SearchViz: An Interactive Visual Interface to Navigate Search-Results in Online Discussion Forums

by

Vishal Vimal Mehta

A Thesis Presented in Partial Fulfillment of the Requirements for the Degree
Master of Science

Approved November 2015 by the Graduate Supervisory Committee:

Ihan Hsiao, Chair
Erin Walker
Mohamed Sarwat

ARIZONA STATE UNIVERSITY
December 2015
ABSTRACT

Online programming communities are widely used by programmers for troubleshooting or various problem solving tasks. Large and ever increasing volume of posts on these communities demands more efforts to read and comprehend thus making it harder to find relevant information. In my thesis; I designed and studied an alternate approach by using interactive network visualization to represent relevant search results for online programming discussion forums.

I conducted user study to evaluate the effectiveness of this approach. Results show that users were able to identify relevant information more precisely via visual interface as compared to traditional list based approach. Network visualization demonstrated effective search-result navigation support to facilitate user’s tasks and improved query quality for successive queries. Subjective evaluation also showed that visualizing search results conveys more semantic information in efficient manner and makes searching more effective.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>v</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Problem &amp; Goal</td>
<td>1</td>
</tr>
<tr>
<td>Motivation</td>
<td>2</td>
</tr>
<tr>
<td>Research Questions</td>
<td>3</td>
</tr>
<tr>
<td>Contribution</td>
<td>3</td>
</tr>
<tr>
<td>2 LITERATURE REVIEW</td>
<td>4</td>
</tr>
<tr>
<td>Discussion Forum Visualization</td>
<td>4</td>
</tr>
<tr>
<td>Visual Approaches of Recommendation &amp; Recommenders</td>
<td>5</td>
</tr>
<tr>
<td>Visualization in Search Context</td>
<td>6</td>
</tr>
<tr>
<td>3 SEARCHVIZ: VISUAL INTERFACE TO NAVIGATE SEARCH RESULT</td>
<td>8</td>
</tr>
<tr>
<td>Provided Search Functionalities</td>
<td>9</td>
</tr>
<tr>
<td>Semantic Information Conveyed by Visual Interface</td>
<td>9</td>
</tr>
<tr>
<td>Viewing Details through Ranked List</td>
<td>14</td>
</tr>
<tr>
<td>4 SYSTEM IMPLEMENTATION DETAILS</td>
<td>15</td>
</tr>
<tr>
<td>System Architecture</td>
<td>15</td>
</tr>
<tr>
<td>Database Design</td>
<td>17</td>
</tr>
<tr>
<td>Data Collection Procedure</td>
<td>18</td>
</tr>
<tr>
<td>5 USER STUDY DESCRIPTION</td>
<td>19</td>
</tr>
<tr>
<td>Task Description &amp; Procedure</td>
<td>20</td>
</tr>
<tr>
<td>Conditions for User Study</td>
<td>21</td>
</tr>
<tr>
<td>Background Analysis of Study Participants</td>
<td>22</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Comparing Information Conveyed via Visual Components</td>
<td>14</td>
</tr>
<tr>
<td>2. Query Statistics Summary</td>
<td>31</td>
</tr>
<tr>
<td>3. An Example of Query Elaboration &amp; Query Concept Elaboration</td>
<td>32</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1. Discussion Forum with Interactive Visual Interface to Navigate Search Results: (SearchViz).................................................................................................................. 8
2. Network Visualization Generated for Query ‘Implement Inheritance’..........10
3. Image Displaying Varying Color Intensity Within Same Cluster to Convey Query Similarity for Each Question Circle.................................................................11
4. Radius of Circle Depicting No. of Answers for Respective Question Circle.....12
5. Width of Links Representing Frequency of Keyword Within Corresponding Question Text.................................................................................................................12
6. System Architecture of SearchViz..................................................................15
7. Class Diagram of SearchViz..........................................................................16
8. Schema Design for SearchViz..........................................................................17
9. Study Procedure..................................................................................................20
10. Chart Displaying Programming Experience of Study Participants...............23
11. Chart Displaying Java Programming Experience of Study Participants........23
12. Chart Displaying Sources of Learning and Problem Solving......................24
13. Precision Differences Between Interfaces.....................................................28
14. Sample Log File of Participant Displaying Role of Visual Interface..............28
15. Average Time per Problem for Task 01..........................................................29
16. Average Time per Problem for Task 02..........................................................30
17. Comparing Accuracy Values for Various Problems Attempted by Two Versions of System..................................................................................................................33
18. Comparing Completeness Values for Various Problems Attempted by Two Versions of System.........................................................................................................34
<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
</table>
Chapter 1: Introduction

Forums or discussion boards are popular troubleshooting technologies, especially for online learning. They are free, open and fast-growing online communities (homework-help sites, discussion forums for MOOCs courses etc.) that draw massive user-initiated efforts to contribute, to consume and to interact with content on the site. In the context of learning programming, such free online discussion sites allow programmers and learners to reach out for help so that they can freely discuss programming problems, ranging from general to specific and simple to complex topics. Examples of such sites are stackoverflow\(^1\), Dream.In.Code\(^2\), Tutorialspoint\(^2\), CodeProject\(^2\), etc. A 2011-2012 analysis of 28 million course papers submitted to Turnitin\(^3\) (2013) revealed that social networking and other user-generated content sites were cited in 23% of the papers written by students in higher education institutions. The same paper also lists online Q&A sites such as Yahoo! Answers and WikiAnswers as second only to Wikipedia among sources used by students. These sites therefore not only throw open unbounded topics in the form of questions and answers, but are especially attractive for open-ended problem discussions.

A. Problem & Goal

The drastic shift in momentum of learning opportunities from traditional learning objects (textbooks, intelligent tutoring systems, worked examples etc.) to

\(^1\)Stack Overflow (stackoverflow.com) is an online community of question and answer site for professional and enthusiast programmers. It's built and run as part of the Stack Exchange network of Q&A sites.

\(^2\)Dream.in.code (www.dreamincode.net) is a reputation-based online community for programmers and web developers. Tutorialspoint (TutorialsPoint.com) is a hub that hosts online programming tutorials for learners. CodeProject (www.codeproject.com) is another online community for programmers.

\(^3\)Turnitin.com is primarily a plagiarism detection service that analyzes student papers uploaded by subscribing institutions, but their analysis also includes identifying the sources students cite in their papers.
community help is becoming prominent but not yet fully supported and comprehended. Among all the discussion forums, search functionality is usually provided for users to filter large volume of discussion posts in an online forum. Some common algorithms are typically deployed to rank the search results, such as PageRank, HITS, or simple keyword matching etc., which present the search results in a form of ranked list. The ranked list representation of search results on discussion forums presents two major problems: 1) it assumes programming learners know how to search; 2) it demands intensive reading-labor to filter the content quality in the large and ever growing corpus, especially inefficient for novices. These trends may end up resulting in expert-oriented communities rather than creating open public available technology for all.

My goal is to investigate alternate ways to represent relevant search information on online discussion forums with aim to reduce user efforts in finding what they are looking for.

B. Motivation

We begin to see more and more intelligent interfaces to support general browsing, exploring, searching and adaptation [15-18, 20-22, 6-8]. However, most of these approaches follow traditional hierarchical clustering paradigm, such as Faceted Browsing, Exploratory Search etc., which utilizes the breadcrumb trails to facilitate searches. Such sites typically still rely a lot on query issuers’ efforts (comprehensive query bank and massive reading efforts to identify relevant documents). They are usually designed with several filtering features, such as sorting, voting, badges and other features to filter the content and help readers sift through massive amount of user-generated contents. However, I argue that these filters tend to point out the extremes (i.e. good/bad or recent/outdated) but not represent the overviews, especially the interrelations (i.e. the 5th answer of Q1 and the best answer of Q4 share the same concepts; the best answer of Q1 can be good enough answers to multiple
other questions too). My goal of this thesis is to investigate an alternative solution to present search results from one of the most fast-growing information retrieval systems i.e. discussion forums. I designed and studied a flat and responsive interactive visualization to navigate search results on large volume of programming discussion forum posts. I implemented a network interactive visualization that visualizes users’ search results in a programming discussion forum. I hypothesized that visualizing search results via network visualization (via various abstract visual cues encapsulated in network visualization) will support user’s navigation in search and exploration process in online programming discussion forums.

C. Research Questions

I will investigate about how visualizations assist users in exploring discussions on online forums followed by researching effectiveness of search-results navigation via an innovative two-dimensional network visualization, with the premise that not-everyone-is-an-expert-searcher. User study is also aimed to reveal various sources for learning programming and getting information to perform programming related problem solving tasks among students.

D. Contribution

A functional visualization interface to navigate search-results on online discussion forums. Interface will provide overview of search-results and can be used to navigate to appropriate posts quickly and efficiently. Findings of user study to investigate usefulness of visualizations in navigating posts on discussion forums.

The next sections present overview of developed visual interface (SearchViz) to explore discussions, system design, experimental design, and evaluation results. Finally, I summarize the work and discuss limitations and future work.
Chapter 2: Literature Review

I reviewed three streams of literature related to my thesis: discussion forum visualizations, visual approaches of recommendation & educational recommenders and visualizations in search context.

A. Discussion Forum Visualizations

Similar attempts to incorporate visual interface on online discussion forums have been made previously. In their work [1]; Heer, Viegas and Wattenberg designed an interface supporting asynchronous collaboration in context of information visualization with underlying assumption that visualization improves ability to process large amounts of data. User study also showed that users were able to identify trends within dataset by discussing, sharing and commenting on visualizations generated for the dataset. This provided me motivation to include visualization interface to provide higher level overview of the search results in online discussion forums because forums contain large amount of data and effective visualization can be used to represent this data concisely.

In [2] Hoque and Carenini presented a novel topic modelling system to extract key points from conversations on online discussion forums. They provided interactive visual interface to revise the model to generate highlights on the fly as per user feedback. This paper presented a novel human-in-the loop topic modelling approach combined with visual interface to support exploration of large conversations. This motivated me to generate the visualization for SearchViz by keeping users in mind and providing them enough flexibility to configure the interface as per their preference by controlling the level of details and position of components.

In their work [3]; Vassileva and Sun studied about impact of visual interface on participation of users in class based online communities. Visual interface shows
contributions of users to an online community to encourage social comparison and
more participation. Through user study they found that visual interface makes users
significantly more proactive in the community by showing increased participation and
engagement. This also provided belief that introducing visual interface (SearchViz) to
navigate search results on online discussion forums will impact user contribution
towards community strongly and positively.

In [4] Hsiao and Awasthi discuss about Topic Facet Modeling (TFM) approach
to extract forum post semantic for uncovering latent structural topics. TFM looks like
a promising approach to automatically generate discourse semantics for large scale
dynamic discussions. The authors also implemented a semantic visual analytics
interface in visualizing forum posts semantics [5].

B. Visual Approaches of Recommendation and Educational
Recommenders

LinkedVis [6], PeerChooser [7], SetFusion [8] are among several other
projects that include richer user interface to provide hybrid recommendations.
Educational recommenders have been successfully deployed across many disciplines
to assist information seeking, exploration, discovery, diversification, enhancing user
experiences and engagement [9]. However, they have not yet been fully exploited in
the educational sector. Currently, majority of educational recommenders are
conceptual designs or prototypes like [13] and [14]. Only a few have been reported
with real system usage as in [12]. Secondly, the most common approach in
educational visual recommenders is to provide relevant reading objects (new
resources, learning partners, or learning object sequences), for instance a suitable
next step to learn a new concept or to help solve a problem as suggested in works
[13] and [14]. There has been less work focused on dynamic recommendations
based on system interactivity.
C. Visualization in Search Context

There are several intelligent user interfaces supporting general browsing, exploring, searching and personalization such as PEx-WEB visualizing content based results similarity [15]; Resultmaps which implements TreeMap to visualize search results [17] which again follow the breadcrumbs trails paradigm to facilitate search.

In their work [16] the researchers made great use of visualization by combining it with web search. They presented a interactive visualization system by combining a number of algorithms to help users analyze and navigate through a collection of web pages. Various case studies carried out revealed that this system was useful for exploration of data. This encouraged me to combine the visualization with online discussion forums to enhance the search result navigation by reducing reading efforts made by users to find relevant content.

The idea of using network visualization for SearchViz by clustering search results for efficient navigation support was also influenced by similar work [18] carried out by Omar, Michael and Recardo. They designed an add-on to combine with traditional information retrieval systems to present and cluster the search results along timelines. In [19] the researchers identify that in exploratory search, there is significant room for improvement in the way search results are returned to the users. Exploratory searches are performed by the users who need to learn, discover and understand complex topics. Hence in such cases presenting web search by enabling users to visualize, manipulate and organize their search results is more effective to serve search purpose of the users. SearchViz aids users in navigating search results on online discussion forums where users may have poorly formed search goals and visualization interface can greatly assist in finding relevant information by summarizing search results and providing efficient navigation to jump to respective results quickly.
In [21], Hearst talks discusses about two methods i.e. Clustering and Faceted categorization that are used to generate useful groupings and aid in designing interface to support exploration using groupings. The idea of utilizing clustering to generate network visualization was also based on information presented in this paper. According to Hearst, advantages of clustering are that it can be easily applied to text collections. Since programming discussion forums have majority of textual content; clustering relevant search results makes sense. Hearst also notes that clustering is useful for clarifying and sharpening vague queries, by showing users dominating themes of search results. Hence the user queries need not to be complete for efficient retrieval. Users can improve their subsequent queries by analyzing query details. This was also one of the factors of choosing clustering to generate network visualization for SearchViz.

Based on literature review of various works on integrating visual interface with discussion forums, recommenders, search context; I learnt about their capability to summarize large data effectively, display clustering, improve user experience/interaction with system and engaging users. This learning gives me enough confidence to study about the impact of combining interactive user interface with online discussion forum with the aim to improve search result navigation and make searching more effective.
Chapter 3: SearchViz: Visual Interface to Navigate Search-Results in Discussion Forum

As you can see in figure 1; on issuing query ‘implement multiple inheritance’ the visual interface displays search results in form of clusters centered around keywords like class, object, inherit, implement, etc. Each cluster is represented with distinct color and recommended question posts are categorize into appropriate cluster depending on their textual content. Visual interface presents higher level overview of the search results and can be utilized to navigate to appropriate discussion forum posts by clicking on the question circles. Apart from clustering; we also represent each question circle by most frequent keywords appearing in the document. The dataset for this sample discussion forum is generated by downloading content from Stackoverflow website.

I used Apache Lucene library to index the questions on forum according to their plain text as well as program code content to support multiple search features and enable efficient retrieval for various user queries.
Various functionalities and visual information conveyed by the interface have been discussed below.

**A. Provided Search Functionalities**

1. **Text based Search**

   In this mode of operation the query string is compared with the textual content of questions in forum posts and related documents are retrieved by searching through text version of index generated by Lucene. Similarity of query to question posts is calculated by using Vector Similarity Model.

2. **Code based Search**

   Posts on discussion forums may also have code fragments associated with them. Code based search enables users to retrieve questions with specified code fragments. User query string is compared with code of questions to retrieve search results. In this mode; the network visualization for retrieved search results is generated based on text content of questions while ranked list presents semantic code similarity measure.

   Users can utilize either of these search functionalities to search for relevant content on discussion forum.

**B. Semantic Information conveyed by Visual Interface**

1. **Network Visualization to provide overview of Search Results**

   Visualization shows search results of query in form of clusters centered around keywords. Each cluster is displayed with distinct color and recommended questions are linked to corresponding keywords according to their textual content to generate network visualization. Link of question circle to a keyword within cluster provides higher level overview about its textual content. I implemented a two-layer view by following Sheiderman’s visual information seeking mantra “overview first, zoom and filter detail on demand” [26]; to display overview of each question circle.
In default view, each question circle displays one keyword per question as shown in figure 2. But users can switch to details view by clicking on ‘Show Cluster Details’ button on left side of visualization to get detailed view displaying three keywords per question as shown in figures 3 and 5. Users can switch back to default view by clicking the same button again.

Figure 2: Network visualization generated for query ‘implement inheritance’

Apart from keyword connection within cluster; each question circle is also represented by other significant words from within its text. User can control the level of detail for question circle representation appropriately via “Cluster Details” button available on left side of the visualization. By default; question circle uses single significant word to represent each question. This can be changed to display three significant words per question in order to get deeper insight about each question.
from the visualization. “Cluster Details” button allows users to adjust such level of
details.

Visualization interface used to navigate search results was implemented using
frontend technologies HTML5, CSS3, D3 library, Javascript and jQuery.

2. Color Intensity to represent Query Similarity

![Image of varying color intensity within same cluster](image.png)

Figure 3: Image displaying varying color intensity within same cluster to convey
query similarity for each question circle

In the visualization; similarity of a question circle to query is represented via
color intensity. Hence question circles with higher color intensity have more Tf-Idf
similarity to query as compared to ones with low color intensity. This information can
be used to make appropriate navigation choices by users.

3. Radius of Question Circle to represent number of Answers

This information is conveyed in visualization through radius of question circle.
Question circle with larger radius contains more answers as compared to question
circle with smaller radius. This conveys information about length of discussion on a
particular question post. Users may be interested in viewing long discussions for
some question topics while opting for brief discussions for others. Visual interface
provides cues to make such judgements to the users.
Questions on discussion forums may have multiple answers by various users.

4. **Width of links to represent Frequency of Keywords**

![Diagram showing links between keywords and question circle]

Figure 5: Width of links representing frequency of keyword within corresponding question text

Width of link between question circle and keyword is used to represent frequency count of that keyword within text of question. Hence wider links have more occurrence of keywords within that question text as compared to question with narrow links.

5. **Comparing Information conveyed via Visual Components**

Visual interface provides multiple cues regarding similarity of questions, length of discussions and higher level overview of question text as well.
Comparing Color Intensities of Question Circles

Question represented by ‘interface’ has higher Tf-Idf similarity as compared to question represented by ‘class’ keyword.

Question represented by ‘class’ keyword has lower Tf-Idf similarity as compared to questions to left of table.

Comparing Radius of Question Circles

Question circle represented with keyword ‘loop’ has less number of answers.

Question circle represented with keyword ‘loop’ has more number of answers.
Comparing Width of links between Questions and Keywords

<table>
<thead>
<tr>
<th>Question circle represented with</th>
<th>Question circle represented with</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘class’ has more frequency word</td>
<td>‘interface’ has less frequency word</td>
</tr>
<tr>
<td>count of keyword ‘method’ within its</td>
<td>count of keyword ‘method’ within its</td>
</tr>
<tr>
<td>question text.</td>
<td>question text.</td>
</tr>
</tbody>
</table>

Table 1: Comparing Information conveyed via Visual Components

6. Using Visualization for Navigating search results

Visual interface provides multiple cues regarding similarity of questions, length of discussions and higher level overview of question text as well. User can use this set of information to navigate to appropriate question(s) directly by clicking on the question circle in visualization. Corresponding question text will be highlighted directly in the ranked list and scrolled into user view for further inspection.

C. Viewing details through Ranked List

Questions matching user query are represented in ranked list according to decreasing Tf-Idf similarity measure towards right side of the visual interface. Users can click on ‘View Answer’ button to analyze answers for desired questions. Many times question posts on discussion forums have code fragments associated with them. Such attached code fragments for questions or answers can be viewed by switching to ‘Code’ view from navigation panel above the ranked list. Default view is set to ‘Text’ view for questions and answers.
Chapter 4: System Implementation Details

This section covers various design and implementation details of SearchViz: An Online Discussion Forum with Interactive Visual Interface to Navigate Search Results. It is a java based web application. Backend has been developed using Spring MVC & hibernate frameworks, Apache Lucene for indexing and MySQL database. User interface has been designed by utilizing HTML5, D3 library, Javascript, jQuery and CSS.

Following sections cover various details like system architecture, database design, data collection procedure for discussion forum and provide details about various algorithms related to information retrieval that have been utilized in designing SearchViz.

A. System Architecture

![System Architecture of SearchViz](image)

SearchViz has been designed using MVC architecture. The advantage of MVC framework is that it differentiates the application logic from user interface design.
Separation of components also provides flexibility in modifying them individually without affecting other areas of the application and providing better code organization.

As shown above, view is comprised of HTML web pages which take user queries and display appropriate results. Controller is a java class which inherits from servlet class that calls appropriate methods of model based on input requests coming through view. Controller class contains a number of methods that call other methods of application logic by analyzing the coming input request. Model consists of various java classes that implement core logic of the application. Below class diagram will provide more details about various business logic of application.

Figure 7: Class Diagram of SearchViz
As you can see in above image; the dispatcher servlet routes user requests from view to the controller class. This class contains various methods that are executed according to user requests. The getresults() method is executed to get results of the entered user query. This method calls other methods that aid in process of result generation.

Recommender systems and search engines try to perform as much computation as possible before serving user requests to generate quick results. Hence the business logic of the application contains different classes that handle involved preprocessing as well as generating results when user requests come in. Preprocessing() method performs the necessary calculations prior to handling incoming user requests. getResults() method takes input queries, calculates appropriate results and forwards them to the view. The data access object (DAO) classes contain methods to input and query data from respective tables in database.

**B. Database Design**

![Database Diagram](image)

Figure 8: Schema design for SearchViz

All the information related to questions and answers are stored in these tables. Question_List table is used to hold all the information regarding different questions. Question_no is the primary key for this table. Answer_List table is used to store all information about answers. Question_no is used as foreign key to map entries of answers to respective questions.
C. Data Collection Procedure

I collected sample data for implementing prototype of online discussion forum was collected from stackoverflow.com website. Stackoverflow is a popular online community with huge active user base containing valuable information about various programming concepts and languages in question-answer form. I used StackExchange API provided by stackoverflow to download questions along with answers related to a few selected topics (loops, class inheritance, variables, exceptions, etc.) related to programming in comma separated file (csv) format. Various data cleaning and transformation operations were applied to the collected data in order to make it suitable for information retrieval.
Chapter 5: User Study Description

A user study was conducted to analyze the effectiveness of network based visualization to navigate search results in online discussion forum. Study was designed to simulate situation of a student using online discussion forum for troubleshooting or problem solving various problems related to computer programming.

Participants were recruited from Arizona State University. They were expected to play role of students capable of using online discussion forums for problem solving tasks related to Java programming and were required to meet following criteria. (1) Majoring in computer science, computer engineering or related field and (2) possessing basic Java programming knowledge. Twenty participants (14 males and 6 females) were recruited and asked to solve two tasks. Each task included two programming questions related to unique concept of Java programming. Figure in next sub-section demonstrates user study procedure.

Students can potentially learn about searching for relevant information by participating in the user study. It is our assumption that users will be able to issue smarter subsequent queries while searching for information. Keyword centered network view of search results will provide enough information and can make it easier to form subsequent queries by incorporating more related keywords while looking for information. Doing the same thing on traditional discussion forums demands comprehensive grasping ability on part of students after reading lengthy textual content to figure out keywords to be incorporated in successive queries. Novice programmers usually need help with forming intelligent queries during initial stages of learning programming and this interface can serve as a learning tool to fulfill that need.
Participants were asked to provide solutions to two tasks by using different versions of online discussion forum as search tool. (1) Version with visual interface to navigate search results and (2) Version of forum with traditional interface. The order of version for each tasks was randomized for participants.

A. Task Description & Procedure

In this section I provide details about programming problems included in two tasks of user study.

The problems tested participant’s knowledge and enabling them to search about concepts like loops, inheritance, constructor and handling exceptions. Each of the problems was assigned difficulty level (E: Easy, M: Moderate, C: Complex)

---

Figure 9: Study Procedure

20
Task - 01: (Time: 20 mins)

1. How can you implement multiple inheritance in Java? Describe an instance where doing so may not be a good option. (Difficulty: M, Concept: inheritance)

2. How can constructor affect class inheritance? (Difficulty: M, Concept: constructor, inheritance)

Task – 02: (Time: 20 mins)

1. Provide few (3-5) examples of problems associated with using iterations. (Difficulty: E, Concept: loops)

2. Provide reasons to create custom exception class. (Difficulty: C, Concept: exceptions)

Difficulty level for a problem was determined based on concepts covered by the problem and number of relevant answers/posts providing information about the problem. Exception problem; which was categorized as difficult has been discussed in only one forum post while the easier problems are discussed in relatively more posts. Each participant solved these problems using assigned versions of the discussion forum. Recording their behavior enabled us to uniquely identify their behavior and track performance for each problem. This provided data to compare participant’s efficiency and performance while using both versions of discussion forum.

B. Conditions for User study

Control group: Use discussion forum with traditional interface

Experiment group: Use discussion forum with interactive visual interface for search-result navigation

I utilized simple repeated measure design to meet requirements of randomization of assigned tasks and conditions.

Participants were informed about the aim and procedure of user study followed by completing background survey form aimed to get information about their
programming experience, sources of learning and troubleshooting for various problems related to programming. 10 minute tutorial on using the visual interface to navigate search results in discussion forums was provided to the participants. This session is very important as it familiarizes participants with interface and also provides learning about interpreting multiple levels of information conveyed by the visualization. While searching on either version of online forum; participants were asked to mark question-answer posts as relevant if it provided them partial or complete information/solution to answer task questions. This exercise was aimed to evaluate search proficiency and provide information about knowledge sources used to provide solutions to task questions.

Finally participants provided subjective evaluations to rate the visual interface on multiple factors like usability, ease of learning, improving search efficiency and satisfaction. They were also interviewed to provide feedback on improving the system and making it more effective and user friendly.

C. Background Analysis of Study Participants

This was aimed to get idea about the educational background of participants along with their programming proficiency. We were also interested in getting deeper insight into how students learn a new programming language and how do they search for problems related to programming. Hence background survey form also had various questions asking them to identify various sources that helped them learn programming and perform various problem solving tasks related to programming. This information is very valuable as we can integrate these sources of learning into the existing system to make the process of learning more effective and simpler for the students.

Charts below provide brief overview about participant’s basic information, programming knowledge and sources of learning. (Note: Number of participants: 20)
As you can see from above charts, we recruited participants with varied programming experience in form of number of programming classes completed and Java proficiency to cover each category. Below chart summarizes various sources of
learning identified by participants. (Note: Participants were allowed to report multiple sources of learning and problem solving)

![Chart displaying sources of learning and problem solving](chart.png)

Figure 12: Chart displaying sources of learning and problem solving

From above statistics we conclude that apart from reference books on various programming languages, various tutorials and online discussion forums are also among major learning sources for programmers. It also adds value to our goal of providing efficient navigation through search results on discussion forums.
Chapter 6: Evaluations

In this section I provide information used in evaluating the effectiveness and usefulness of visualizations in navigating search-results over online discussion forums on basis of user study. I also provide details provided by students about various sources of learning and getting information about performing various problem solving tasks related to programming.

Following measures were used to evaluate the effectiveness and usefulness of the visual interface as compared to traditional list based representation.
1. User performance derived by analyzing statistics and behavior information gathered while using both versions of discussion forum interface
2. Determining efficiency of interface by evaluating user responses to various tasks
3. Analyzing usefulness of interface by considering subjective evaluations provided by study participants

To compute values for above measures, I analyzed subject’s system usage details from system logs and subjective feedback to assess the quality of the visual interface. Results of all those analysis have been discussed in the following sections.

A. System Usage Analysis

In our target context, discussion forums, are considered as traditional information retrieval systems, which typically present search results with one-dimensional ranked lists and put the emphasis on relevant items to be placed in the top of the lists in supporting sequential access. SearchViz implements dynamic interactive visualization to support two-dimensional search results in a network view, which not only encourages users to explore flat and responsive representation on large volume of search-results, but also capitalizes network visualization informatics to present relevant document clusters and their relations. To evaluate how successful
SearchViz can facilitate navigation, I explore two parameters to measure interface effectiveness, Search Performance and Search Efficiency.

1. **Search Performance**

   According to system logs, I found that on an average users viewed 9.15±6.67 number of answers per task in traditional ranked list interface as compared to only 7.40±3.80 in SearchViz. I did not find significant differences in average views between these two groups. However, I observed that users read much more coherently while using visual interface as compared to reading sporadically when using traditional ranked lists. This was an important clue that SearchViz may affect users to effectively find relevant information and resulted in cohesive viewing pattern. On the contrary, in the traditional ranked lists group, user’s viewing patterns were greatly varied, which resulted in possibly some slow and some quick readers or some impatient and some attentive ones.

2. **Search Efficiency**

   Since the search performance included only one parameter (answer views), I considered another parameter (time) to evaluate the effects of search efficiency. I computed the amount of time each user took to identify the first relevant document. I found that users spent significantly more time in visual interface (Mean = 328.61 seconds, Standard Deviation = 239.20 seconds) than in traditional ranked lists interface (Mean = 141.72 seconds, Standard Deviation = 183.80 seconds), $t(17) = 2.2997$, $p=0.022$.

   This result was originally counter-intuitive to my understanding; which I assumed that visual interface would have needed less time. However, the caveat of identifying a relevant document should also take into account the accuracy of whether the marked item was actually relevant. It led me to look at the quality matrices of user’s searches in examining interface impacts.
B. User Performance Analysis

To follow through the clue of visual interface effects, I examined user’s performances by assessing to what extent the system guides users in improving their performances in recognizing relevant posts on discussion forums. I used the following matrices.

*Precision* measures the number of answers that were accurately marked as relevant for a particular task. Precision is defined as fraction of marked answers that are actually relevant, according to equation stated below. The ground truth for calculating answer relevancy was collected by two expert judges manually by thoroughly examining the entire selected corpus (Cohen’s Kappa= 0.589).

To evaluate the ground truth; both experts formed extensive queries for each task problems. For every query, all the 15 search result question posts were reviewed to mark the relevant answers by each expert. Finally ground truth was collected by considering the answers marked relevant by both the experts.

\[
Precision = \frac{|\text{Answers}_{\text{actual relevant}}|}{|\text{Answers}_{\text{marked relevant}}|}
\]

I found that users achieved significantly higher precision in identifying relevant answers for tasks while using visual interface \((M = 0.849, SD = 0.186)\) as compared to traditional ranked lists view \((M = 0.544, SD = 0.388); t(18) = 3.577, p=0.0022\). This demonstrated that users were able to successfully and effectively identify relevant information while using network visual interface to navigate search results.
Analyzing User Logs to justify aid of Visual Interface in Identifying Relevant Posts

In order to get further evidence of usefulness of visual interface in finding relevant answers; I analyzed user logs of participants while using SearchViz to find all instances where they navigated to a question post via visual interface prior to marking that post as relevant. Below image shows a sample log file of a participant displaying such behavior.

```
User Query: multiple inheritance  Mon Oct  5 2015 20:50:08 GMT-0700 (MST)
User clicked on Q1405.txt circle: Mon Oct  5 2015 20:50:14 GMT-0700 (MST)
User clicked on Q65.txt circle: Mon Oct  5 2015 20:50:18 GMT-0700 (MST)
User clicked on Q522.txt circle: Mon Oct  5 2015 20:50:18 GMT-0700 (MST)
User clicked on Q522.txt circle: Mon Oct  5 2015 20:50:20 GMT-0700 (MST)
User clicked on Q522.txt circle: Mon Oct  5 2015 20:50:23 GMT-0700 (MST)
User clicked on Q522.txt circle: Mon Oct  5 2015 20:50:25 GMT-0700 (MST)
User clicked on Q522.txt circle: Mon Oct  5 2015 20:50:27 GMT-0700 (MST)
User clicked on Q1196.txt.circle: Mon Oct  5 2015 20:50:27 GMT-0700 (MST)
User viewed Text-answer(s) of Question: Q1196.txt: Mon Oct  5 2015 20:50:33 GMT-0700 (MST)
User found Text: Q1196.txt-A1196.0.txt to be relevant: Mon Oct  5 2015 20:50:33 GMT-0700 (MST)
User found Text: Q1196.txt-A1196.2.txt to be relevant: Mon Oct  5 2015 20:50:34 GMT-0700 (MST)
User found Text: Q1196.txt-A1196.3.txt to be relevant: Mon Oct  5 2015 20:50:35 GMT-0700 (MST)
User found Text: Q1196.txt-A1196.6.txt to be relevant: Mon Oct  5 2015 20:51:05 GMT-0700 (MST)
User Query: inheritance : Mon Oct  5 2015 20:51:15 GMT-0700 (MST)
User clicked on Q1270 .txt circle: Mon Oct  5 2015 20:51:18 GMT-0700 (MST)
User clicked on Q1405.txt circle: Mon Oct  5 2015 20:51:18 GMT-0700 (MST)
User clicked on Q1405.txt circle: Mon Oct  5 2015 20:51:22 GMT-0700 (MST)
User clicked on Q1405.txt circle: Mon Oct  5 2015 20:51:22 GMT-0700 (MST)
User clicked on Q1417.txt circle: Mon Oct  5 2015 20:51:22 GMT-0700 (MST)
User clicked on Q1417.txt circle: Mon Oct  5 2015 20:51:24 GMT-0700 (MST)
User clicked on Q1417.txt circle: Mon Oct  5 2015 20:51:25 GMT-0700 (MST)
User clicked on Q1387.txt circle: Mon Oct  5 2015 20:51:26 GMT-0700 (MST)
User clicked on Q1278.txt circle: Mon Oct  5 2015 20:51:26 GMT-0700 (MST)
User clicked on Q1316.txt circle: Mon Oct  5 2015 20:51:33 GMT-0700 (MST)
User clicked on Q1316.txt circle: Mon Oct  5 2015 20:51:33 GMT-0700 (MST)
User clicked on Q1316.txt circle: Mon Oct  5 2015 20:51:33 GMT-0700 (MST)
User clicked on Q1316.txt circle: Mon Oct  5 2015 20:51:33 GMT-0700 (MST)
User clicked on Q1196.txt circle: Mon Oct  5 2015 20:52:26 GMT-0700 (MST)
User clicked on Q1294.txt circle: Mon Oct  5 2015 20:52:28 GMT-0700 (MST)
```
Extracting such patterns provides information about sources of navigation to question posts being marked as relevant by the users while using SearchViz. On analyzing the user log files for SearchViz I found that on an average, approximately 65% of the answers marked as relevant had visual interface as source of navigation.

We already saw that users identified relevant information more efficiently while using SearchViz and this pattern extraction confirms visual interface as major source of navigation for identifying that relevant information.

2. Time Analysis per Task

I analyzed time spent by students in searching answers for individual task problems while using both versions of the system. For each task; I analyzed all queries issued by participants to identify what task question were participants working on throughout allocated time. I analyzed time between each subsequent queries to calculate total time spent per question of task. Charts below show results of time analysis.

Figure 15: Average time per problem for Task 01
From the above charts, I observed that no significant difference was noted in average time spent per problem while using either versions of the discussion forum. However we already observed that for exception problem, participants were able to identify relevant posts more precisely while using visual interface as opposed to traditional interface of discussion forum. Also from figure 16, I observed that participants using visual interface to solve exception problem spent slightly more time searching for answers. For same problem, traditional group spending less time to look for answers can be attributed to them being overwhelmed with too much textual information to grasp.

C. Query Quality

In this section I discussed the about impact of discussion forum interface on user’s queries. What were the choices of query terms? Were their query expansions? Were they meaningful? I found that users on average issued 7.35 queries while using visual interface (SearchViz); which was slightly more than that observed in traditional interface(6.40 average). However, it was not a significant difference. There was also not much difference in average amount of words per query while using either interface. Query statistics are summarized in Table-1 below. To
understand the variability of queries, I analyzed them qualitatively. Thus, I further examined the query’s semantics. I measured the concepts encapsulated in each query by overlapping each query with Java Ontology, which was developed and applied in [24, 25] and can be retrieved here⁴.

<table>
<thead>
<tr>
<th>Average</th>
<th>Traditional</th>
<th>Visual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Queries</td>
<td>6.40±4.01</td>
<td>7.35±4.74</td>
</tr>
<tr>
<td>Words per Queries</td>
<td>3.23±0.90</td>
<td>3.33±1.08</td>
</tr>
</tbody>
</table>

*Table 2. Query Statistics Summary.*

I observed that there were patterns of *Query Elaboration and Query Concept Elaboration*. *Query Elaboration* counts the subsequent queries for same question/domain that show an increase in total words included in query but those words not being conceptual words. For example:

q1 = {for loop} and q2 = {how to write for loop}

Above set of subsequent queries display Query Elaboration phenomenon.

*Query Concept Elaboration* considers the subsequent queries for the same question/domain that show an increase in total conceptual words included within them.

For instance: q3 = {problems iteration} and q4 = {problems iteration arraylist}

Thus, q1 and q2 show no conceptual expansion, but q3 and q4 do. Table-2 below demonstrates a concrete example of *Concept Elaboration* for one of the subjects. *Concept Elaboration* query patterns are presented in bold fonts. I found that 40% of subjects who used visual interface (SearchViz) to navigate search results; improved their query quality for successive queries. This behavior was also observed in students using traditional interface, but in only 10% of the cases. The results showed that with the support of network visualization, users were able to

---
⁴Source of Java Ontology: http://www.pitt.edu/~paws//ont/java.owl
expand the queries by adding more conceptual words rather than simply extending length of query string. The outcome can attribute to the succinct network nodes view in the visualization, which displayed each relevant document summarization in one or three keywords. The content summarization provided users to quickly grasp document’s relations with key concepts by recognizing only a few “new” terms, instead of traditionally reading through lengthy texts and mentally forming relations to pick out “new” concepts for next query. I also found evidence that participants heavily interacted with the visualization. Averagely each user made 32.30 clicks on the network visualization nodes corresponding to various questions, keywords and 17.75 unique node clicks. These findings demonstrated that interactive network visualization helped to improve user’s query quality and enhance navigation quality.

| Traditional: | {q1= probl}, {q2= problems associated with iterations}, {q3= iteration + problems =}, {q4 = custom exception class}, {q5= why is it better to create a custom exception class}, {q6 = iterations} |
| Visual: | {q1 = multiple inheritance}, {q2= inheritance + constructor}, {q3= constructor after inheriting a class}, {q4= implement}, {q5= constructor}, {q6= parent class constructor}, {q7= inheritance constructor} |

Table 3. An example of Query Elaboration and Query Concept Elaboration

In above table we can see that the participant did not show any concept elaboration in queries issued while using traditional interface. While in case of queries issued via visual interface, we observe concept elaboration in queries q2, q3 and q5, q6.

D. Analyzing User Responses with respect to Task complexity

Based on the concepts involved for the problems in given tasks, I categorized them according to complexity in three categories i.e. Easy, Moderate and Complex. I
measured completion and accuracy of task responses for all participants, where completion was calculated based on entirety of task-response in covering all aspects of problem and accuracy is the measure of correctness of the response.

I found that users using SearchViz outperformed ones using traditional interface in both parameters accuracy and completion for Exception problem (complex). I did not find such patterns for problems related to easier topics (Figure 11 & 12). This showed that interactive visual interface was especially helpful for complex problems.

![Figure 17: Comparing Accuracy values for various problems categorized as Complex (C), Moderate (M) or Easy (E) attempted by two versions of system. Error bars indicate Standard Deviation.](image)

33
Figure 18: Comparing Completeness values for various problems categorized as Complex (C), Moderate (M) or Easy (E) attempted by two versions of system. Error bars indicate Standard Deviation.

**E. Subjective Evaluation**

This section summarizes the subjective evaluations provided for various criteria by students to visual interface. Participants provided ratings to access usefulness, ease of use and ease of learning of visual interface.

**Accessing Usefulness**

While accessing usefulness; students stated that visual interface conveys multiple levels of information concisely and effectively. It was also helpful in finding relevant information more effectively as compared to traditional list based view.
Interactive visual representation of search results conveys more semantic information as compared to traditional list based view.

**Accessing ease of Use and Learning**

Participants reported that they did not take much time in getting familiar to the user interface and found it relatively easy to learn and understand. They reported that interface helped them in finding relative information easily and effectively.

![Subjective feedback for dynamic forum with visual search-result navigation interface. Error bars indicate Standard Deviation.](image)

Users also provided feedback regarding various components of visual interface by suggesting use of separate shapes to display keywords and question circles to differentiate them more easily.
Chapter 7: Conclusion and Future Work

This is an era of interaction. Newer devices are increasingly making use of touch interface and gestures to get things done. Moreover in case of discussion forums; it is very necessary to present the information in concise and efficient manner due to ever increasing volume of posts and reading efforts needed to find relevant information. In this case, the visual approach proposed in my thesis presents a novel way to deliver required information in concise and user friendly manner. It encourages user interaction, provides flexibility and multiple views of information to help user to focus search and make smart choices.

A. Discussion

In my thesis I identified the need for providing effective ways to find relevant information in fast growing user-generated content of online discussion forums. I proposed a novel interactive visualization interface; capable of providing higher level overview of results along with efficient search-result navigation to address that need.

I also carried out user study to analyze the usefulness and impact of visual interface on search performance and efficiency of participants. I observed that participants were successfully and effectively able to identify relevant information while using visual interface as opposed to traditional interface. I also noticed improved query quality along with higher rates of accuracy and completion of responses while solving complex tasks (Exception problem) using discussion forum with visual interface.

B. Contribution

SearchViz is a novel approach which looks at alternative way of navigating search results on online discussion forums. From the user study results; I conclude that this approach is useful in navigating search results effectively and conveys multiple levels of semantic information more concisely as compared to traditional
ranked list representation in online discussion forums. Background information collected through study participants also provides evidence that online tutorials and discussion forums are newest forms of learning tools that are being utilized in addition to traditional sources like books and other paper based resources.

Students can potentially learn about searching for relevant information by using SearchViz. Keyword centered network view of search results provides enough information to make it easier to form subsequent queries by incorporating more related keywords while looking for information. Novice programmers usually need help with forming intelligent queries during initial stages of learning programming and this visual interface can also serve as a learning tool to fulfill that need.

C. Limitations

There are a few limitations with this approach of navigating search results via visual interface. Current network visualization only considers content semantics ignoring user’s connections, which can be a potential alternative multimodal network visualization in enhancing current work. Another improvement possibility is to enhance network features, such as sort the network as hive plots etc. which will provide more detailed analysis and views.

D. Future Work

As a future work; I will briefly discuss about various improvements that can be made in the existing system as well as implementing this approach in areas other than discussion forums.

1. Improvements to current System

As far as improvements in existing systems are concerned; keyword extraction algorithm can be made more efficient to extract more meaningful keywords that would help users in making smart choices about which document
clusters to examine. User interface can be improved to provide more dynamic components to view the results in various formats according to user choice.

2. Using Interactive Visual Interfaces in other areas

Visual representation approach can be combined with popular search engines like Google, Bing, Yahoo, etc. Search results for various queries often contain results from a diverse range. For example: when users query google for “football score”; they may be referring to either what Europeans call soccer or American game of football. So search results referring to this query can be divided in two clusters (assumption). It is not a good idea to predict user’s mind and show results from either cluster only. Ideally results from each cluster should be presented in efficient manner with relevant information to make user’s choice easier.

The visual representation approach can be utilized in this case by search engines to present their results in efficient manner and letting users make smart choices. Thus; I believe that this approach will greatly help users and impact the field of search engines.

In spite of limitations, study analysis and subjective feedback led me to believe that network based visualization interface is effective in navigating search results in online discussion forums by eliminating efforts for extensive reading and grasping capabilities to find relevant information. This novel approach of improving search precision and efficiency in online discussion forums by integrating visual interface with traditional list based result-view should be explored further.
REFERENCES


