Although around half in both instances were comfortable with this aspect of the system, around 30% had trouble with it. The post-experiment interviews provided two main causes for the diverging opinions. First, the RBG color model was not familiar to some. These users cited heavy exposure to the RBY model as being a source of confusion when interpreting the displays. The increased unfamiliarity left those individuals feeling more comfortable using a text-based search.

The other difficulty reported by users was in determining the relative value of one result with another. This could be, in part, a result of the human eye being more sensitive to some colors rather than others. For example, green-yellow colors have the strongest reception, which could mislead a user into considering a result with this color to have more overall value than another when that may not be the case [15]. Researching and examining alternate color models will be one of the challenging task for future work on the VIBS system.

The other area for future work is with regard to content analysis. As books and similar works move toward electronic rather than traditional hard-copy formats, the use of indexes in the traditional sense may become obsolete. For VIBS to be viable in such a scenario, full content analysis would be an alternative for classifying information. Instead of relying upon pre-existing documentation of content, the system could examine the work in full, determining such items as key words and topics, along with their frequency.

In a similar vein, the inclusion of a thesaurus-like aspect would also be a potentially valuable addition. The current Detail view arranges sub-terms solely by content size, a redundant expression to the size of the nodes. Instead, sub-terms could be grouped according to similarity, which would serve to strengthen the exploratory aspect of the system, and improve the user's experience.

By working toward the solutions and ideas presented here, continued work on the VIBS system will be better able to address some of the difficulties presented through the experiments. This will also serve to improve upon its strengths, and allow for a more comprehensive alternative to more traditional text-based search methodologies.

References

1. Good, L., Popat, A., Janssen, W., Bier, E.: Fluid Interface for Personal Digital Libraries. In: Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries, pp. 162–173 (2005)
2. Veerasamy, A., Heikes, R.: Effectiveness of a graphical display of retrieval results. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 236–245 (1997)
3. Dushay, N.: Visualizing Bibliographic Metadata – A Virtual (Book) Spine Viewer. D-Lib Magazine 10(10) (2004)
4. Hawkins, D.T.: Information Visualization: don't tell me, show me! Online 23(1), 88–90 (1999)
5. Kim, B., Johnson, P., Huarng, A.: Colored-sketch of Text Information. Journal of Informing Science 5(4), 163–173 (2002)
6. Borner, K., Chen, C.: Visual interfaces to digital libraries: motivation, utilization, and socio-technical challenges. In: Börner, K., Chen, C. (eds.) Visual Interfaces to Digital Libraries. LNCS, vol. 2539, pp. 1–9. Springer, Heidelberg (2002)
7. Shen, R., Vemuri, N., Fan, W., Torres, R., Fox, E.: Exploring digital libraries: integrating browsing, searching, and visualization. In: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, pp. 1–10 (2006)
8. Silva, N., Sánchez, A., Proal, C., Rebollar, C.: Visual exploration of large collections in digital libraries. In: Proceedings of the Latin American conference on Human-computer interaction, pp. 147–157 (2003)

9. Shneiderman, B., Feldman, D., Rose, A., Grau, X.: Visualizing Digital Library Search Results with Categorical and Hierarchical Axes. In: Proceedings of the 5th ACM conference on Digital Libraries, pp. 57–66 (2000)
10. Marks, L., Hussell, J., McMahon, T., Luce, R.: ActiveGraph: A Digital Library Visualization Tool. International Journal on Digital Libraries 5(1), 57–69 (2005)
11. Lin, X.: Graphical Table of Contents. In: Proceedings of the first ACM international conference on Digital libraries, pp. 45–53 (1996)
12. Baudich, P., Good, N., Bellotti, V., Schraedley, P.: Keeping Things in context: A comparative Evaluation of Focus Plus Context Screens, Overviews, and Zooming. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 259–266 (2002)
13. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 175–186 (1994)
14. Allen, R.B.: User Models: Theory, Method, and Practice. International Journal of Man-Machine Studies 32, 511–543 (1990)
15. Foley, J., Van Dam, A., Feiner, S., et al.: Computer Graphics: Principles and Practice, 2nd edn. Addison-Wesley Publishing Company, Reading (1996)