SUPPORTING INFORMATION SEEKING AND SENSEMAKING
IN ISSUE-BASED KNOWLEDGE CRYSTALLIZATION

A Dissertation in
Information Sciences and Technology
by
Feng Sun

© 2018 Feng Sun

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2018
The dissertation of Feng Sun was reviewed and approved* by the following:

Guoray Cai  
Associate Professor of Information Sciences and Technology  
Dissertation Advisor, Chair of Committee

Mary Beth Rosson  
Professor and Associate Dean of Information Sciences and Technology  
Interim Graduate Director

Xiaolong (Luke) Zhang  
Associate Professor of Information Sciences and Technology

Alan M. MacEachren  
Professor of Geography

*Signatures are on file in the Graduate School.
Abstract

People often face the challenge of having to sift through a large volume of data and make sense of them in order to achieve actionable insights on complex issues. We formalize this challenge as a class of problem, *Issue-based Knowledge Crystallization (IBKC)*, and address this problem in the context of local civic engagement.

This dissertation focuses on understanding the information seeking behavior of IBKC and identifying design issues for supporting effective knowledge crystallization. Towards these goals, we propose a conceptual framework that characterizes the information seeking processes and associated information and cognitive barriers in IBKC. It was informed by prior theories and models, and by observing practices of IBKC in Community Issue Review (CIR). We found that effectively extracting information nuggets in IBKC requires collaborative information foraging and sense-making where workspace awareness and activity awareness are important for judging the completeness of nugget extraction task. Based on such insight, we employed a design-based research approach to investigate the design issues of NuggetLens as a visual analytical method to support group information seeking in IBKC.

The design of NuggetLens requires that we advance our understanding of two questions: (1) how do people judge completeness in nugget extraction tasks? (2) how should people be assisted in dealing with cold-start problems when social clues are lacking? The first question was answered by observing small-group nugget extraction experiments and semi-structured interviews with participants. The second question was partially answered by generating an initial information landscape through an interactive topic modeling method. Using a variety of datasets, we demonstrated that interactive topic modeling can enable non-expert users to quickly make sense of topics and generate good summaries. The resulting topics, serving as information scent, can help users better understand document space structurally, providing a solid foundation for nugget extraction.

This work contributes to the science and solutions of information seeking by presenting new research frontiers in IBKC problems. Our findings have significant implications in enabling informed citizens engagement in public decision-making.
Table of Contents

List of Figures viii

List of Tables x

Acknowledgments xi

Chapter 1
Introduction 1
1.1 Motivation 2
1.2 Issue-based Knowledge Crystallization (IBKC) 4
1.3 Challenges for Supporting Information Seeking in IBKC 5
1.4 Research Scope 8
1.5 Research Objectives and a Design-based Research Approach 9
1.6 Contributions 13
1.7 Outline 14

Chapter 2
Understanding Information Seeking in IBKC 16
2.1 Theories, Concepts, and Models of Information Seeking 16
2.1.1 Information Seeking 17
2.1.2 Collaborative Information Seeking 24
2.1.3 Information Sensemaking 32
2.1.4 Information Foraging 34
2.1.5 Information Need 36
2.1.6 Summary 38
2.2 Background on Knowledge Crystallization 39
2.3 A Conceptual Framework of IBKC 42
2.3.1 Problem Context: Community-level Issue Review 44
2.3.2 Goal: Complete and Concrete 45
2.3.3 Data: Large, Unstructured, and Heterogeneous 46
4.2.5.4 Representative Decision Points .......................... 93
4.3 User Study ....................................................... 94
  4.3.1 Study Setup ................................................. 94
  4.3.2 Procedure .................................................. 95
  4.3.3 Data Analysis ............................................... 95
  4.3.4 Results ..................................................... 96
4.4 Summary and Discussion ....................................... 100

Chapter 5
Designing Interactive Topic Modeling Interface for Knowing about an Information Landscape 101

5.1 Related Work .................................................... 102
  5.1.1 Characterizing Petition Data and Information Needs .......... 102
  5.1.2 Information Visualization Approaches for Sensemaking ...... 105
  5.1.3 Interactive Topic Modeling .................................. 106
    5.1.3.1 Topic Modeling Techniques ............................. 106
    5.1.3.2 Human-in-the-loop Topic Modeling ...................... 108
    5.1.3.3 Disconnection Between Implementation and User Perception ......................................................... 111
  5.1.4 Interactive Topic Modeling for Non-expert Users ............ 111
5.2 Design Considerations ........................................... 115
5.3 FILA: System Implementation ..................................... 119
  5.3.1 Data Preprocessing .......................................... 119
  5.3.2 Interface Design ............................................ 120
  5.3.3 Initial Topic Discovery .......................... 123
  5.3.4 Supported Operations ...................................... 123
    5.3.4.1 Split a Topic ........................................ 124
    5.3.4.2 Merge Topics ......................................... 125
    5.3.4.3 Move Topic Words .................................... 125
    5.3.4.4 Set the Number of Topics ............................ 125
    5.3.4.5 Navigate through Topic Models ......................... 126
  5.3.5 Usage Scenarios of Topic Refinement ...................... 126
    5.3.5.1 Move Intruder Words .................................. 126
    5.3.5.2 Split a Multi-theme Topic ............................ 129
    5.3.5.3 Merge Semantically Similar Topics .................... 132
  5.3.6 Interacting with FILA ...................................... 133
    5.3.6.1 Inspect a Single Topic ................................. 135
    5.3.6.2 Visually Investigate Clusters of Documents ............ 135
    5.3.6.3 Drill-Down in Context of Topics ....................... 136
5.4 User Study ....................................................... 136
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4.1 Methods</td>
<td>137</td>
</tr>
<tr>
<td>5.4.1.1 Datasets</td>
<td>137</td>
</tr>
<tr>
<td>5.4.1.2 Participants</td>
<td>138</td>
</tr>
<tr>
<td>5.4.1.3 Scenario Description and Task Design</td>
<td>139</td>
</tr>
<tr>
<td>5.4.1.4 Procedure</td>
<td>140</td>
</tr>
<tr>
<td>5.4.1.5 Data Analysis</td>
<td>142</td>
</tr>
<tr>
<td>5.4.1.6 Scoring Performance</td>
<td>143</td>
</tr>
<tr>
<td>5.4.2 Results</td>
<td>143</td>
</tr>
<tr>
<td>5.4.2.1 Performance Results</td>
<td>143</td>
</tr>
<tr>
<td>5.4.2.2 Behavioral Patterns and Strategies</td>
<td>144</td>
</tr>
<tr>
<td>5.4.2.3 Subjective Feedbacks</td>
<td>146</td>
</tr>
<tr>
<td>5.4.3 Discussion</td>
<td>147</td>
</tr>
<tr>
<td>5.4.3.1 Topic Quality Improvement</td>
<td>147</td>
</tr>
<tr>
<td>5.4.3.2 Topic Refinements</td>
<td>148</td>
</tr>
<tr>
<td>5.4.3.3 Document Collection Visualization</td>
<td>149</td>
</tr>
<tr>
<td>5.4.3.4 Limitations</td>
<td>150</td>
</tr>
<tr>
<td>5.5 Summary and Discussion</td>
<td>147</td>
</tr>
</tbody>
</table>

Chapter 6

Conclusions

6.1 Research Contributions                                             153
6.2 Limitations and Future Work                                       155

Bibliography                                                          157

Appendix A: The Questionnaire for Evaluating the Design of TopicLens    180

Appendix B: The Task Description and Instructions, and the Questionnaire for Assessing the Design of QILA 187
List of Figures

1.1 The overview of community issue review as a knowledge crystallization implementation ........................................ 5

2.1 Wilson’s nested model of information behavior (adapted from [Wilson, 1999]) .................................................. 19
2.2 Wilson’s 1981 model of information behavior (Adapted from [Wilson, 1981]) .................................................. 21
2.3 Krikelas’s information seeking model (adapted from [Krikelas, 1983]) .............................................................. 22
2.4 Kulthau’s information search process (ISP) modal (adapted from [Kuhlthau, 1991]) .............................................. 23
2.5 A comparison of Ellis’s behavioral framework and Kulthau’s information search process (ISP) (adapted from [Wilson, 1999]) .................................................. 24
2.6 Foster’s non-linear models of information seeking behavior (adapted from [Foster, 2004]) ....................................... 25
2.7 Collaborative information seeking (CIS)’s relationship to the information seeking, collaboration, and information retrieval fields, in context. (adapted from [Shah et al., 2014]) .............................................. 26
2.8 Karunakaran, Reddy and Spence’s model of collaborative information behavior in organizations. (adapted from [Karunakaran et al., 2013]) .............................................................. 27
2.9 Pirolli and Card’s sensemaking process for analyst (adapted from [Pirolli and Card, 2005]) ........................................... 33
2.10 Cole’s theory of information need (adapted from [Cole, 2011]) ................................................................. 37
2.11 Knowledge crystallization (adapted from [Card et al., 1999]) ................................................................. 39
2.12 The sensemaking framework for supporting Knowledge Crystallization ................................................................. 43
2.13 The overview of community issue review as a knowledge crystallization implementation ........................................ 51
2.14 Face-to-face meeting on Day 1 .................................................. 52

3.1 Nugget extraction workspace 1.0 .................................................. 66
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>The nugget extraction task structure.</td>
<td>69</td>
</tr>
<tr>
<td>4.1</td>
<td>User interface with NuggetLens enabled (used in the 33-participant</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>user study)</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>DocuMap</td>
<td>86</td>
</tr>
<tr>
<td>4.3</td>
<td>Parallel Sets plot</td>
<td>87</td>
</tr>
<tr>
<td>4.4</td>
<td>Zoomable bar chart</td>
<td>88</td>
</tr>
<tr>
<td>4.5</td>
<td>Extracting an information nugget.</td>
<td>91</td>
</tr>
<tr>
<td>4.6</td>
<td>The screenshots of NuggetLens at representative decision points</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>(four scenarios of use)</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>NuggetLens user interaction log</td>
<td>97</td>
</tr>
<tr>
<td>4.8</td>
<td>Results of questionnaire (the x-axis represents the elicited four</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>evaluation criteria, and the y-axis represents the degree to which</td>
<td></td>
</tr>
<tr>
<td></td>
<td>participants consider the corresponding criteria are strong)</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Topic analysis using topic coherence. The context vector-based</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>topic coherence of fitting the online petition data is shown for CorEx</td>
<td></td>
</tr>
<tr>
<td></td>
<td>models, where the number of topics varies from 10 to 185</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>The user interface for exploring a large document collection based</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>on interactive topic modeling</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Move topic word “health” from topic 4 to topic 0</td>
<td>128</td>
</tr>
<tr>
<td>5.4</td>
<td>A comparison of visualized petitions before and after moving words</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>between topic 0 and topic 4</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>Split topic 6 into two topics topic 6 (6.1) and topic 7 (6.2)</td>
<td>130</td>
</tr>
<tr>
<td>5.6</td>
<td>A comparison of visualized petitions before and after splitting topic</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>6 130</td>
<td></td>
</tr>
<tr>
<td>5.7</td>
<td>Merging topic 0 and topic 9</td>
<td>133</td>
</tr>
<tr>
<td>5.8</td>
<td>A comparison of visualized petitions before and after merging topic</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>6.1</td>
<td>Positioning the major contributions of this dissertation in IBKC</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>framework.</td>
<td></td>
</tr>
</tbody>
</table>
List of Tables

2.1 CIS research summarization (adapted from [Foster, 2007]). . . . . . . 28
2.2 Differences between IIB and CIB across different items (adapted from [Reddy and Jansen, 2008]). . . . . . . . . . . . . . . . . . . . . . . . . . . 29
2.3 Differences between IIB and CIB across different levels (adapted from [Reddy and Jansen, 2008]). . . . . . . . . . . . . . . . . . . . . . . . . . . 29

5.1 Selected topics (#topics = 20) . . . . . . . . . . . . . . . . . . . . . 127
5.2 Datasets Description . . . . . . . . . . . . . . . . . . . . . . . . . 138
5.3 Performance Comparison . . . . . . . . . . . . . . . . . . . . . . . 144
5.4 Time Spent Comparison . . . . . . . . . . . . . . . . . . . . . . . . 144
5.5 Subjective Ratings . . . . . . . . . . . . . . . . . . . . . . . . . . . 146
Acknowledgments

There are no proper words to convey my deep gratitude and respect for my thesis and research advisor, Professor Guoray Cai. I could not have accomplished all these things without his continuous support, patience, and critical thinking. He inspired me to become an independent researcher.

My earnest thanks to my committee members: Dr. Mary Beth Rosson, Dr. Alan M. MacEachren, and Dr. Xiaolong (Luke) Zhang. They provided tremendous help in various aspects of my research work.

I deeply thank my parents, Jun Sun and Fuying Chen for their unconditional trust, timely encouragement, and endless love despite the long distance between us.

Finally, I would like to thank my dear wife, Xiao Liu, and our lovely daughter, Skye, for their endurance, support, and love.
this study focuses on designing technological support for citizens to better seek information and extract important information in local civic engagement.

1.1 Motivation

Meaningful participation in public decision-making premised on the ability of the participants to form reasonable and well-informed opinions [Chambers, 2003, Dewey and Rogers, 2012]. It is important for the success of democratic decision-making processes [Gastil and Levine, 2005]. However, such ability is often hindered by a lack of exposure to a diverse marketplace of ideas [Druckman et al., 2012, Price and Neijens, 1997].

In addition, it is widely known in the literature of deliberative democracy [Gudowsky and Bechtold, 2013] and policy communication studies [Richards et al., 2013] that the public is facing a tremendous information barrier to public participation. This barrier is a result of both the large amount of heterogeneous knowledge that must be made explicit in different formats at different stages (information overload) [Elliman et al., 2007] and the insufficient political knowledge most Americans have (civic knowledge deficit) [Burkhalter et al., 2002, Carpini and Keeter, 1997, Jacobs et al., 2009].

In order to better participate, participants have to deal with the daunting task of seeking, interpreting, and synthesizing a large number of documents, comments, and claims, along with other materials to gain insight on which opinions are based [Gross, 1964, Toffler, 1990]. Such in-depth civic participations usually require much effort but citizens who are not enthusiastic participants can only allocate limited time and energy.

Recently, deliberative innovations have been developed, such as mini-publics [Goodin and Dryzek, 2006] and random assembly [Gastil and Richards, 2013], which can delegate in-depth and information-intensive public participation tasks to a small
group of citizens who will engage in detailed investigation, analysis, and deliberation on an issue and report their findings back to the broader public. Mini-publics are very suitable for this purpose because they attempt to achieve impartiality, preference change, and public reasoning by bringing together a random selection of non-partisan citizens to discuss key issues [Elstub and McIaverty, 2014]. Mini-publics are made up of ordinary, non-partisan, lay citizens and are “designed to be groups small enough to be genuinely deliberative and representative enough to be genuinely democratic” [Goodin and Dryzek, 2006]. In mini-publics, either a random or a stratified sample of the population is selected to achieve a microcosm of the population, with each citizen having an equal chance of being selected [La-font, 2015]. They are usually asked to deliberate on the issue given evidence and advocacy provided by subject matter experts [Elstub and McIaverty, 2014]. In the past, mini-publics were mostly made use of for the purpose of writing policy recommendations to decision-makers, rather than crystallizing knowledge for the public to be informed.

The Oregon practice of mini-publics [Gastil and Knobloch, 2010], named Oregon Citizens Initiative Review (OCIR), is an exception. OCIR randomly selects a panel of voters to evaluate a ballot measure through reviewing the information gained from campaigns and policy experts. The selected small panel gather what they know about a public issue and, based on all the data that can be found, present such knowledge in a concise fashion [Gastil and Richards, 2013]. The final outcome is a set of statements conveying the most important findings of the measure, which serves as an easily accessible resource for informing voters. OCIR was designed to be an institutional tool for the state-level vote on ballot measures, and the practice of OCIR typically involves a face-to-face meeting that lasts 4 or 5 days, which may exclude a certain population to participate [Gastil et al., 2014, Kropczynski et al., 2015b]. In other words, the practice of deep deliberation is still time-, effort-, and attention-consuming for the selected small panel.
1.2 Issue-based Knowledge Crystallization (IBKC)

OCIR uses mini-publics to analyze issue-based community knowledge, discover useful knowledge from the information, and contextualize it to inform the broader public [Rowley, 2007]. This practice is formalized as a general class of problems called Issue-Based Knowledge Crystallization (IBKC) and maps the challenges of supporting OCIR into a series of scientific questions where rigorous research can be developed by studying the Community Issue Review (CIR).

IBKC is a panel-based deliberation process and attempts to reduce the participation barrier by crystallizing knowledge about a pending government or community issue. Data about an issue are usually scattered in multiple venues and media forms and are not directly usable by the broad public in a community (due to being redundant, unsystematic, conflicting, ill-structured, unreliable, and/or incomplete). IBKC takes advantage of a small group of citizens to analyze the issue in-depth through a multi-day public review process. Panelists are either randomly or strategically selected from citizens, depending on the nature of the issue. As representatives of a community, panelists are granted access to a large amount of data from various sources concerning a given issue. IBKC aims to generate an informative briefing of an issue to provide the broader public with insights concerning the issue so that every citizen is able to form an opinion effectively and efficiently.

The procedure illustration of IBKC is provided in Figure 1.1. Over the course of practicing IBKC, the panel of panelists needs to understand an issue and related materials, seek information from the available data, refine collected information and make claims, and present the outcome to the public. Crystallizing knowledge from issue knowledge involves a set of information behaviors, including identifying and refining information needs, making sense of information landscape, seeking and collecting information nuggets, and using the information to inform more people.
IBKC can be treated as an information seeking task.

In this iterative divergent-convergent process of IBKC, the finest granularity of information is information “nuggets”. Compatible with existing literature in information sciences [Goecks and Cosley, 2002, Imran et al., 2013, Pousman and Stasko, 2006], information nuggets is defined as brief, discrete, and self-contained information pieces in documents that are pertinent to describing a given community issue. In our case, claims regarding the community issue can benefit from aggravating and assembling information nuggets. Examples of information nuggets include phrases, sentences, even paragraphs, as long as they are self-contained and are useful for making claims, they are considered as information nuggets.

1.3 Challenges for Supporting Information Seeking in IBKC

Despite all advantages, IBKC is still an unexplored instance of information seeking with special constraints imposed by the context of applying mini-publics to reviewing community issues. Compared with traditional information seeking, IBKC is usually collaborative and requires multiple participants to work together; the goal of IBKC is to produce a compact and accurate description of knowledge, rather than retrieving most relevant information; and IBKC involves many information behaviors other
than searching.

There is a knowledge gap between the understandings of well-studied traditional information seeking processes and not fully elucidated information seeking in the context of IBKC. Although much work has been done, both in fields of information seeking (theories, models, concepts, and relevant information behaviors) and digital government (methods and infrastructures for collecting and sharing ideas, galvanizing interest and eliciting participation) [Hartz-Karp and Sullivan, 2014], a conceptual and systematic description of IBKC is still missing, which prevents us from examining information foraging behaviors in the context of IBKC and designing tools to support them.

This section describes the challenges of supporting IBKC from the aforementioned dimensions: it mandates collaboration among participants, brings unique requirements of completeness, requires in-depth understanding beyond berrypicking (a model of online searching behavior [Bates, 1989]), and involves many information behaviors other than searching.

In terms of its collaborative nature, IBKC requires a strategically selected citizen panel to collect all the available data about a public issue and deliver the compacted description of the information to the general public [Gastil and Richards, 2013]. During the process, the group of people needs to work together through communication, coordination, and cooperation in order to efficiently generate meaningful collective intelligence. Collaborative information seeking (CIS) has been studied in various settings [Jensen, 2009, Morris and Horvitz, 2007, Paul and Morris, 2009b] other than community-level civic engagement.

IBKC can benefit from collaboration, but little attention has been paid to collaborative information seeking in knowledge crystallization. In particular, maximizing the benefits of collaboration while minimizing the costs of information seeking in IBKC is challenging as collaboration can bring benefits to information seeking while introducing additional costs. Researchers are still investigating the
influence collaboration has on information seeking [Hyldegård et al., 2015, Prekop, 2002, Shapira et al., 2001, Tao and Tombros, 2014a]. The deliberative and democratic nature of IBKC makes it even more complicated. Systems for supporting CIS activities are usually designed for a specific application scenario (e.g., web search [Morris and Horvitz, 2007], health care [Reddy and Jansen, 2008], crisis management [Bjurling and Hansen, 2010], patent processing [Hansen and Järvelin, 2005], library [Twidale et al., 1997], academy [Sonnenwald et al., 2004], software design [Poltrock et al., 2003]) since the context plays a significant role in shaping the requirements of system design. Although these tools cannot be directly applied to IBKC, possible design problems can be inferred, providing valuable insights for supporting IBKC [Hearst, 2014].

In addition, IBKC brings the additional requirement of ensuring a comprehensive information coverage (i.e., completeness) in collaborative information seeking. Compared with some settings whose information space appears to be infinite (e.g., library database [Brown and Ortega, 2005], the internet [Morris and Horvitz, 2007]), participants in IBKC are expected to go through and make sense of a finite collection of data by collaboration.

Previous research tended to investigate information searching tasks [González-Ibáñez et al., 2013, Kuhlthau, 1991, Morris, 2008], with the goals of providing a thorough description and explanation of the process in terms of user behavior and strategies [Brown and Ortega, 2005, Davies, 2007, Robinson, 2010, Tao and Tombros, 2014b], and designing tools to support the process [Krishnappa, 2005, Morris and Horvitz, 2007, Shah, 2010b]. The unique evaluation requirement of IBKC makes information searching behavior (which is the design focus of many existing tools) less important; instead, the ability to constantly assess the completeness of information with regard to the given issue becomes vital [Ma et al., 2017].

Furthermore, judging the relevance of data in information seeking in IBKC is a human-directed activity. The complexity of the available data requires individuals
to have an in-depth understanding and interpretation of the data, which cannot be achieved simply with computational power [Hagen et al., 2015b]. Comparatively, judging the relevance of content in some settings is a relatively simple action to take. Although technology can not replace human beings in terms of judging relevance, it can provide support for facilitating the process.

This is especially the case when the dataset is large. There are many text analysis and visualization technologies and tools currently available [Atterer and Lorenzi, 2008, Cutting et al., 2017, Dou et al., 2013, Gretarsson et al., 2012, Koch et al., 2014] to enable human beings to read, make sense of, and summarize large data collections. However, whether they are suitable for this particular context needs further study. Not only the selection of technologies but also how the selected technology is tailored to meet the needs of sensemaking in IKBC should be considered carefully. Ideally, the adopted technology should be well-performing, transparent, easy-to-understand for non-expert users, and with iterative user interaction [Lee et al., 2017].

Although considerable effort has been made, we are still far from understanding information seeking in the community decision making context. All of these problems increase the complexity of designing systems to support information seeking in IBKC. In summary, we lack a thorough understanding of IBKC and this prevents researchers from practically designing and developing tools to support the additional requirements imposed by the unique characteristics of IBKC.

1.4 Research Scope

Information seeking is a broad term encompassing the ways individuals identify information needs and, seek, assess, select, collect and use the needed information [Majid et al., 2000]. IBKC, in which information seeking is widely observed, is a very complex process that includes several phases, a variety of information behaviors, information intermediates, and all kinds of interactions between these
entities. It falls within the scope of collaborative information seeking with specific requirements introduced in a local CIR context.

Based on the challenges of supporting information seeking in IBKC identified in the previous section, there are multiple directions we can take to start this work. Among them, the collaborative nugget extraction phase of IBKC is the focus of this dissertation, and it includes conceptualizing nugget extraction in the context of IBKC and identifying possible challenges and difficulties of existing practices so as to help users better forage information during nugget extraction phase from the perspective of system design. I designed and implemented tools to support collaborative nugget extraction and they serve as a vehicle to develop a qualitative understanding of the phenomenon in IBKC. The above-mentioned characteristics make it difficult to directly apply existing solutions of information seeking to IBKC context.

As the first phase of IBKC, and the focus of this dissertation, nugget extraction is itself an information seeking task. It involves activities of navigating through a collection of documents, making sense of the documents, recognizing relevant contents, and extracting information nuggets from the documents. Nugget extraction is a challenging task because it is a collaborative process that requires constant evaluations of situation, activities, and information landscape. Also, the topics discussed and extracted are community issues that are difficult to reach consensus.

1.5 Research Objectives and a Design-based Research Approach

The essential research questions of this work are, “What is special about IBKC as a kind of information seeking? What role can technology play in supporting nugget extraction.” This is an exploratory question and addressing such a question requires both the design and theory of supporting nugget extraction in IBKC to be
mutually developed through the research process. The goals of this work include not only designing solutions for supporting nugget extraction but also refining design principles. Therefore, a design-based research (DBR) approach is adopted to elicit design issues of supporting the process. Wang and Hannafin [Wang and Hannafin, 2005] define DBR as:

“a systematic but flexible methodology aimed to improve educational practices through iterative analysis, design, development, and implementation, based on collaboration among researchers and practitioners in real-world settings, and leading to contextually-sensitive design principles and theories.”

DBR is appropriate for situations where research is conducted on practical problems and theories are not well-established for directly guiding problem solving, rather, iterative processes of testing and evaluation of solutions are necessary to deliver better design principles. As such, DBR provided a process that is well-suited for developing solutions to support information seeking behaviors in local civic engagement process. The design process using DBR approach is grounded in both existing theories and practices of information seeking in IBKC context. It is interactive, iterative and flexible, implying that the interventions (developing solutions) tend to be refined continuously. Also, the design process needs to be contextualized for findings to adapt to other settings.

These essential research questions are further decomposed into three sub-questions that correspond to the challenges described in Section 3.2. These questions are described as follows along with objects and corresponding DBR approaches:

1. Conceptualize IBKC and identify the characteristics of IBKC. Several research questions need to be addressed are: What are the general characteristics of IBKC as an information seeking instance? What are the special characteristics of IBKC that make it different from commonly studied information searching
tasks? What are the difficulties and challenges of implementing and practicing IBKC?

This work starts with a conceptualized framework to describe IBKC from multiple directions. On the one hand, theories and models of information seeking are applied to describe information phenomena in IBKC. On the other hand, the unique characteristics and context of IBKC that make it different from commonly studied information seeking tasks are highlighted and elaborated on. Specifically, IBKC is conceptualized by characterizing the entities involved in the process, the linkages among these entities (how information behaviors transform data), contexts, and the evolutionary structure of the process. By identifying its unique requirement of completeness, the information behavior of browsing, and its collaborative nature, a solid foundation is laid for designing tools to support these special needs.

2. Identify the design issues for supporting collaborative nugget extraction in IBKC. Related research questions are the following: What are the challenges and difficulties information foragers may encounter during collaborative nugget extraction? What are the design characteristics of a practical and effective tool that enables participants to be aware of ongoing activities and products in nugget extraction?

First of all, nugget extraction was isolated from other phases and was analyzed as a self-contained process based on previous literature. Then a formative study (described in Section 3.3) was conducted and several information scents users use to evaluate nugget relevance, task process, and potential opportunities during nugget extraction were identified. In particular, in order to promote collaborators’ awareness of activity in terms of both processes and products, I designed and implemented a visual analytical tool called NuggetLens. NuggetLens is composed of multiple coordinated views that
provide a depiction of collaborative activities and products in real time. A user study was conducted to demonstrate the usefulness of NuggetLens, while it also helped find some problems that challenge its effectiveness. I further conducted research to address one of these issues regarding the induction of initial schemas that can prepare participants for nugget extraction by providing a structural summary of the information landscape.

3. Explore an approach that enables information foragers to have a structural understanding of an initial information landscape for better nugget extraction. Related research questions are the following: What are the difficulties for ordinary people to quickly and effectively make sense of an information landscape? Which techniques have the potential of tackling those difficulties? What is the disconnection between user perceptions and current implementations? And in particular, how can we help information foragers to quickly know about an information landscape for nugget extraction in IBKC by joining the forces of computational power and human knowledge?

Inspired by a problem identified in investigating the nugget extraction process, I found there is a need for information foragers to quickly and effectively have an understanding of a given information space. I analyzed existing technologies for large document sensemaking and found that the interactive topic modeling technique has the greatest potential. Based on existing works of interactive topic modeling, I identified the disconnections between user perceptions and current implementations. These disconnections encouraged me to develop an interface called Fast Information Landscape Analyzer (FILA) for non-experts users to quickly generate a structural understanding of the information landscape of a given document collection. A user study was conducted to demonstrate the usefulness of FILA and to help elicit design issues for supporting the sensemaking of an information landscape.
1.6 Contributions

In this dissertation, I claim three major contributions.

First, I propose a new framework for conceptualizing information behavior in IBKC. This framework was applied to a more concrete implementation, i.e., CIR. I designed an online system to support this implementation and conducted user studies with recruited citizens. The practice demonstrated the effectiveness of this process, highlighted multiple features this process brings and identified some drawbacks as well. Although this work is specifically conducted within our community, the idea of taking advantage of a “mini-public” to inform citizens at the community-level can be easily generalized to a variety of applications and scenarios.

Second, a visual dashboard that allows users to learn about collaborative information seeking behavior, NuggetLens, was developed and tested. The preliminary research conducted for understanding collaborative information seeking in IBKC is reported. Specifically, I focused on the information extraction phase using CIR as an example of collaborative information seeking to investigate how such a task can be supported. This includes a preliminary study that helps elicit design requirements and a user study that helps evaluate the usefulness of NuggetLens. This process helps us better understanding collaborative information seeking in CIR and how technology can play a role.

Third, I designed and implemented an interactive topic modeling tool to provide a structural summary of a given document collection that can provide nugget extraction with an initial schema. Compared with existing approaches, this tool is designed especially for non-expert users by providing straightforward interactions and avoiding direct parameter manipulation. In addition to topic modeling techniques, the introduction of visualization enables users to make sense of the information space more intuitively. The back-and-forth capability allows users to iteratively refine topic models and explore documents with more flexibility. In short,
this tool has the potential to bridge the disconnection between user perceptions and technical implementation details, which can also benefit other domains in addition to supporting citizens’ sense-making of large documents.

My research provides new insights in technological support of information seeking and sensemaking in the public participation domain where information, technology, and people intersect. In order to tackle information seeking and sensemaking challenges in civic engagement, technologies were introduced to provide assistance while people who are not technology experts can still steer technology and make decisions.

1.7 Outline

The dissertation is structured following the DBR process [Wang and Hannafin, 2005]: it starts with an analysis of CIR practices with a theoretical framework, and it leads to the development of a visual analytical solution. Evaluation and testing of the solution provide several insights while also reveal certain limitation, which results in another round of refining problems, solutions, and methods. Several design principles, as well as lessons learned during the process, are produced in the end.

In more details, this dissertation essentially investigates the following questions: “What is special about IBKC as a kind of information seeking, and how to employ technologies to support it in terms of nugget extraction?” Specifically, this work explores ways to help citizens be better involved in community-level public participation in terms of seeking information about community issues. It starts with a conceptual formalization of IBKC. A formative study with an implementation of IBKC, CIR, is then conducted, allowing us to have a better understanding of this process and find the difficulties and challenges it faces. In particular, I designed and implemented visual analytical support for the first phase of CIR, collaborative
nugget extraction. I achieved promising results while also encountering challenges with the initial representation of the information landscape. This led me to the exploration of an interactive topic modeling technique as an approach for inducing an initial schema for cold-starting of nugget extraction.

As outlined in the above rationale, the rest of this dissertation is organized as follows. Chapter 2 will introduce the framework IBKC and one of its concrete implementation, CIR, serving as the context of this study. An illustrative study of CIR and a formative study that concentrates on the nugget extraction activity of CIR are described in Chapter 3, and they lead to several implications for tool design. Based on those implications, Chapter 4 focuses on the support of information seeking and sensemaking in nugget extraction - multiple coordinated views of collaborative information behavior in supporting the evaluation of the completeness of extracting information nuggets. Chapter 5 focuses on the beginning of nugget extraction when social scents are lacking: an interactive topic modeling approach is proposed in supporting better sensemaking and summarizing the information landscape. Chapter 6 will discuss and conclude my work.
Chapter 2  
Understanding Information Seeking in IBKC

This chapter takes the first step by providing a conceptual framework of IBKC based on what we know from the literature. More specifically, I frame IBKC as an information seeking instance and review pieces of literature on information seeking, collaborative information seeking, sensemaking, information foraging and information needs. I also review the concept of knowledge crystallization used in information visualization field. Besides the reviewed literature, I highlight the unique dimensions of IBKC, along with the descriptions obtained from the literature, providing a systematic and comprehensive description of IBKC that reflects its characteristics.

2.1 Theories, Concepts, and Models of Information Seeking

This section provides an overview of the historical background of the work on information behavior in terms of concepts, theories, and models, including the models, theories, and concepts that can contribute to the conceptualization of IBKC.

Information seeking is an important concept that involves information foraging
theory and sensemaking is an important aspect of information seeking processes. Information foraging theory is the study of how human users seek, gather, and consume information in an “information environment [Pirolli and Card, 1999]”. While information seeking is theorized and applied differently in a variety of contexts and in different domains, the results, and findings from this research can provide useful references and valuable insights, serving as the basis of this study.

2.1.1 Information Seeking

IBKC falls into the category of information seeking. Therefore, much of our understanding of information seeking, either in a universal sense or in a concrete context, should be still valid for IBKC. This section selectively introduces the current popular theory and concepts of information seeking that can be used to help describe IBKC. These features and properties apply to most information seeking, including IBKC. Before introducing the uniqueness brought by IBKC, reviewing these studies can help us understand the nature and characteristics of IBKC as a kind of information seeking. Although these characteristics are not unique, they are necessary for conceptualizing IBKC.

Researchers have been studying information seeking for decades, and many theories and models have been developed to describe information seeking. This study positions sensemaking and information seeking in a collaborative and deliberative environment. Therefore, a good understanding of information seeking can provide a solid foundation to examine what parts of information seeking can benefit from introducing collaboration and studying sensemaking.

Information seeking models provide frameworks for understanding information seeking and related behavior by representing and organizing involved processes and their relationships from various aspects. Most of the models are pragmatic and descriptive, and they vary in terms of their assumptions, structure, purposes, and scope [Case, 2012]. One way that differentiates models is whether a model focuses on
active or passive information seeking. Models of passive information seeking apply to non-work settings (e.g., everyday life) [McKenzie, 2003] while active information seeking occurs mostly in work settings (e.g., scholars or professionals). Another way that divides models by their structures. Many models adopt flowcharts [Krikelas, 1983, Wilson, 1981] to describe the sequence and relationships of activities involved in information seeking. There also exists models that are not linear [Foster, 2004]. In addition, depending on different levels of specification, some of the models focus on more specific and fine-grained problems than others [Wilson, 1981, 1999]. These variances usually result from different research purposes, e.g., cognitive and affective factors [Kuhlthau, 1991], contextual factors [Foster, 2004], and personal variables [Longo, 2005]. Although these models are difficult to compare directly, they all contribute to the understanding of information behavior and sometimes can be used together to interpret information seeking process. For example, several models focus on the relationship of information seeking and uncertainty [Krikelas, 1983, Kuhlthau, 1991, Wilson, 1999], providing the basis for evaluating the performance of information seekers using World Wide Web as an information resource [D’Ambra and Wilson, 2004]. These models might also be useful in different stages of system development, for example, Ellis’s model has been applied to academic lawyers to generate tool design suggestions [Makri et al., 2008] and Foster’s model provides a framework of coding to analyze results [Foster and Urquhart, 2012].

The reviews of models of information seeking in this section pay particular attention to models of active information seeking in which voluntary action is triggered by recognizing information needs that can provide guidance for the design, development and evaluation of a system. Rather than a thorough review of every aspect of these models, we focus on the parts that can benefit the formalization of IBKC.

The scope of the field of information behavior can be illustrated using Wilson’s nested model [Wilson, 1999]. As described in this model, information behavior,
information seeking, and information searching are placed hierarchically as different levels of fields (See Figure 2.1). Information seeking is “particularly concerned with the variety of methods people employ to discover, and gain access to information resources”, while its subset, information searching, is about “the interactions between information user and computer-based information systems”. Information behavior represents a more general field that contains the previous two fields.

Figure 2.1: Wilson’s nested model of information behavior (adapted from [Wilson, 1999])

The three related terms, information behavior, information seeking behavior, and information searching behavior, were then formally defined by Wilson in 2000 [Wilson, 2000]. He described information behavior as “the totality of human behavior in relation to sources and channels of information, including both active and passive information seeking, and information use”. He described information seeking behavior as “the purposive seeking for information as a consequence of a need to satisfy some goal”. And he described information searching behavior as “the ‘micro-level’ of behavior employed by the searcher in interacting with information systems of all kinds.”

Summarizing the available definitions of information seeking behavior, Case [Case,
2012] describes information-seeking behaviors as “a process of either discovering patterns or filling in gaps in patterns previously recognized”, which differs from other information behaviors in its emphasis on purposive activity and close relation to the information need. Information seeking usually involves behaviors of identifying information needs, seeking information to satisfy certain information needs.

Our current understanding of information seeking origins from Wilson’s information seeking model, one of the pioneering models of information behavior that firstly systematically describes the process of information seeking. The model incorporates important concepts within the field, such as information need, information seeking, and information use, are organized in the form of diagrams that attempt to describe the relationships among stages in information-seeking behavior. To be more specific, as shown in Figure 2.2, the model suggests that recognizing information need is the trigger that motivates an individual to seek information. In order to satisfy the perceived information need, the individual makes a demand on information sources or systems. If successful, the gathered information can either fully or partially satisfy the information need. Partial satisfaction may lead to further information seeking. This systematic description of information seeking indicates that information seeking can be modeled as an iterative process driven by constant evaluation of the satisfaction of information needs.

Similar to Wilson’s model, but resides in the context of library search, Krikelas’s information seeking model is one of the first models that depict information seeking explicitly [Krikelas, 1983]. Krikelas’s model represents information seeking with a simple flowchart (see Figure 2.3). However, actions and entities in such representation has no bidirectional or nested relationship. It also emphasizes the important role of uncertainty as to motivate information seeking. The comments received from other researchers include, for example, the environment should serve as the context of information seeking by surrounding the other entities rather than
Figure 2.2: Wilson’s 1981 model of information behavior (Adapted from [Wilson, 1981])

appearing as a box in the flowchart [Henefer and Fulton, 2005]. Case highlighted one of the model’s contributions to be “placing the use of literature in the context of other sources of information, such as other people’s observations and memories” and it implies the need for collaboration.

Krikelas’s model emphasizes the importance of elements other than information behaviors in shaping information seeking. Elements such as disciplines (e.g., library science), settings (e.g., synchronous or asynchronous, collocated or distributed), data (e.g., structured), and attributes of participants (e.g., skilled, knowledgeable), can play important roles in practicing a specific information seeking task.

Information seeking can be represented in a manner similar to a state machine, while from the process perspective, these states can also be grouped into phases according to the of activities involved. The behavioral framework proposed by Ellis
Figure 2.3: Krikelas’s information seeking model (adapted from [Krikelas, 1983]) describes information seeking activities as consists of several “phases”, including starting, browsing, chaining, monitoring, differentiating, extracting, verifying and ending [Ellis, 1989]. He carefully pointed out that the sequence of these activities might vary and be iterative, rather than follow a fixed order. This model has been extended or used to study information seeking behavior of social scientist [Meho and Tibbo, 2003], economists and business analysts [Thivant, 2005], and humanists [Baruchson-Arbib and Bronstein, 2007]. However, this model does not describe the context of the activities and how information needs are generated [Järvelin and Wilson, 2003].

Similarly, Kuhlthau proposed a six-stage Information Search Process (ISP) model that includes stages of initiation, selection, exploration, formulation, collection, and presentation 2.4. In addition to actions (physical), Kuhlthau describes information seeking from a user’s perspective by considering feelings (affective) and thoughts (cognitive) [Kuhlthau, 1991]. The focus on actions, cognition, and emotions makes
it distinct from the other information seeking models which describe and analyze information seeking at different levels. She also sees uncertainty as a motivation for information retrieval (similar to Krikelas’s model).

Figure 2.4: Kulthau’s information search process (ISP) modal (adapted from [Kuhlthau, 1991])

By comparing the work of Ellis and Kuhlthau by also presenting Ellis’s characteristics as a process-based model, as shown in Figure 2.5, Wilson pointed out that the two models share strong similarities while Ellis’s model “specifies the modes of exploration or investigation” [Wilson, 1999], making them address issues at various levels. Nevertheless, these two models imply that information seeking can be conceptualized as a multi-step process where each step has relatively clear inputs, process, and goals.

Godbold [Godbold, 2005] summarized the previous models [Brookes, 1980, Dervin, 1998, Ellis, 1989, Kuhlthau, 1991, Wilson, 1981, 1999] and pointed out that information behavior described in these models, especially flowchart models of information seeking that illustrate the deterministic sequence of actions, could be multi-directional rather than a linear multi-phase process, and there might exist alternative strategies for bridging information gap between identifying information needs through sensemaking and starting to seek information.

There are also non-linear models of information seeking behavior developed
that attempt to provide a vertical description of information seeking. Foster’s model [Foster, 2004] of information seeking behavior describes components involved in information seeking as a nested square (see Figure 2.6), including external context, internal context, cognitive approach, and core processes (from outer to inner). Core processes consist of an opening - orientation - consolidation triangle, which represents constructs that are directly observable. This types of models can provide another viewpoint for researchers to design study and analyze observation.

### 2.1.2 Collaborative Information Seeking

This section describes the recent transition from individual information seeking (IIS) towards collaborative information seeking (CIS). Recently researchers are starting to explore the collaborative nature of information seeking due to that the more accessible ubiquitous information, the growth of information volume and more complicated context make information-intensive more complex for individual to handle [Byström and Järvelin, 1995].

Studies of CIS have been conducted within a variety of settings, domains, and disciplines. Research into CIS draws from work in computer-supported cooperative work (CSCW), computer-supported collaborative learning (CSCL), information
Figure 2.6: Foster’s non-linear models of information seeking behavior (adapted from [Foster, 2004])

science (IS), information retrieval (IR), and human-computer interaction (HCI) [Foster, 2007, Shah et al., 2014]. CIS’s relationship to information seeking, information retrieval, and collaboration is illustrated in Figure 2.7. Researchers have focused on a range of aspects relevant to an understanding of CIS. Related studies are summarized by Foster, as shown in Table 2.1. In addition to searching, CIS behaviors might also include collaborative querying, filtering, and navigating [Foster, 2007], co-browsing [Gerosa et al., 2004], collaborative sensemaking [Paul and Reddy, 2010], and collaborative-grounding activities [Hertzum, 2008]. CIS is an important component of CIB while other types of CIB also contribute to the shaping of CIS, thus a better understanding of CIS cannot be achieved without considering CIB as a broader context.

Foster [Foster, 2007] defined CIS as “the study of the systems and practices that enable individuals to collaborate during the seeking, searching, and retrieval of information”. In a more recent work, CIS was defined as “an activity in which two
or more individuals work together to seek needed information in order to satisfy a goal” [Karunakaran et al., 2013]. The definitions of CIS suggest that CIS is essentially an information-seeking activity that involves collaboration but the focus of CIS is on enabling collaboration for information seeking.

Derived from distinct domains, researchers are working on developing models to better understand collaborative aspects of information behavior. Based on findings from two ethnographic field studies of patient care teams in two different hospitals [Reddy and Jansen, 2008], Reddy and Jansen developed a model that describes the major difference between individual information behavior (IIB) and collaborative information behavior (CIB) by conceptualizing the information environment along behavior- (from lower-level information searching to higher-level information seeking) and contextual- (from individual to collaborative) axis. In addition to
describing differences between IIB and CIB, this model can also identify various
factors that trigger the transition among different levels of information behaviors.
The major differences between IIB and CIB in terms of communication, triggers,
and information retrieve technology are summarized in Table 2.2. Descriptions of
different levels of information behaviors, both individual collaborative, are given
in Table 2.3. Karunakaran, Reddy, and Spence further elaborated this situated
model by conceptualizing CIB to be comprised a set of activities organized into
three major phases: problem formulation, CIS, and information use [Karunakaran
et al., 2013], as shown in Figure 2.8. Besides activities that are specific to certain
phases, information sharing and evaluation and collaborative sensemaking occur in
all phases. Activities labeled “shared” are the ones that differ CIB from IIB.

![Figure 2.8: Karunakaran, Reddy and Spence’s model of collaborative information
behavior in organizations. (adapted from [Karunakaran et al., 2013])](image)

By reviewing and analyzing previous work on collaboration, Shah proposed a
Table 2.1: CIS research summarization (adapted from [Foster, 2007]).

<table>
<thead>
<tr>
<th>Settings</th>
<th>Aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>web search [Morris and Horvitz, 2007]</td>
<td>awareness of the different types of information and knowledge that are required by group members to support collaborative work [Sonnenwald et al., 2004]</td>
</tr>
<tr>
<td>health care [Reddy and Jansen, 2008]</td>
<td>different types of collaborative information behavior [Talja, 2002]</td>
</tr>
<tr>
<td>crisis management [Bjurling and Hansen, 2010]</td>
<td>strategies for collaborative information seeking [Poltrock et al., 2003]</td>
</tr>
<tr>
<td>patent processing [Hansen and Järvelin, 2005]</td>
<td>factors that affect the efficacy of information seeking [González-Ibáñez et al., 2013, Gorman et al., 2002]</td>
</tr>
<tr>
<td>library [Twidale et al., 1997]</td>
<td>problems of and support for collaboration during information seeking [Hyldegård et al., 2015, Shapira et al., 2001]</td>
</tr>
<tr>
<td>academy [Sonnenwald et al., 2004]</td>
<td>the design of appropriate technology [Sonnenwald et al., 2004]</td>
</tr>
<tr>
<td>industry (e.g., software design) [Poltrock et al., 2003]</td>
<td>collaborative sensemaking in CIS [Paul and Morris, 2009a, Paul and Reddy, 2010]</td>
</tr>
</tbody>
</table>

The model is built on the idea that concepts within collaborative work such as communication, contribution, coordination, cooperation and collaboration (C5) could be organized and interpreted using a layered structure in which layering represents dependency relationship. In C5 model, the differences among these five activities are summarized using six variables, including interaction, intent,
### Differences between IIB and CIB across different items (adapted from [Reddy and Jansen, 2008]).

<table>
<thead>
<tr>
<th>Items</th>
<th>IIB</th>
<th>CIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>Limited to Questions &amp; Answers</td>
<td>Plays a more central role</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triggers</td>
<td>Gap between current situation and future task demands; Lack of information</td>
<td>Complexity of information need; Fragmented information resources; Lack of domain expertise; Lack of immediately accessible information</td>
</tr>
<tr>
<td>IR Technology</td>
<td>Primary medium to search for information</td>
<td>Plays a supporting role, supports coordination/collaboration among information seekers</td>
</tr>
</tbody>
</table>

### Differences between IIB and CIB across different levels (adapted from [Reddy and Jansen, 2008]).

<table>
<thead>
<tr>
<th>Level</th>
<th>IIB</th>
<th>CIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information behavior</td>
<td>Simple information problems, direction interaction with a single system</td>
<td>Complex information Problems, importance of communication, interaction with multiple systems</td>
</tr>
<tr>
<td>Information seeking</td>
<td>Use of a single system. Relatively little interaction with other people or systems</td>
<td>Use of multiple agents (people and system). Significant interaction with other people and systems</td>
</tr>
<tr>
<td>Information searching</td>
<td>Direct (Q&amp;A) Interaction Mode</td>
<td>Conversational interaction to address exploratory search, problem solving, decision making</td>
</tr>
</tbody>
</table>

trust, human involvement, the symmetry of benefits, and level of awareness. This model could help identify what level of collaboration is required, or to what level collaboration should be supported given a predefined task.

Other than vertical models, Nonaka and Takeuchi develop a procedural model that describes the transformation from information to knowledge in CIS process as four steps: internalization, socialization, externalization and combination [Nonaka
and Takeuchi, 1995]. Knowledge is firstly acquired by individuals through information provided from various sources. Then gained knowledge is shared across group members. After that, collaborators assemble this information in order to satisfy the information need. Finally, the resulting product is organized appropriated depending on the goal of the task. Socialization is the key step that transforms process-based model from individual to collaborative.

Primary motivations of Collaboration include requirement or setup, the complexity of the information (e.g., ambiguity, large volume, various sources), role-based distribution of information (lack of immediately accessible information), and diversity of skills (lack of domain expertise) [Karunakaran et al., 2013, Paul and Reddy, 2010, Shah, 2015a]. The last two reasons can be categorized as the intent of benefiting directly from collaborates. The complexity of information could also be a higher-level motivation that raises the need for dividing labors and taking advantage of individual expertise.

In an empirical study conducted in an emergency department (ED) environment, Paul identified ambiguity of information a trigger for collaborative sensemaking [Paul and Reddy, 2010]. Her study revealed that the ambiguity of information could be caused by missing information, incorrect and inconsistent information.

In the same study [Paul and Reddy, 2010], Paul also pointed out that a person might in demand of collaboration because he is unable to interpret information correctly. For example, disease symptoms are experienced by patients while a doctor is capable of interpreting the information, and a patient can be cured by patient-doctor collaboration. Researchers have designed several systems to support information seeking by leveraging individuals’ expertise. For instance, Answer Garden can forward a question to a person who is equipped with required expertise [Ackerman and Malone, 1990]. However, designing systems to “facilitate a sharing of knowledge in a team and allow searchers to benefit from synergetic effects by leveraging diverse sets of knowledge brought in by different people” is
still challenging due to the difficulty for performing effective and efficient collaboration [Shah and González-Ibáñez, 2011]. In these cases, collaboration is in need since the completion of tasks requires diverse expertise and skills that an individual is not equipped with.

In some information-intensive tasks where there is a need to “speed up” information seeking, an effective division of labor is desired. Pirolli [Pirolli and Card, 1999] presented a field study of knowledge crystallization task: a group of MBA students seeks information in order to compose a strategic business analysis report. The amount of information is too large that is beyond an individual’s processing power, thus assigning the workload to each group members is in need. As claimed by Morris, an effective division of labor “may result in advantages such as increased coverage of the information space, higher confidence in the quality of their findings, and greater productivity due to a reduction of unnecessary redundant work” [Morris and Horvitz, 2007].

In addition, Simon Knight and Karen Littleton discussed CIS from an educational perspective [Knight and Littleton, 2015]. They claimed that “CIS has the potential to bring together rich collaborative, and multimodal, contexts in which important learning processes may take place.” That is to say, in addition to the benefits mentioned before, collaborators could also learn from each other through CIS activities. This also applies to some political settings where collective intelligence and democratic outcome are desired.

Sharing is an important approach that contributes to collaboration. Artifacts can be shared across collaborators to support cooperative awareness. Traditionally, shared artifacts can be “documents, tools or working resources of any kind [Fuchs et al., 1995]”. In visual analytics research, shared artifacts can also be common workspace where findings, hypotheses, and evidence can be shared [Mahyar and Tory, 2014]. In CIS research, shared artifacts can also include search product and search process [Morris, 2008].
Collaboration brings benefits as well as costs in terms of communication, coordination and time. However, it is difficult to measure the benefits and costs of collaboration precisely. Shah raised two interesting questions: “if two people working together can find twice as much information as either of them working independently, was that a good thing?”; “The participants may not be able to find twice as many results, but what if they achieved better understanding of the problem or the information due to working in collaboration?” [Shah, 2012]. That is to say, in addition to task performance, there might be other factors, such as learning, engagement, and social interactions that are more important depending on the goal of the task.

2.1.3 Information Sensemaking

While the term “sensemaking” has been used in a wide range of disciplines and sometimes is defined differently [Weick, 1995], this work refers it as “understanding the meaning of” [Paul and Reddy, 2010]. Sensemaking is often modeled as an important information behavior of information seeking tasks [Dervin, 1998, Pirolli and Card, 2005, Pirolli et al., 1996, Russell et al., 1993, Savolainen, 1995] as making sense of information happens all the time over the course of information seeking. For examples, making sense of available information landscape before searching information can help users make plans of seeking information, and users select and collect materials for certain information needs by making sense of retrieved searching results.

According to Dervin’s sensemaking approach [Dervin, 1998], information sense-making aims to narrow the gap between the contextual situation and the expected outcome by identifying the difference between them. The approach emphasizes the importance of reducing uncertainty in enlightening information seeking, and highlights the importance of contextual factors and awareness of the situation in shaping information seeking.
Pirolli and Card [Pirolli and Card, 2005] analyze sensemaking from another angle by describing sensemaking as a process of transforming information into knowledge based on studies with intelligence analysts. In this model, as shown in Figure 2.9 sensemaking is an iterative loop that consists of both deductive top-down processes and inductive bottom-up ones. Note that the rectangular boxes represent information flow and the circles represent processes. The top-down approach emphasizes the role of a priori representation (schema) in guiding sensemaking by starting with answering questions required by a task, while bottom-up approaches focus on information exploration and schema construction by starting with data. In this model, information seeking is an earlier sub-process of a broader sensemaking process.

Figure 2.9: Pirolli and Card’s sensemaking process for analyst (adapted from [Pirolli and Card, 2005])

Researchers have also developed other models of information seeking to under-
stand the process, but these models are less relevant to my study. For example, Shenton and Hay-Gibson’s [Shenton and Hay-Gibson, 2011] model is applicable to a specific group of individuals (children and youth), Johnson’s [Johnson, 1997] and Williamson’s [Williamson, 1998] models come from fields other than information science, Byström and Järvelin’s [Byström and Järvelin, 1995] and Savolainen’s [Savolainen, 1995] models are developed for use with everyday life (non-work) context.

While earlier studies are able to capture and explain various important dimensions of information seeking behavior, they treat information seeking as an intrinsically individual activity by conceptualizing it primarily from the individual perspective, ignoring the collaborative aspects of certain tasks and activities [Paul and Reddy, 2010]. However, these literature highlights the importance of contextual and situational awareness of information environments in terms of supporting sensemaking.

Sensemaking happens in each phase of information seeking. This dissertation pays specific attention to the sensemaking of collaborative progress in nugget extraction phase. It concentrates on enabling collaboration by taking advantage of other people’s activities. A review of existing literature on sensemaking provides insights on designing support from both theoretical and applicable perspectives.

### 2.1.4 Information Foraging

To understand how users seek for information in sensemaking activities, information foraging theory was developed to modal human users’ information seeking behavior borrowing the ideas from optimal foraging theory [Pirolli, 2007]. This theory recognizes that information is usually distributed in a variety of data sources in a patch manner. Information foragers usually need to navigate through the patchy information environment, recognize and collect relevant pieces, and make decisions on whether to stay or leave for a new patch, to pursue one information source or
another in order to maximize their gain. The theory makes a parallel between food diets and different types of information that people collect. The task confronted by the foragers is to make predictions the utility of unseen information “food” based on proximal cues. These proximal cues may come from the contents of the foraging environment and from activities of the foragers. Proximal cues serve as the information scent that can help predict whether or not the distal contents are desired. A foraging environment is considered “richer” if the corresponding information scent reveals more prevalent or profitable and thus will attract more foragers.

Information foraging theory recognizes two interrelated environments, information environment, and task environment [Pirolli, 2007, Pirolli and Card, 1999]. The dynamically changing information landscape, including the covered landscape and the landscape yet to be explored, affects how well we adaptively engage task environment by structuring our interactions with available content. Information needs emerging from the embedding task environment will, in turn, determine information environment as well. Therefore, it requires a constant evaluation of the environments for information foragers to better seek information.

The patch-based information landscape, the concept of information scent and diet, and the typical information foraging strategies and behaviors, provide a methodology for analyzing information seeking behaviors and capturing the important aspects of information seeking tasks. Information foraging theory can then be used directly to guide the development of information seeking system, as long as we can figure out how these elements are instantiated in a specific context.

A document browsing technique, Scatter/Gather, was developed by applying the classical patch model from information foraging theory to user behavior in iteratively browsing information space represented by clusters of documents [Cutting et al., 2017, Pirolli et al., 1996]. Initially, users are presented with clusters of documents as patches and a summary is provided for each cluster as information scent. Users
will select some clusters of the documents and clustering is then applied to these
selected documents to produce new clusters. This iterative process continues until
information needs are satisfied, e.g., expected documents are found.

2.1.5 Information Need

Information need is an essential concept in information sciences. However, this
concept is largely ignored by studies about information-based tasks: existing work
either investigates the context or situation from which information need arises to
understand information tasks [Case, 2012, Kuhlthau, 1991, Wilson, 1999] or design-
ing information retrieval systems that assume a known information need [Jensen,

In order to address the problem, Cole developed a theory of information need
which links information access to knowledge formation by relating eight informa-
tion need “surrogates or adjacent concepts” (information seeking, search and
use; problem, problematic situation and task; sensemaking and information forag-
ing) [Cole, 2011] and vertically interpreting Taylor’s classic four-level information
need model [Taylor, 1967].

Taylor’s model defines four levels of information need: from the deepest Q1
level of unconscious and visceral need, to Q2 level of the need that a brain can
describe, to the need with a formal expression Q3, and finally to top Q4 level of the
need with a compromised expression (the question presented to the information
system). The top-level information need is at command level (computer language)
while the other three deeper levels of need are at question level (human language).
In Taylor’s model, formulating the correct query in information search task is a
process to negotiate questions that consist of Taylor’s four information need levels.

Cole’s information need model is depicted in Figure 2.10. The concreting of an
information need is a process that transforming a human-generated (probably vague)
query to one that an information system can take. Transformational information
use events, such as generic knowledge, social context, and past experience, may hook Q4 level need into deeper ones. In particular, the deepest level of information need is reached by going through a “tunnel” from an evolutionary adaptation perspective [Lewis-Williams and Pearce, 2005].

Figure 2.10: Cole’s theory of information need (adapted from [Cole, 2011])

Cole’s information need model highlights several important characteristics of information need. Firstly, although it is commonly assumed in information science (as described in previous descriptions of information seeking models) that information need may evolve over the course of a task, Cole emphasizes that information need, once instantiated to its deepest Q1 level, will remain unchanged. Instead, over the course of the exploration stage (Kuhlthau’s ISP Model) of information seeking, the aspects of topics a user intends to investigate evolve.

Secondly, Cole’s conceptualization of information need reveals that Q1 level need is an unconscious, intangible, and visceral black box that users are not able to express or specify using linguistic terms to information retrieval systems. However, the model also indicates that sensemaking, as a continuous and existential-level
quest that bridges the gap between contextual situation and expected output, constitutes a major part of the deepest Q4 level information need. In other words, sensemaking can help make information needs concrete if they are difficult to be externalized.

2.1.6 Summary


Theories and practices of collaborative information seeking illustrate the benefits of enabling collaboration, including easing the complexity and ambiguity of information, making information more accessible, taking advantage of the diversity of skills, and allowing participants to learn from each other and perform tasks more efficiently. While it also comes with costs, such as requiring activity awareness and coordinating works.


Information foraging helps the modeling of human behavior in information seeking [Pirolli and Card, 1999], while sensemaking is a key component that helps people to select the best strategy [Pirolli et al., 1996].

Since IBKC is a special version of information seeking in local civic engagement of which the goal is to achieve better information coverage, the review of previous studies on information seeking in general or in other contexts can provide valuable insights into the understanding of information seeking in a different setting. Although some adjustments have to be made, the understanding form the basis for the conceptualization of IBKC framework later in this chapter.
2.2 Background on Knowledge Crystallization

Knowledge crystallization is a task that primarily attempts to address the “big data” problem [Shneiderman, 2008]. Knowledge crystallization is aimed at finding the most compact description possible for a set of data relative to a predefined task. In information-intensive tasks, the problem could imply, on the one hand, the volume of data is too big, on the other hand, data is difficult to interpret. In both cases, individuals are not capable of performing the task and collaboration is needed. Collaboration in knowledge crystallization has roots in social construction theory since during the process group members collaboratively construct shared knowledge through a social negotiation of meaning [Gunawardena et al., 1997]. Collaborative learning is also possible since personal knowledge can be constructed and shaped through group interaction.

![Figure 2.11: Knowledge crystallization (adapted from [Card et al., 1999])](image-url)
Card and colleagues [Card et al., 1999] first coined the term “knowledge crystallization” for the purpose of explaining how information visualization amplifies human cognition. They define knowledge crystallization as a task in which a person gathers information for a purpose, makes sense of it by structuring the discovered knowledge into a schema, and then packages the findings into some form of communication or action (see Figure 2.11). The results could be a briefing, a short paper, or even just a decision. Knowledge crystallization consists of the ill-structured activities required to sift through information available from a set of heterogeneous sources and develops a product that crystallizes the information into a more easily assimilated form. They recognized that effective knowledge crystallization requires several subtasks to be performed: foraging, searching for schema, representing findings using the schema, and compacting them into clear statements that reduce the amount of communication to a level that is amenable to human mental processing [Resnikoff, 1987]. Recently, a modified version of knowledge crystallization was used to address the task of prioritizing patients, serving as a foundation to improve the delivery and utilization of information [Pollack et al., 2014].

Knowledge crystallization is usually achieved by making sense of the data, constructing a representational framework, and packaging it into some form of communication or action. The value of the knowledge crystallization model lies in its ability to improve the efficiency of acquiring and processing data in order to accomplish a stated task [Pollack et al., 2014]. Knowledge crystallization was initially envisioned as a tool to create a schema for information visualization models. However, use of knowledge crystallization extends to most information seeking activities.

Knowledge crystallization is motivated by information overload problem [Gross, 1964]. Information overload refers to the difficulty a person can have making sense of massive complex information due to limited human perceptual and cognitive
capacity. The amount of stored information is growing exponentially nowadays, making such problem more challenging. In this context, knowledge crystallization becomes one of the solutions to information overload problem by producing a most compact description possible for a set of data.

In addition, a huge amount of data available on the Internet is a heterogeneous collection of data that have no rigid structures [Uno et al., 2004]. Worse still, data in the real world is usually problematic for its inaccuracy, incompleteness, or ambiguity [Strong et al., 1997], and thus difficult to comprehend directly. Knowledge crystallization becomes one of the solutions by discovering knowledge hidden in the data and structuring it by a reasonable schema.

In order to examine information foraging behavior, Pirolli conducted a study that involves both individual and group knowledge crystallization tasks. In the individual setting, a business analyst is required to write a business intelligence newsletter by foraging information from multiple sources. In the collaborative setting, a group of MBA students is asked to compose a strategic business analysis report. In both settings, they found that a knowledge schema is useful for them to forage and make sense of incoming information. They also found that the foraging activities usually involve a process of judging the potential relevance of information, and one that refines, improves and enriches information. They observed that “the analyst judges the relevance of articles by scanning titles and skimming, rather than by fully reading them”, and they considered the role titles play here is so-called “information scent”.

The above-mentioned tasks of knowledge crystallization and IBKC share some common features: they are motivated by the information overload problem; they need schemas to help frame the information need and guide various information behaviors; the major operations include extract, collect, and assess. However, community-level civic engagement context, stratifiedly selected citizen panels as participants, and the nature of issues introduce additional constraints that charac-
terize IBKC, which are covered in the next section.

### 2.3 A Conceptual Framework of IBKC

A conceptual framework is an illustration of linked concepts. They are extremely useful for understanding complicated phenomena [Jabareen, 2009]. IBKC is such a phenomenon, as it involves connecting multidisciplinary bodies of knowledge. This section develops a conceptual framework for understanding IBKC and it serves as my basis for us to design, implement, and evaluate the use of tools that support IBKC, particularly paying attention to the nugget extraction phase of IBKC.

Based on the definitions [Case, 2012, Wilson, 1999, 2000], an essential part of IBKC is information seeking as IBKC involves various information behaviors, including making sense of data, extracting information nuggets, evaluating the completeness of extracted information nuggets, making claims using collected nuggets, and presenting improved statements to inform the public. Therefore, IBKC is generally conceptualized as a special case of information seeking with the need of achieving better information coverage and in the specific context of local civic engagement.

Although the actions and steps involved in information seeking vary depending on the application, scenario, and constraints, some commonalities are obvious. As revealed by theories described in the previous sections, information seeking is a stage-oriented process that usually involves phases of identifying information needs, exploring information, collecting information, and using information [Ellis, 1993, Krikelas, 1983, Kuhlthau, 1991]. The process proceeds in a generally sequential manner with feedback loops between phases [Foster, 2004]. However, the sequence is not necessarily determined and alternative strategies may exist to link these phases as the practice evolves [Godbold, 2005, Wilson, 1981].

Inspired by Pirolli’s sensemaking loop for intelligence analysis [Pirolli and Card,
In 2005, the process of IBKC is also conceptualized as a series of phases with feedback loops that improve information structures. As shown in Figure 2.12, looking from the lower left- to the upper right corner, IBKC transforms raw data to accessible products through a series of phases while each phase has an input, information behavior, and output. A phase could proceed in a reverse way if the required information is not fully elaborated. In short, IBKC is conceptualized mainly as a process that transforms raw data into accessible products using connected entities and these entities are connected with information behaviors.

Figure 2.12: The sensemaking framework for supporting Knowledge Crystallization

The rectangular boxes represent entities involved in the process. The arrows represent flow relationships among them. This process has four small loops and has one set of loops that cycles around knowledge evaporation and another that cycles around knowledge condensation, with plenty of interaction between these. A
complex task is decomposed into a chained sequence of subtasks, where the output of one becomes the input to the next. This process proceeds under the guidance of a schema and helps refine the schema. The processes and data are arranged by the amount of effort and degree of knowledge crystallization. A bigger rectangular task wraps the entire process and serves as the context.

The IBKC process has many dimensions. Here I focus on the dimensions that most differ IBKC and other prevalent information seeking tasks, especially those that are particularly interesting from the perspective of system design. The dimensions include the problem context of local civic engagement, the goal of completeness, the expected transformation of data, the characteristics of participants, the role schematization plays, and involved information behaviors.

### 2.3.1 Problem Context: Community-level Issue Review

It is widely recognized that capturing all the relevant information and making it easily accessible to local citizens is an important step to facilitate civic engagement [Dewey and Rogers, 2012]. Providing people with access to more information is not enough, digital government technologies must go beyond an information repository, and provide help to citizens who can maximize their attention to information that will be useful for their judgment and the formation of their opinion. In the context of civic engagement, knowledge crystallization takes all the data that we can collect about a particular issue or subject, and puts them through a systematic process of distilling relevant nuggets, purifying, abstracting, and compacting to create a best and most accessible form of knowledge that can be understood and trusted by the public.

We call this type of knowledge crystallization *Issue-Based Knowledge Crystallization* (IBKC) since it is a special version developed for community issues. It is consistent with the work of Card et al. [Card et al., 1999] who recognized that effective knowledge crystallization requires several subtasks to be performed: infor-
mation foraging, knowledge schematization (searching for a schema and representing findings using the schema), knowledge compaction, and knowledge communication. IBKC goes beyond the previous conception of knowledge crystallization by Pirolli and Card [Card et al., 1999, Pirolli and Card, 1999] in two significant aspects:

- IBKC extends knowledge crystallization from individuals to a group activity. As an ill-structured problem, crystallizing knowledge from a large volume of data can benefit from group work by a divide-and-conquer approach, complementary expertise and perspectives, and enriched judgments.

- IBKC is designed for the knowledge crystallization problem for analyzing issue-based community knowledge to inform public deliberation. In the practice of democracy, the perception of “good information” is not universal [Dervin, 1994]. Democratic deliberation requires that public opinion should be informed by a full and balanced understanding of a community issue and associated solutions. That imposes a strong constraint, i.e., that the outcome of IBKC cannot lack any essential knowledge that is present in the data.

2.3.2 Goal: Complete and Concrete

Knowledge crystallization is a solution to problems of overwhelming information, poor data quality and limited human cognitive resources by producing a compact description of original data that can prompt further decision-making. Traditional search-oriented information access applications [Kuhlthau and Tama, 2001] aim to retrieve relevant information given a query. The expected outcome of IBKC, however, is a compact knowledge representation that covers the provided information space. The information goal of IBKC is to achieve better coverage of issue-relevant information in terms of content and structure with a compact presentation.

In search-oriented applications, information access usually requires the formulation of a query to express (aspects of) an information need. The difficulties of
generating a query lie in the lack of a complete definition of information needs and users’ vocabulary to express information needs precisely. In IBKC, users do not look for any specific thing, but intend to learn more about the content of the document collection [Cutting et al., 2017]. That is to say, information tasks in IBKC lean more towards the information access concept of browsing.

2.3.3 Data: Large, Unstructured, and Heterogeneous

Government decision-making processes, both at state-level and at community-level, need to embrace big data [Bertot and Choi, 2013]: for examples, analyzing input from the crowd for policy making; aggregating information from multiple sources around certain issues; and presenting government datasets to the general public. In addition to the amount of data, issue-relevant data are often piece-wise, difficult to connect, redundant, and inconsistent. They may include false information, overly detailed information, or unverified (untrustworthy) information that creates more confusion than it is informing. Therefore, useful knowledge has to be extracted from the information and be contextualized for certain tasks [Rowley, 2007]. Furthermore, the textual data are unstructured and require in-depth understanding by humans, making it difficult for automatic approaches to conduct a semantic-level analysis.

2.3.4 Participants: Stratifiedly Selected Citizens

On the other hand, a variety of technological, educational, social, and cognitive barriers may prevent citizens from accessing the necessary knowledge required for public participation [Maciel et al., 2016]. Especially at the community level, a lot of information and discussion related to an issue is dispersed across a wide spectrum source [Kavanaugh et al., 2014], making it challenging for ordinary citizens to get involved effectively in government decision-making processes. In IBKC, the difficulties caused by “civic knowledge deficit” are passed on to the small strategically selected citizen panel that represents representatives of a community,
making the process more controllable.

### 2.3.5 Schema: A Terse, Structured Summary

Schematization is defined as information processing that forms representations of an environment [Olivier and Gapp, 1998]. Such representations enable people to navigate through the environment without the need of going into details [Herskovits, 1998]. From the perspective of cognition, externally representing the information environment in an organized, integrated, and structured fashion can reduce memory load and thus facilitate more efficient internal information processing [Tversky, 2003].

The construction and population of an information schema is an indispensable component of sensemaking that helps streamline information foraging activities [Heer and Agrawala, 2008]. It needs to be pointed out that generating a perfect schema is usually difficult as it has to be contextualized to specific needs [Freksa, 1999].

IBKC usually requires a schema that reflects the structure of the data. As this dissertation concentrates on the information seeking activities in IBKC, the schema is considered as a given in the form of a set of user-generated themes that depict different aspects of an issue [Card et al., 1999]. IBKC schematization is able to efficiently assist participants to make sense of data, to rapidly formulate a compact description of the data and to accessibility present the outcome to the intended audience. The knowledge crystallization schema is important to information foraging activities but it is not the focus of this work. Therefore the scheme in this dissertation is simplified to a set of user-generated themes that describe different aspects of a community issue.

### 2.3.6 Information Behavior

In this section, a description of information behaviors involved in IBKC is provided using a metaphor borrowed from chemical engineering, where the goal of crystallization is to produce a highly purified and ordered crystal lattice from raw materials.
through the processes of purification and condensation [Beckmann, 2013].

Sensemaking is a key information behavior that helps people to select the best strategy [Pirolli et al., 1996] during information seeking, which happens over the course. Previous literature highlighted the importance of contextual and situational awareness of information environments in terms of supporting sensemaking [Morris and Horvitz, 2007, Paul and Morris, 2009b, Paul and Reddy, 2010]. Therefore, in addition to the procedural description of the framework, factors described in previous sections (schema, data, task goal, participants) are positioned in Figure 2.12 in order to give shape to the framework.

In the beginning, as provided data are usually not directly usable for supporting effective knowledge delivering, decision-making and problem-solving, information nuggets are introduced as an intermediary. An information nugget refers to a piece of knowledge in the data that has the potential to be useful for constructing knowledge crystals. The usefulness of a nugget is determined by its relevance concerning the schema and its potential of contributing to claim-making.

After extracting nuggets, analysts are able to produce knowledge crystals, which represent clear, solid and self-contained information, in light of collected nuggets. A knowledge crystal is expected to be constructed by collecting one or several nuggets and organize them in a meaningful way.

Knowledge crystals can be reduplicated, conflicting and redundant and thus cannot meet the requirements of the tasks. A variety of operations can be performed in this phase to improve the quality of knowledge crystals to make them more compact, defensible and understandable. For example, duplicates can be resolved by merging and paraphrasing; redundant knowledge crystals will be discarded.

Knowledge crystals of high-quality are considered to be exquisite knowledge crystals. Exquisite knowledge crystals are expected to be presented in an appropriate way, depending on the particular goal of the task, the characteristics of the resulting crystals, and the intended audiences. For example, if analysts involved in the
knowledge crystallization process are experts while targeted audiences are non-
experts, the presentation of exquisite knowledge crystals should be more accessible
and understandable.

2.4 Community Issue Review (CIR): an IBKC im-
plementation

The conceptual framework of IBKC described in the previous section provides an
instrument to understand information seeking in IBKC and characterize the unique
aspects IBKC from a theoretical perspective. Community Issue Review (CIR),
as a concrete implementation of IBKC, can help people analyze the difficulties
encountered in the practice of IBKC.

CIR is a community-level panel-based deliberation process for crystallizing
knowledge about a pending community issue [Kropczynski et al., 2015a] that
is specially tailored to the need of informing the public on local policy issues.
However, some of the actual difficulties encountered are not captured in a theoretical
simplified approach (insights can be gained though) and can only be discovered
through practice. This motivated me to conduct an illustrative case study using
CIR. Reflections on the practice provide design implications for supporting nugget
extraction and document sensemaking (before extracting nuggets) in CIR. Finishing
this study, nugget extraction was further investigated in terms of the collaborative
patterns and foraging strategies of participants with a formative study. Findings
from the study directly lead to the design of NuggetLens introduced in the next
chapter.

Oregon Citizens’ Initiative Review (OCIR) and our Community Issue Review
(CIR) are two instances of IBKC. OCIR is an institutional process whose primary
purpose is to enable informed voting on citizens’ initiatives while the outcome
of CIR is expected to inform the general public and allow them to be aware of
the controversy surrounding the issue. OCIR issues are at state-level while CIR issues are at community-level and are usually directly related to citizens. Different from OCIR, participation is voluntary in CIR, as long as the selected mini-public can represent the interests of stakeholder groups. As to how the documents and materials are collected and compiled, OCIR assigns a dedicated team to take up those responsibilities, while the selected citizen panel in CIR is allowed to modify them. Last but not least, OCIR is still relatively complicated and requires a supporting team to facilitate the process. By taking advantage of a hybrid approach, CIR is expected to be easy-to-moderate.

In order for a CIR process to happen, three conditions must be satisfied: (1) there is a public issue that is pending for decision and is drawing the attention from the public, (2) sufficient expressions of concerns, policy choices, and preferences has been openly exchanged among the public, experts, and policy advisors, and (3) all the information sources about a policy issue have been identified and the information is collected. The product of conducting CIR is a set of Citizens’ Statements that can answer all the questions community members have [Boudjelida and Mellouli, 2016]. In the meantime, the Citizens’ Statements must be short, concise and comprehensive to minimize the time and effort required from citizens, and they must be agreed upon by the public if the process is to be a non-partisan analysis based on the full consideration of public goods. In particular, CIR re-interprets the information provided by experts and translates this into easily accessible expressions without technical jargons and unnecessarily sophisticated analytical details. CIR uses a citizen panel as issue analysts who are likely to evaluate and deliberate on the issue the same way as their peer citizens, which can increase trust and acceptability of the statements presented to the broader citizen audience. The products generated through CIR can provide the community with insights concerning the issue so that everyone in the community is able to form opinions effectively and thus has an impact on the public decision-making process.
CIR guides a group of panelists to in-depth review an issue relevant to the community through a multi-day public review process. Panelists are either randomly or strategically selected from a community. As representatives of a community, panelists are given access to a large amount of data from various sources concerning a given issue. Their task is to make sense of the data, extract important information nuggets from the data, make claims based on the extracted nuggets, and improve and organize the claims to generate an informative briefing of the issue in order to provide the community with insights concerning the issue so that everyone in the community is able to form opinions effectively and efficiently (see Figure 2.13).

This dissertation focuses on the processes of making sense of issue-related documents and extracting important information nuggets. The purpose of sensemaking in CIR is to provide participants with an overview of the contents of the given issue-related documents in summarized form, enabling users to decide what actions to take before formally harvesting information nuggets.

Nugget extraction in CIR is aimed at reliably recognizing and collecting all nuggets relevant to the pending issue, and it is the prerequisite for subsequent tasks of knowledge crystal formation, refinement, and compaction. The task involves a number of cognitive and physical actions. First, the CIR panel members are
charged with investigating a policy issue that is usually complex and controversial. At the same time, they are provided with a document collection that is considered to contain all the data we can find about the issue. Other than published reports, websites, and news articles, the document collection also contains interviews with subject matter experts and their written statements. During this phase, a citizen panel gathers information nuggets relevant to a policy issue through an online analytic forum where individuals can access all the documents to be analyzed, extract nuggets, and tag them by a particular theme. The collected nuggets are expected to cover all the information that can be found in terms of content.

2.5 Practicing CIR: an Illustrative User Study

In order to demonstrate the feasibility of conducting CIR with proposed IBKC framework and learn about the challenges and difficulties of supporting sensemaking and nugget extraction, we conducted a descriptive case study of practicing CIR to derive design considerations and principles.

Figure 2.14: Face-to-face meeting on Day 1
2.5.1 Study Setting and Design

This study is a part of GeoDeliberator Project. The community issue used for this study is “inflationary tax indexing”. This issue is a proposal to council that real estate tax should be increased by at least inflation every year just to keep pace with the cost of providing services to the Borough. Our research team consists of 1 professor, 1 postdoctoral scholar, 2 doctoral students (including me), and 2 undergraduate students. We recruited participants from local (State College) community by sending out emails and paper fliers to local residents with the help of State College Borough. 14 participants were selected from the those who are interested in participation considering stakeholders of the issue. Among these participants, 3 students are from college organizations, 3 students rent in the borough, and 8 are homeowners.

The study simulates the process of conducting CIR while adding certain sessions for exploratory purpose. In general, the study lasted for 10 days, including two face-to-face meetings held on both the first and last day. In the first-day meeting, the postdoctoral scholar from our team and an official from borough government who is familiar with the issue presented the issue to the participants. There was also a training session on the introducing the features provided by the online platform. The undergraduate students in our team helped troubleshoot any problem participants had when playing with the system during the session. In the following days, participants perform CIR activities using the online system and our team sent out an email digest everyone that summarizes the progress and provides further instructions. In the last-day meeting, participants wrapped up their work and came up with a set of statements as the outcome. As there were still some controversial cases, our team further organized 3 follow-up sessions in which participants were encouraged to reflect on their experiences. Our team was on call during the study and thus participators were able to get necessary support when needed.

The direct observation and interview methods were used in this study. In
face-to-face meetings, undergraduate researchers observed participants in terms of seeking information using the online platform and took notes on represented or unusual behaviors. During the period between the face-to-face meetings, our team members kept track of participants’ online activity and intervened when necessary (e.g., a features is misused). After the CIR task was completed, our team interviewed the participants, and encouraged them to reflect on what they did in the study.

Several findings based on the participants’ feedback and our observations have been previously discussed [Sun and Cai, 2017]. For example, we observed that, in the context of collaborative information seeking, participants communicated quite well in face-to-face meetings, while they worked almost individually in the online environment, relying on a variety of communication channels (including a chatroom, question panels, and discussion boards). As mentioned by a participant, this disparity is due to time delays in asynchronous communication, while people do expect immediate responses or in-time notification. This phenomenon was explained by Curtis and Lawson [Curtis and Lawson, 2001]. A solution to address this problem in asynchronous information foraging environments is to enable implicit collaboration using user traces [Li et al., 2014]. However, in the beginning, there was no trace of user foraging activities on the information surface, which prevented implicit collaboration.

On the other hand, when participants were asked to come up with an initial set of themes to cover important aspects of a given community issue before reading document contents on the first-day meeting, they had little confidence because of lacking related knowledge. Such feedback calls for an approach to automatically analyze the information space and provide an overview that summarizes its content.
2.5.2 Reflections on Practicing CIR

IBKC is empowered by a mini-public. Currently, mini-publics are widely used in deliberative events, such as citizens’ juries, consensus conferences, planning cells, deliberative polls, and citizen assemblies [Elstub and Mclaverty, 2014]. Mini-publics contribute to these decision-making processes by providing collective recommendations and opinions based on the feedback obtained from a sample of citizens. Instead of making decisions directly, in CIR, the mini-public functions as an information processor and is tasked with analyzing the documents without expressing any opinions. Since the present study focuses on the early phases, the articulation of the text and the degree of compactness are not discussed further here. The expected outcome should cover all important information related to an issue objectively, allowing a broader citizen base to form opinions.

In addition, online participation plays an important role in CIR. It offers the flexibility of time and place by providing citizens with a much wider range of channels for participating in deliberative democracy process. However, it also limits the potential for generalizing the findings to a broader population due to the differences in computer skills [Min, 2007]. Our study participants were college students, who possess better computer skills than the general public, and this might have influenced the outcome of online participation. Our implementation of CIR addressed such limitation through a training session that was provided to all participants on the first-day of the study, as well as technical support throughout the process. Nonetheless, it still requires panelists to have basic computer skills, which an increasing number of the general public is acquiring.

From the technical perspective, in addition to the basic functions necessary for reviewing community issue, incorporating visualization features that are more intuitive for citizens to visually interact with the content in the system would be advantageous. Some effort has been made for supporting nugget extraction phase visually [Sun et al., 2016]. Preprocessing of information should also be useful to
enhance people’s analytical capability.

Our work in designing and supporting information seeking in CIR was fully informed by the recent advances in deliberative democracy theories [Curato and Böker, 2016]. Still, it clearly moved beyond the face-to-face deliberation paradigms of mini-publics. Towards supporting CIR online, we abstracted out the unique aspect of CIR and formulated into a type of knowledge crystallization problem. It is within this context that we seek to understand information foraging activities and develop approaches to support sensemaking and nugget extraction by combing human and machine intelligence.

2.5.3 Design Implications for Supporting Information Seeking in CIR

I have provided the basis for portraying IBKC as a process that involves information seeking and sensemaking, and that can be modeled using the information foraging theory. By reviewing existing works on conceptualizing, practicing, and supporting information-intensive tasks, I found that the understanding of collaborative information seeking in the civic engagement domain is far from sufficient, indicating that computational practices can be developed to support the process.

In this section, knowledge crystallization derived from information visualization was introduced, and the process was extended to the community-level civic engagement context. Specifically, CIR was illustrated as an implementation of IBKC. The findings from the illustrative case study indicate the possibility to enable implicit collaboration and the possibility to provide an overview of the document content are important design considerations.

Tools and systems for CIS have not “enjoyed widespread success” by far [Hearst, 2014]. Hearst described three scenarios current CIS tools are unable to address: (1) selecting a few from many similar choices, (2) covering a topic thoroughly, and (3) discovering unknown information [Hearst, 2014]. The knowledge evaporation phase
of IBKC can be mapped to the second scenario, while the knowledge condensation phase within IBKC is similar to the first scenario. This study focuses on the knowledge evaporation phase, the goal of which is to cover the provided information space thoroughly. Compared to the widely-studied information searching tool support (reviewed in Section 4.1), tools designed for supporting IBKC have several different requirements:

1. Searching is the design focus of many information seeking tools while identifying information needs and use of information serve as a (less important) context. In IBKC, searching is embedded in sensemaking process: during the process of making sense of provided source information, users need to seek information to elaborate the schema. Keyword-based search (or similar) becomes optional.

2. Information space appears to be infinite (e.g., library database, internet) for information seekers in many information seeking activities. Users only search and retrieve the potentially relevant information that can answer certain questions. Comparatively, in IBKC, users are expected to exhaust the whole information space. That is to say, in addition to satisfying information needs, another goal of the task is to ensure completeness.

3. In many information seeking processes, collaborators can learn from each other in terms of searching capability. In IBKC tasks, collaborators can benefit from others’ sensemaking results and use of information. In this sense, the learning behavior in IBKC shares some similarity with social tagging and bookmarking.

4. In terms of evaluation, identifying information needs is trivial in many information seeking tasks and thus judging the correctness of the products is easy. However, since IBKC is aimed at ensuring the outcome to be a compact version of input data that incorporates collective intelligence, assessing the
quality of the outcomes becomes more difficult than question-and-answer cases.

5. Another feature introduced by IBKC is the importance of schema for identifying information needs. Due to the nature of IBKC, a schema is required to organize both the intermediate information and the final products. It is expected to be composed initially when identifying information needs and to keep evolving during the process.

Although existing tools designed for CIS cannot be directly adopted in the IBKC context, this study can still benefit from them in terms of how to support communication, coordination, and sharing, how to promote situational awareness, and how to apply models and theories to direct the design and development of support systems.

Similarly, although text analysis combined with visualization has been studied in dealing with the understanding of large document collections (Section 5.1.2), the question how the techniques should be tailored to fit the CIR context remains challenging and thus requires further investigation.

2.6 Summary and Discussion

This section provides a conceptual framework for systematically describing IBKC by reviewing previous information seeking studies and characterizing the unique aspects of IBKC. It also derives multiple design implications by implementing IBKC as CIR and practicing the process of CIR. While there are many directions that this research stream can take, this dissertation particularly focuses on understanding the nugget extraction phase of IBKC and designing tools to support it. In order to do that, in the next chapter, I will formalize nugget extraction on top of the IBKC framework and highlight the difficulties users may have during the process so that we can design tools to address the identified problems.
Chapter 3  
Design Considerations for Supporting Nugget Extraction

The task of nugget extraction in CIR is fundamentally a problem of collaborative information seeking. It has been intensively studied in information science [Foster, 2006, Shah, 2013, 2015b, Shah et al., 2014]. As an information seeking problem, the task is driven by the information need to justify policy choices that address a policy issue. It shares some similarity with the form of “post-decision-making information seeking” recently studied in [Mishra et al., 2015]. However, decision-making in the public domain is different from the expert driven decision-making explored by Mishra et al. [Mishra et al., 2015] in the sense that there is a need to seek information after experts proposed a solution and before the authoritative body (such as a city council) makes a decision. This information seeking is driven by citizens’ need to understand the impact of a policy proposal in order to express their opinion. Information needs of such information seeking are unique in that they are driven by the agendas of public decision-making processes and there is greater uncertainty when it comes to assessing the future impact of public policies on diverse stakeholders. To our knowledge, crystallizing knowledge by mini-public for public decision-making (for surveys of research in information need, see [Case, 2002, 2006]) has not been studied and there is a need to observe, theorize, and understand it in order to support such activities.
3.1 Conceptualizing Nugget Extraction as Collaborative Information Foraging Activity

*Nugget extraction* in CIR is aimed at reliably recognizing and collecting all nuggets relevant to the pending issue, and it is the prerequisite for subsequent tasks of knowledge crystal formation, refinement, and compaction. The task involves a number of cognitive and physical actions. First, the CIR panel members are charged with investigating a policy issue that is usually complex and controversial. At the same time, they are provided with a document collection that is considered to contain all the data we can find about the issue. Other than published reports, websites, and news articles, the document collection also contains interviews with subject matter experts and their written statements. During this first phase of CIR, a citizen panel gathers information nuggets relevant to a policy issue through an online analytic forum where individuals can access all the documents to be analyzed, extract nuggets, and tag them by a particular theme. The collected nuggets are expected to cover all the information that can be found.

Nugget extraction CIR is a process of information foraging. Applying information foraging theory to nugget extraction, a patch refers to a group of “nearby” (closely related) sections within documents. The number of collected nuggets can be used for directly measuring gain and the within-patch activities benefit the cumulatively gained knowledge nuggets. There exist other types of gains as well. In order to optimize the outcome of nugget extraction, a group of panelists must select an appropriate strategy to fit the constraints of the environment and adjust the strategy to accommodate the changing environment to speed up the within-patch seeking and minimize the between-patch foraging costs. The task environment serves as the context of the information environment. It carries the information concerning the entities involved and their relationship, the evolution of the extracted nuggets and gained knowledge, and the task flow.
In other words, the foragers in nugget extraction, aside from assessing the relevance of a single nugget, also need to make predictions about the utility of unseen distal knowledge nuggets based on proximal cues. These proximal cues, in this study, involve both information directly related to the contents (e.g., headers, the table of contents) and information derived from activities (e.g., highlighted texts, reading heatmap), and serve as the information scent that can help predict whether or not the distal nuggets are desired. For example, the paragraphs that contain more nuggets, or within which the distribution of nuggets is more intensive, appear more appealing to information foragers.

3.2 Benefits and Costs of Collaboration in Nugget Extraction

Nugget extraction can benefit from collaborative work in many ways [Heer and Agrawala, 2008, Shah, 2015b]. First of all, the issue under consideration can be a deep and long-standing community concern that is complex and multi-faceted. The diversity of expertise and experiences panelists have related to that issue is beneficial to the analysis. Panelists must share their findings, discuss evolving hypotheses, generate a collaborative interpretation of the information, and contribute contextual knowledge that deepens the understanding. Second, the amount of data (documents in particular) can be so large that a thorough exploration by an individual is unlikely. A group of people can cover more documents in the same amount of time. Third, the number of themes to be considered can surpass what one can handle with active memory. In this case, a divide-and-conquer approach helps break down a complicated task into smaller ones and thus makes the task possible to accomplish.

In nugget extraction tasks that involve multiple persons, the process becomes more arduous [Grudin, 1988]. On the one hand, the change of the information environment becomes unpredictable for each individual. On the other hand, the
activities of group work become an indispensable context that shapes the information landscape.

Nugget extraction requires a group of users to constantly assess whether the collected nuggets can satisfy the information need. Such evaluation is driven by maximizing the \textbf{gain} per unit \textbf{cost} and they need to constantly plan and adjust strategies to guide their actions. The requirements of collaborative work bring the additional cost of establishing and maintaining awareness of one another’s activities and outcomes [Carroll et al., 2003]. The gain in this information foraging game is measured in two valued aspects:

G1 The quality and quantity of the extracted nuggets.

G2 The extent to which the group has depleted all the opportunities for extracting more novel nuggets.

According to the goal of IBKC, the task is considered perfectly done if G2 is 100%. The cost of foraging nuggets involves the cost of:

C1 reading documents and recognizing nuggets.

C2 extracting nuggets, selecting and tagging.

C3 moving the document view window from one part of the document to another, and navigating across documents. This is analogous to the cost of \textit{within-patch move} and \textit{between-patch move} in the foraging metaphor.

C4 planning foraging actions by assessing the information landscape and identifying the best opportunities for additional foraging.

C5 maintaining awareness and coordination among CIR panelists.

C6 assessing the task status in relation to the completion state.
The analysis of collaborative nugget extraction using information foraging theory calls for a support of promoting awareness of activities and information, and this provides an opportunity to incorporate a more efficient collaboration approach. Activity awareness refers to the awareness of joint efforts of promoting one collaborators’ intentions, actions, and results during complex tasks [Ganoe et al., 2003]. This chapter discusses how collaboration can be leveraged to support activity awareness in nugget extraction, when the information landscape is too large, the task structure is too complicated, and diverse expertise is required.

In order to develop a nugget extraction supporting tool that can take advantage of the power of collaboration while reducing the costs it imposes, we must seek a deeper understanding of CIR nugget extraction as a collaborative information foraging activity. From the theoretical lens of information foraging [Pirolli and Card, 1999] as information foraging principles apply to CIR nugget extraction activity, when CIR panelists collaborate to decide how to best allocate their efforts, they can benefit from good support in situation awareness [Chalmers, 2002, Dourish and Bellotti, 1992]. In an information-seeking situation, awareness refers to the information seeker being aware of various aspects of the searching and sense-making processes, including the task and its context, past and present actions, and various attributes of the information objects and the system [Shah, 2013]. This is a salient aspect of a collaborative information-seeking process when the information landscape is complex and dynamically changing, and the task lasts several sessions.

Theories of information foraging and collaborative information seeking provide high-level insights into the human behaviors and collaboration patterns, highlight the need for collaboration, and emphasize the importance of being aware of the foraging activities in CIR nugget extraction task. How such behaviors, collaborations, and group dynamics are shaped by the new context, however, remains unclear and needs investigation to elucidate specific requirements for awareness support. The identified requirements can further guide the design of support tools
among alternatives.

3.3 Understanding Analytical Reasoning in Nugget Extraction

Nugget extraction is a collaborative information foraging activity (section 3.1). The information foraging theory provides an abstraction of the goals of its sub-tasks (in the context of knowledge crystallization), the procedure and the list of actions necessary to perform the task. However, the details of the task structure are still vague, leading to the first question:

Q1. How do users mentally approach the nugget extraction task? In particular, how do they collaborate, coordinate, and communicate during the process? What are the decisive moments that collaboration is needed?

Also, the theory suggests the necessarily of making decisions of foraging, but it is unclear when and how exactly participants make such assessment. These problems lead to the second question:

Q2. What are the difficulties that prevent users from efficiently collaborate? How do they coordinate strategies to increase the chance of completion? What are the factors and indicators needed for them to make decisions (e.g., important cues and situational dynamics)?

To answer Q1, an exploratory formative study was conducted to explore how participants approach a relatively complete collection of nuggets in general. On that basis, we conducted a second study to answer Q2 and drill down further into their foraging strategies and reasoning process of evaluating situations through a semi-structured interview.
### 3.3.1 Study Setup

The community issue used in this study was borrowed from a real-world debate called *Collegiate Housing Overlay* (CHO). It was around a pending ordinance that allows taller than normally allowed buildings under the current commercial zoning that can incentivize commercial builders to include housing features (diverse apartment sizes, commercial areas on the first floor, etc) [Hartley, 2015]. Becoming knowledgeable about this issue enables the general public to vote on this proposal confidently.

Our research team for this study involves 1 professor and 2 doctoral students (including me). We recruited 6 students from an undergraduate visual analytic course and were organized into three pairs. A minimum number of group size was chosen in order to better control the study conditions. To ensure the reliability and consistency of our observations, the same system, documents, and instructions were used in the three sessions. Ideal participants should be selected from the community, while here students are recruited for exploratory purposes and convenience.

There was a training session before the tasks. We presented the participants with the features provided by the online platform. We also explained the concepts and common practices of nugget extraction in IBKC. During the tasks, group members extracted information nuggets from the given documents collaboratively until they believed all important information is covered. After each task, participants were encouraged to reflect on their behaviors. In particular, they were asked to explain their rationales for determining the completeness of nugget extraction.

Each panelist was given an account to use the nugget extraction workspace (see Figure 3.1). User’s interactions with the system, including traces of users’ reading and extraction activities, were captured and saved. The interface provides the functions necessary to perform nugget extraction: navigating through the document set, highlighting a piece of text for extraction, assigning a theme to the extracted knowledge nugget, and a chat box for users to coordinate their tasks and
The document set for the formative study included the proposed ordinance and three subject matter experts’ opinions. The four documents had a total word count of approximately 12,800. The participants were required to learn about these documents and explore them from two aspects: intended purpose (what the ordinance is designed to achieve), and proposed action (what exactly will be done if the ordinance is approved). The instructions given to the students were designed to simulate the real-world CIR nugget extraction task, in which panelists work asynchronously at home, without face-to-face communication. The students were allowed to coordinate their work in any way in order to achieve the best efficiency. Certainly, they were encouraged to take advantage of the system as much as they could.

Figure 3.1 shows the user interface of the system (version 1.0) used in this formative study. Individuals can explore/read documents in the Document View and extract nuggets by selecting a piece of text judged relevant. The extracted nuggets are listed in the Nugget List view. One can review any nugget by clicking...
on it, which will trigger a linking action to the Document View to show the context of that nugget. The panel is provided with a Chat Room that supports spontaneous communication, asking questions, and leaving notes for each other. Users can also communicate asynchronously through the chat room. The drop-down menu at the top of the interface enables users to filter nuggets by different themes.

The **Document View** is the container for a collection of documents. When the document view is in focus, a table of contents of the collection of documents pops up. Users can browse all documents in sequence or quickly retrieve one by navigating through the table of contents. In this case, a table of contents is more useful than search box as it illustrates the logical structure of the given document collection and can help users formulate the scheme (and plan) to crystallize knowledge through browsing documents more efficiently. In this view, participants can browse documents and extract nuggets via simple highlighting and tagging actions. Once a segment of text is selected, it will be highlighted with yellow color and invoke a pop-up menu with three available actions: comment, raise a question, or extract as a nugget. For “extract” action, there is one more step that asks users to assign this information nugget to a particular theme.

All the extracted nuggets are collected into the **Nugget List**. Once a nugget is newly extracted, it will immediately appear in the top of the list. In case a nugget is incorrectly assigned (to a theme), users can modify the assignment through the “reassign” action that is available from the context menu invoked by right-clicking the nugget. A nugget can also be removed if it is a duplicate or a mistake. Nugget List View is actively linked to the Document View, allowing users to trace back to where a nugget originates in document contexts. Capturing the relationship between nugget and its origin in documents effectively makes it possible to replay and review the analytical process later on [North et al., 2011].

While an individual extracts nuggets, he or she can maintain awareness of what nuggets have been extracted by others and from where in the documents. Nuggets
in the document view are coded in blue if it is self work; nuggets extracted by others collaborators are colored as green.

During the study, researchers conducted detailed observation and took notes on the observed patterns of participants interaction with the system in terms of document reading and nugget extraction, as well as their collaboration. In the middle of the task, we gently interrupted them and ask about their judgments and rationales, indicating their strategies, e.g., *How do you determine the degree to which your task is accomplished? How do you know where to look for information next?* After that, they resumed their task and researchers continued observation and activity recording. After task accomplished, we conducted another round of interview with the students to ask them reflect on their mental process, task performance, and working strategies. Applied Cognitive Task Analysis (APCT) [Militello and Hutton, 1998] was leveraged to elicit critical cognitive elements, especially the task structure, the decisions points, the decisions that must be made, and the cues used for assessment.

### 3.3.2 Data Analysis and Findings

The mental process of the participants behind the nugget extraction task was clarified and is provided in Figure 3.2. Before generating an initial version of information needs, the participants have a basic understanding of the document collection and they then come up with a set of themes to organize the content. With the information needs in mind, the participants make sense of the documents and, recognize and extract relevant and important information nuggets. The participants assess the completeness of the extracted nuggets regularly to decide whether the task is finished; otherwise, the information needs will be updated and further nugget extraction activities are required.

The formative study combined observations, system logs, and semi-structured interviews to dive deep into the analytical reasoning process. It formalized the
Figure 3.2: The nugget extraction task structure.

data analysis into a coding scheme for analyzing critical actions captured in video recordings and chatroom messages, including changing focus in the document space, hesitating to move, communicating with others, letting people know where they are, drawing others’ attention to things that are difficult, coordinating foraging strategies, and assessing their progress. We attempted to analyze the captured actions and interpret their intentions, some of which were later confirmed and explained through interviews. Findings drawn from the exploratory study are the following:

F1 Observations show that participants divided their work by either document subsets or themes in order to avoid duplicate work. However, dividing by
document subsets is preferred, as one participant summarized: “We split up the documents for doing both themes. I felt like if you try splitting up by theme you definitely need to re-read everything twice through.”

F2 While the task is a mixture of individual and collaborative work, the frequency of collaboration increased as they advanced towards the end of the task. This finding was mainly derived by analyzing the content and number of chat messages among group members. The participants did, however, communicate at the beginning to come up with an agreed plan.

F3 In order to ensure that no relevant information was omitted they tried various means to exhaust all content carefully and thoroughly. One participant said during the interview: “I was just going top-to-bottom and tried not to miss anything or jump around at all.” Segments of documents were considered complete if they were covered by highlights and more attention was paid to documents without any highlights. One participant said: “I put the filter to ‘all’ to see all she had highlighted, so I only read where she didn’t highlight to see if there is anything in that text that can be important.”

F4 Judging if the task is complete is considered difficult. Every group mentioned that they had to go through all documents to claim every important piece of information was gathered. Also, group members hesitated to declare the completion of the tasks, and they exchanged their opinions before the completion. Even so, they tended to have little confidence in the judgment since there was little conclusive evidence to support.

These findings helped us to optimize the collaborative information seeking process, especially the procedures participants follow, and the challenges and difficulties participants are confronted using the provided design.

Specifically, taking a document as the basic unit for organizing all the materials is preferred for ease of dividing work among a group, as suggested by F1. F2 confirmed
that the collaboration is especially needed in the later phase. As participants become more familiar with the materials, generate more traces in the workspace, and have difficulty in proceeding to further steps, collaboration can be conducted with the produced information and a better understanding of the data to address the difficulty. The major difficulties include avoiding duplicate work (F1), and the perception of uncertainty in evaluating task progression (F4).

Based on the analysis above, the system structures the provided data as documents, rather than themes, with a table of contents for navigation. A visualization tool is anticipated to provide support for evaluating the completeness of collaborative work and optimize the entire process effectively. Especially in the later phase, the tool will be used frequently.

To design the visualization tool, we need to understand what are the indicators with regard to the information and task environments to which panelists refer when making decisions for the purpose of more efficient information foraging. These questions are explored in the study described in the next section.

### 3.3.3 Indicators for Reasoning about Completion

We made several assumptions on the participants’ strategies and rules for foraging information during nugget extraction by observing their interactions with the system. The after-task interview allowed us to dive deep into their thoughts on the reasoning about situations, and to confirm our assumptions.

In general, participants tend to evaluate from two perspectives, i.e., the content on which the activity is centered and the individuals involved. For the former, the information space is considered covered if each of the included documents has been investigated. Depending on the relevance, a document will be judged to be sufficiently investigated based on the allocated time, the effort involved in making sense of the document and the extracted information nuggets as the outcome.

For example, evaluating the completion of the task involves the subtask of
evaluating coverage of document sensemaking, and can be further decomposed into
tasks of assessing the coverage of individual documents. That is to say, if a document
receives a great coverage in terms of the time spent and the nuggets extracted, the
document is more likely to be considered well-investigated. As more documents are
considered investigated thoroughly, the extraction task proceeds towards completion.
The coverage of a document, as well as the number of collaborators on the document,
can also serve as an indicator that the document might need further investigation.

According to the information foraging theory [Pirolli and Card, 1999], the
coverage represents the information landscape explored, the traces left by the
participants can be used to indicate the profitability of a information cluster,
and the awareness of these traces can reduce the cost of both within-patch and
between-patch information seeking. Previous interpretations of information theory
indicated (1) that more time spent on patches represents a higher likelihood a patch
contains relevant information and (2) that the profitability of a patch is defined by
the relevant information contained in the patch [Chen et al., 2002]. The findings
derived in this study further elaborate on these suggestions by incorporating more
details specified by the nugget extraction task.

It is worth noting that such analytical processes involve uncertainties. For
example, a panelist might interpret the information incorrectly or inaccurately,
certain evidence might not look convincing enough, and different individuals might
perform reasoning in different ways. As long as we take into account such uncertainty,
the general reasoning process of the situation could still serve as a persuasive
guideline for researchers to develop systems that can facilitate panelists to conduct
the process.

The four major indicators panelists tend to use for making decisions about
movement within an information landscape are compiled into the following rules:

ST1 nuggets have been intensively extracted from a piece of text (the opportunity
of foraging more nuggets depletes);
ST2 there are few nuggets extracted in a part of a document but this part has been frequently read;

ST3 the frequency of nugget extraction activity decreases;

ST4 the efforts are widely distributed across all documents.

If all four rules are satisfied, it is more likely that the information landscape has been explored thoroughly. Otherwise, the task is considered incomplete and participants will continue the work and select a strategy based on these rules.

ST1 points out that the number and distribution of nuggets can assist users to make choices of within- and between-patch movement. From the level of the document, ST2 reveals that users believe the reading pattern and the extraction pattern can jointly become an indicator of completeness. ST3 and ST4 both describe the situation from the task level. ST3 focuses on the change of the number of extracted nuggets overtime while ST4 considers the distribution and coverage of efforts. Sometimes some of the rules are considered together, for example, as mentioned by a participant, ST1 becomes stronger if ST3 is satisfied.

3.4 Design Implications for Supporting Nugget Extraction in IBKC

The reasoning processes, including the four statements, emphasize the importance of maintaining awareness of the dynamically changing information landscape. By providing such information at hand, users will be able to make a comprehensive evaluation of the current progress with confidence each moment. Awareness of peer activities is also important for assessing contributions of oneself in relation to collaborators’ contributions. On the one hand, providing information about how individuals extract nuggets from documents and how these nuggets are assigned to different themes could help examine their contributions from an overview perspective.
On the other hand, being aware of what collaborators are currently working on can provide information on the focus of attention. All this information is beneficial for the communication of the progress and the coordination of further collaboration. Therefore, to support such awareness, the information landscape and task structure need to be represented in a more accessible way, e.g., through interactive visualization.

Inspired by both the features advocated by previous studies [Hearst, 2014, Tao and Tombros, 2014b] and the findings described in the formative study, collaborative nugget extraction calls for support to increase the awareness of the dynamically changing information environment. Making the information landscape and the task status immediately accessible enables collaborators to adaptively adjust the strategy of nugget extraction. As users explore a collection of documents, the information landscape is changing under the influence of collaborative foraging activities, and the information of the distribution of extracted nuggets within documents and how these nuggets contribute to different scheme entries should be better represented.

Specifically, based on the derived evaluation criteria, the following information should be better represented: the distribution of extracted nuggets ($ST1 \rightarrow D1$), the reading frequency of content ($ST2 \rightarrow D2$), the trend of nugget extraction activities ($ST3 \rightarrow D3$), and the time spent on each part of documents ($ST4 \rightarrow D4$).

My goal is to design appropriate interaction to make the required information more accessible at the right time. The design implications described above motivated the design and implementation of NuggetLens, a visual analytical tool that supports collaborative nugget extraction in IBKC by providing a better representation of information seeking activities. The design rationale and concrete implementation of NuggetLens, and a user study using NuggetLens, will be described in the next chapter.
Chapter 4  Designing Visual Analytic Support for Collaborative Nugget Extraction in IBKC

The previous chapter showed that collaboration was needed as the task became complicated, and users had difficulties identifying relevant information nuggets and evaluating the task process. It also illustrated the design implications for supporting nugget extraction in IBKC which were derived from observations of how users make judgments and take actions during the process.

This chapter further investigates the design issues of supporting nugget extraction from a practical perspective. I start with a review of technologies and tools developed to support CIS. These works, along with the insights gained from the formative study, provide a guideline to set design goals and in turn enumerate certain design choices. I designed a visual analytical tool, NuggetLens, following this guideline. A user study was conducted to demonstrate the usefulness of NuggetLens, as well as to identify the problems that require further attention. The design process of NuggetLens serves as a vehicle that helps us learn about the design issues of tool support for nugget extraction.
4.1 Systems for Supporting Collaborative Information Seeking (CIS)

Much Research has been conducted on the design of tools dedicated to support collaborative information foraging. This section reviews these tools, including CoSense [Paul and Morris, 2009a], SearchTogether [Morris and Horvitz, 2007], CollabSearch [Yue et al., 2012], Cerchiamo [Golovchinsky et al., 2008], Querium [Golovchinsky et al., 2012], MUSE [Krishnappa, 2005], Coagmento [Shah, 2010b], Footprint [Wexelblat and Maes, 1999], MrTaggy [Kammerer et al., 2009] and Dogear [Millen et al., 2006]. Although the designs focus on a range of aspects of CIS, the tools allow participants to share content and workspace, see collaborators’ activities, and communicate in context.

Many tools are designed for web-search tasks. In a study aimed at designing interfaces for multiple users to perform investigational web search tasks collaboratively, Paul and Morris argued that a group of collaborative information seekers should make sense of not only products of searching results, but also the process [Paul and Morris, 2009a]. As such, they developed a tool called CoSense that focuses on the representation of activities about a given search session and concluded that CoSense can help handoff of sensemaking between asynchronous information seekers. In the same vein, Morris and Horvitz designed a collaborative web searching prototype: SearchTogether [Morris and Horvitz, 2007]. Through analyzing an example usage scenario, they demonstrated that SearchTogether’s collaborative features, including increasing awareness, dividing labor, and persisting knowledge repository, can help avoid undesired duplication of work and assist novice searchers by making MrTaggy experts’ searching behavior accessible. Other systems for supporting collaborative exploratory online search were also developed. These include CollabSearch [Yue et al., 2012], Cerchiamo [Golovchinsky et al., 2008], and Querium [Golovchinsky et al., 2012].
Coagmento [Shah, 2010b] supports interactive and collaborative information seeking process in various scenarios in addition to web-search, e.g., collaborative writing using found information. It is a browser add-on that provides a workspace in which users can interact with collected information, including organizing, sharing, and visualizing for better sensemaking and reusing of information. Coagmento was later extended to incorporate new methods for capturing users’ action (activity history) in online search to which collaborators can refer [González-Ibáñez and Shah, 2011]. Shah and his team have been currently working on extending Coagmento for better support collaboration in information-intensive tasks [Mitsui and Shah, 2016]. During the process, Coagmento also serves as a data collection tool.

Tools have also been developed to support information seeking in the academic setting. Krishnappa [Krishnappa, 2005] designed a tool called Multi-User Search Engine (MUSE) to support collaboration in information retrieval in information systems. The interface of MUSE includes search, share and chat components, allowing an individual user to search independently, share search results and chat with other collaborators. A preliminary evaluation of MUSE using a biomedical article database, PubMed, identified the prominent role of chat in supporting collaboration within information seeking activities, which leads to better search results through enhancing better understanding of group search process.

Information seeking behavior itself could also be used for facilitating information seeking. Researchers have designed several systems to utilize social activities and relationships to prompt users to seek information, including Footprint [Wexelblat and Maes, 1999], MrTaggy [Kammerer et al., 2009] and Dogear [Millen et al., 2006]. These systems keep track of users’ activities and their social relationships, and this information serves as clues to which users can refer for more effective and efficient information seeking. Some systems, such as [Wexelblat and Maes, 1999], transform interaction history into navigation tool based on the idea of history-enriched digital objects. Other systems, such as MrTaggy [Kammerer et al.,
2009] and Dogear [Millen et al., 2006], utilize social bookmarking and tagging for information retrieval and information discovery.

Nevertheless, researchers believe that tools dedicated to CIS are still far from success. Collaborative work mandate making sense of not only the “product”, but also the “process” [Paul and Morris, 2009b]. Therefore, an ideal tool is expected to also help make sense of group activities, understand the group dynamics, and improve group performances [Dourish and Bellotti, 1992], or in another word, support activity awareness. In short, a number of features are still missing from existing tools and systems identified by Hearst [Hearst, 2014] and Tao [Tao and Tombros, 2014a] are listed below:

- Bring awareness of what has already been gathered, viewed, and gathered by anyone in the group, enabling easy comparison of them to avoid duplicate and redundant work by providing a clear depiction of the landscape covered and the landscape yet to be explored;

- Support the tracking of knowledge development and corresponding activities, the enhancement of human cognitive capability in awareness interpretation, and the awareness of both produces and processes;

- Offer the analyst a sense of accomplishment, a feeling of walking over the information landscape and have their bearing on the terrain as they work, enabling them to strategize about the allocation of work based on their understanding of the information environment;

- Provide efficient ways of organizing, structuring and prioritizing collected information to alleviate people’ internal cognitive burden.

While NuggetLens introduced in this chapter is designed for a specific collaborative information seeking activity, the design implications derived from the aforementioned formative study are actually consistent with the desired features
for designing systems to support CIS. Specifically, researchers call for tools that can: support awareness of shared workspace; help allocate work; keep track of task progress; and provide flexible ways of organizing information products. NuggetLens is designed to support these features with combinations of various visualization components that represent the distribution of extracted nuggets (D1), the reading frequency of content (D2), the trend of nugget extraction activities (D3), and the time spent on each part of documents (D4).

4.2 System Design

We designed a visual analytic tool, NuggetLens, to enhance individual and collaborative work when CIR panel conducts nugget extraction (Phase 1 of CIR). Visualization is adopted in this study for its potential as external cognition in distributed cognitive tasks. It can offload part of the cognitive processes in terms of collaborative activities from the heads of individuals to external artifacts [Scaife and Rogers, 1996], enabling users to make judgments more quickly, effectively, and accurately.

In this section, we firstly describe the design rationales derived from both prior literature and the design implications gained from the formative study (section 3.3). We present design choices for each goal and justify them by considering design space, constraints, and alternative techniques.

4.2.1 Design Rationales

Previous section identifies the information related to user activities that need to be reasonably represented (D1 - D4) in order to support the evaluation in nugget extraction. It leads to a set of candidate design choices. In this section, we discuss the design rationales behind current design of our tool and introduce the design choices we made which are aligned with the high-level design goals.
Figure 4.1: User interface with NuggetLens enabled (used in the 33-participant user study)
In order for users to be able to make sense of a changing information environment quickly and easily, the information environment should be represented in an abstract fashion while still can deliver the most valuable information that can enable users to make choices concerning foraging strategies. In our case, in addition to the documents provided at the beginning, the information environment of CIR also include the changes made by the users. As the users work on these documents, each patch appears to vary depending on the number and distribution of nuggets (D1) within it, the amount of attention it draws (D2), and the categories of themes associated with it (D4). As a result, in order to achieve the design goal, the reciprocal influence exerted on the original collection of documents by the users should be represented as well.

One way of supporting such representation is to visualize users’ important activities and incorporate the visualization into scroll bars. An attribute-mapped scroll bar designed by Hill et al. shows a history of use by using marks to portray how frequently sections of the document have been read and where parts of the document have been modified, named Edit Wear and Read Wear respectively [Hill et al., 1992]. They claimed that Edit Wear and Read Wear can better facilitate small-group collaborative work compared with group process control through supporting effective and efficient information exchange. However, by now, Read Wear and similar designs are still used in an individual fashion [Alexander et al., 2009, Atterer and Lorenzi, 2008].

Aggregating those attribute-mapped bars provides an overview of group activities. One example is Seesoft, which is a visualization tool that maps each line of code into a thin row in a vertical bar that represents a file [Eick et al., 1992]. Different colors are used to encode each line depending on statistic associated with the line of code in this reduced representation. The key concepts of Seesoft include providing an overview using reduced representation, using color to encode attributes by position, and allowing users to directly interact with the representation.
Aforementioned work represents one aspect of a dataset, while we extended their ideas by positioning two attributes of activities side by side since comparing the distributions of allocated time and completed work can deliver richer information for making judgments. In document view, only one column concerning current reader is shown. It functions as a scroll bar while also providing some more information about individual work. In DocuMap view (Heatmap Bars), a collection of columns that represents all documents are shown and they are used to show group activities and collaborators’ attention focuses by providing at-a-glance overview. We also add the ability of data filtering which allows an individual to examine activities under different constraints.

In addition to information environments, being aware of task environments is also important. It can benefit collaboration in nugget extraction by revealing the group effort on workspaces (D4) to help coordinate tasks and communicate findings. The task environment in CIR nugget extraction involves three entities, documents, panelists, and themes. These entities are connected through activities of nugget extraction in each a panelist extracts a nugget belongs to a theme from a document. Nugget extraction activities are multidimensional and categorical. In order to make the connections between those entities more accessible, Parallel Sets [Kosara et al., 2006] (Sankey Diagram) is adopted in this study. There exist other alternatives for visualization multivariate categorical data, e.g., the mosaic plot (or Marimekko chart) and parallel coordinates [Unwin et al., 2006]. The former is more suitable for spotting small disparities for its laying out subsets side-by-side, while the latter for multivariate ordinal data.

For representing the trend of nugget extraction activities (D3), timeline-based visualization fits for its capability of revealing temporal pattern of a collaborative task. For example, Classroom BRIDGE is designed to support distributed group projects among different classrooms where students rarely meet in person [Ganoe et al., 2003]. The tool is a resulting product of a long-term participatory study in
which a group of public school teachers is involved in the tool development process. They choose timeline-based coordinated visualizations on as the primary user interface since they consider it could better facilitate planning and activity awareness by integrating as much activity information as possible. Another visual analytics tool, VAET, helps analysts to explore time-of-saliency by selecting transactions through specifying time interval using time window [Xie et al., 2014]. Classroom BRIDGE demonstrates the importance of temporal dimension of activities while VAET utilizes time window as a filter for detail exploration. As described in [Eick and Karr, 2002], this technique could increase scalability through interactivity. Therefore, zoomable bar chart is used to depict the count of nugget extraction activities (D3) over time. Applying filters allows users to examine the data generated by particular users or relevant to specific documents. The bar chart itself also serves as a filter to specify a time period.

Similar to VarifocalReader [Koch et al., 2014], a multi-layer visualization is embedded in our tool to allow panelists to explore and navigate through documents in a hierarchical fashion (from overview to details). The Parallel Sets plot provides an overview of all documents by aggregating the number of extracted nuggets organized by either authors or themes. Then heatmap bars are used to provide a more detailed overview by displaying the location of extracted nuggets and the distribution of frequently viewed contents. When hovering over a document, either the node in the Parallel Sets plot or heatmap bar, the bar chart is updated for the single document and a word cloud is shown as a preview of the document. By clicking the document, the contents of the document are retrieved. Besides, nugget list allows panelists to share their work with each other. We will describe each component in more details in the following sections.
4.2.2 Revised Nugget Extraction Workspace

Figure 4.1 shows the user interface of the current system. The current version was designed and implemented by improving the nugget extraction workspace deployed in the formative study and incorporating a new visual dashboard called NuggetLens. In this section, we describe the new component added to the previous nugget extraction workspace, a heatmap-based scroll bar, as shown in Figure 4.1b. The red button located at the top-right corner is for evoking NuggetLens on demand. A more detailed description of NuggetLens is provided in the next section.

4.2.2.1 Heatmap-based Scroll Bar

Heatmap represents documents as columns. Extracted nuggets are represented as rows within the columns. It is on the edge in the document view. It can also function as a traditional scroll bar in which a blue arrow indicates the currently viewed part of the document. Usually, the lengths of documents are longer than the height of browser window height thus scrolling will be necessary to go through all contents. The height of the bars corresponds to the length of the entire document.

The heatmap bar is beyond a traditional scroll bar by also incorporating two types of heatmaps. They subdivide the heatmap bar vertically into two parts and provide powerful capabilities that can facilitate document reading. The larger left part shows a heatmap of the distribution of extracted nuggets. As a nugget extraction activity occurs, relevant information is collected and saved. The heatmap will be updated correspondingly.

The extracted nuggets within a document are represented using horizontal stripes of different heights and colors. A stripe is placed according to its corresponding nugget’s vertical offset within a document view. Each horizontal stripe corresponds to one extracted nugget and the height of the stripe depends on the length of the nugget. Light blue is used to indicate the length of a document and serves as
the background while blue indicates extracted nuggets in terms of their relative positions and lengths.

The smaller, right part shows a heatmap that visualizes how individual sense-making time is distributed across each part of a document. This heatmap is initially white while a document has not been visited before. As a user starts and continues to interact with the document by reading certain parts, scrolling around, performing annotations, and extracting nuggets from it, the amount of time spent on different parts of the document are collected and the heatmap is updated in real time correspondingly. The more time spent on a certain part of a document, the corresponding part within the heatmap will become darker, until red. A threshold time can be set for the heatmap according to specific needs introduced by different tasks. The time counter will pause once a user stops interacting with the document. The inactivity is recognized through monitoring mouse movement within the document view.

The hybrid heatmap allows a user to keep being aware of where extracted nuggets are located and how time is spent within the currently viewed document. It serves as the basic component for composing DocuMap that is designed to support group awareness of the information environment.

4.2.3 Visualization Components

NuggetLens provides multiple features to prompt panelists to tackle some difficulties they encounter and gain more insights into user interaction more quickly, confidently and reliably. The interface of NuggetLens is shown in Figure 4.1a. It contains DocuMap consisting of a set of heatmap bars, a Parallel Sets plot, and a document preview panel. Filters and a zoomable bar chart are placed at the top. DocuMap is designed to provide an overview of the information landscape while also provides enough detailed information when interacting. The Parallel Sets plot illustrates the task structure through nugget extraction activities. Filters and zoomable bar chart
enable users to explore both the information and task environments from various dimensions.

4.2.3.1 DocuMap

In order for users to make sense of a changing information environment quickly and easily, in addition to the documents provided at the beginning, the information environment also include the changes made by the users. As the users work on these documents, each patch appears to vary depending on the number and distribution of nuggets within it, the amount of attention it draws, and the categories of themes associated with it. As a result, the reciprocal influence exerted on the original collection of documents by the users should be represented as well.

![Figure 4.2: DocuMap](image)

While previous work represents one aspect of a dataset, we extended their ideas by positioning two attributes of activities side by side since comparing the distributions of allocated time and completed work can deliver richer information for making judgments. Specifically, the horizontal stripe corresponding to each document represents the distribution of attention and extracted nuggets respectively, allowing a user to keep being aware of where extracted nuggets are located and how time is spent within the currently viewed document. It serves as the basic component for composing DocuMap that is designed to support ST1 and ST2.
4.2.3.2 Parallel Sets

In addition to information environments, being aware of task environments can benefit collaboration in nugget extraction by providing necessary evidences for coordinating tasks and communicate findings. The task environment in CIR nugget extraction (ST4) involves three entities, documents, panelists, and themes. These entities are connected using nugget extraction activities: an activity is described as a panelist extracts a nugget from a document and the nugget is assigned to a theme. Nugget extraction activities are multidimensional and categorical. Parallel Sets [Kosara et al., 2006] is adopted in this study for its capability of visualizing categorical data [Kosara et al., 2006]. The Parallel Sets plot in NuggetLens supports activity overview, flow tracing and detail on demand through cross-linking among sets of categories.

Figure 4.3: Parallel Sets plot

In the Parallel Sets plot, each node represents an entity appear in the information nugget extraction task. We use different colors to encode different categories of entities. Green is used for encoding documents, red for panelists and yellow for
themes. The number of nuggets determines the width of the nodes and edges. The category tabs within filters allow users to choose the categories of nodes that interest them most.

4.2.3.3 Zoomable Bar Chart

The zoomable bar chart shows the trend of the number of extracted nuggets in an user-adjustable time interval (ST3). The smaller bar chart shows the overview of the number of nuggets and a user can select a specific time range through click-and-drag action within the overview bar chart. After a time window is set, the bigger bar chart will only show the detailed number of extracted nuggets within that time range. the Parallel Sets plot and DocuMap will also be updated accordingly.

4.2.4 Multiple Coordinated Views

When NuggetLens is launched, DocuMap immediately provides a global view of group activities by showing nugget distribution and time spent for distinct segments of documents, as well as peers’ latest focused region within documents. The Parallel Sets plot connects documents, panelists and themes using nugget extraction activities at the moment. Bar chart gives an overview of the numeric change of extracted nuggets.
Filters consist of three lines of toggles that map to three categories of entities (document, panelist, and theme). Users can choose entities they are interested in by toggling. As described in section 4.2.3.3, the zoomable bar chart also functions partially as a time-windows selector, which can be used to specify a specific region of interest for detailed examination.

The Parallel Sets plot supports flow tracking, allowing users to follow nugget flows across the graph. By clicking a node, all relevant edges and nodes are highlighted. For example, in Figure 4.3, a panelist node is chosen. As a result, one document that the panelist has worked on and three themes that the panelist has contributed to are highlighted.

The multiple views are coordinated. When a document node within the Parallel Sets plot is clicked, the corresponding bar with DocuMap is highlighted, and vice versa. After clicking other categories of nodes in the Parallel Sets plot, the bar chart will be updated by showing only activities related to the selected entity. If a user finds a part of documents and clicks that part within a heatmap bar, the document will be opened in a preview panel with word cloud and is directly scrolled to the corresponding region.

### 4.2.5 Supporting Decision-making in Nugget Extraction

This section demonstrates the functionalities of NuggetLens using four scenarios of use. The major functionality of the system and the user-system interaction are illustrated on three aspects: (1) supporting nugget extraction based on collaborative sensemaking of the available documents; (2) exploring aggregated activities for task coordination; and (3) making sense of the information landscape to evaluate task progress. For illustrative purposes, let’s call the users Alice, Bob, Colin and they work as a group. Their actions are composites based on the observed behavioral traces of real users from the aforementioned formative study.

During the process, there are several time points that users have to evaluate
the information foraging progress and make decision about further actions. Four representative screenshots of NuggetLens are attached to illustrate how they are capable of supporting completeness assessment and decision making. These screenshots also demonstrate the change of patterns of collaborative information foraging activities over time. Additional details about the system are described in the end.

4.2.5.1 Selecting Information Nuggets based on Collaborative Sense-making

Before touching the system, Alice has participated a tutorial session where she learned basic knowledge about the community issue Collegiate Housing Overlay (CHO), and she agreed on the schema (themes) used for structuring information nuggets. She is willing to help produce more accessible information through extracting nuggets using the system.

When Alice arrives at the system, she clicks on the icon of “collegiate housing overlay” among multiple community issues. The link directs her to the corresponding nugget extraction workspace. Alice gets a general sense of information landscape by reading the table of contents and skimming some documents quickly. She feels that it is difficult for her to go through all the documents carefully in a limited amount of time. Therefore, Alice opens the chat room and discusses with the other two collaborators, Bob and Colin. They decide to start with the documents with obvious positions (pro, con, or neutral) respectively. Specifically, Alice begins with the documents that appear to support the issue.

Alice starts with a document that provides a general introduction of the points that support CHO. She notices that a sentence “It encourages the highest and most efficient use of a limited land resource, and supports recommendations from the Downtown State College Master Plan to add value to the community overall” can help explain the benefits of supporting the issue in terms of the “value”. Thus she includes the contents in the nugget list through highlighting the corresponding
texts and attaching the theme “value” to the information nuggets (Figure 4.5). She notices that the newly extracted nugget appears immediately on the top of the nugget list.

As she learns more about the issue, Alice realizes that some nuggets extracted earlier might not be informative enough that can be replaced by more solid and representative ones encountered later. As a result, Alice removes the nugget extracted earlier from the nugget list and adds the new one.

Alice continues adding information nuggets and modifying earlier ones until she thinks all pro contents have been viewed. In order to increase her confidence of making such judgment, she reviews her finished work with the help of the scrollbar. She is interested in locations where few nuggets extracted and/or little time spent since they indicate the opportunity of finding more valuable information.

After completing her assigned job, Alice then browses through the neutral documents, which have been touched by Bob. She notices one nugget is more appropriate than the one extracted by Bob. She adds it to the nugget list and makes a comment about Bob’s nugget.
4.2.5.2 Making Sense of Activities for Task Coordinations

Thus far we have described how the system enables nugget extraction through both individual and collaborative sensemaking. In this section, we describe the system components that facilitate group coordination by promoting activity awareness.

After completing her assigned job, Alice asks, in the chat room, whether others need any help in the chat room. Bob replies that he has to leave and he asks Alice to take over the job he is currently working. Alice accepts Bob’s request. She starts by opening NuggetLens and applying filters to check Bob’s work. Resulting DocuMap shows that Bob has spent a lot of time on the documents he has read. He already covers much information space, and the distribution of the extracted nuggets is intensive. By referring to the NuggetLens, Alice becomes familiar with the contents touched by Bob. She also finds that most of the nuggets already extracted by Bob question whether the proposal actions can achieve CHO’s intended purpose. Parallel Sets confirms her impression, showing that not much the work by Bob is concerned with “safety”. Since the “safety” has been identified as one of the important themes (on the first-day meeting) that need to be addressed and elaborated, Alice starts to take over Bob’s work but adjust her strategies to focus more on contents related to the theme.

4.2.5.3 Depicting Information Landscape for Evaluation

Alice feels that the task is complete since important contents appear to be extracted as she randomly browses through documents. She interacts with NuggetLens in order to confirm her judgment. The bar chart shows that the number of extracted nuggets no longer increases. The Parallel Sets shows that all themes have been explored. Alice notices that Colin spends a lot of time on some paragraphs but few nuggets are extracted. Colin then explains in the chat room that the document provides valuable thoughts for understanding the issue but is not directly relevant. After discussion, the group agrees that the nugget extraction task is completed.
4.2.5.4 Representative Decision Points

Figure 4.6 depicts four representative examples that best demonstrate the NuggetLens usage scenarios. At the beginning of the task, the group employs a divide-and-conquer strategy (See Figure 4.6a). Each of them starts with a different document. With the assistance of NuggetLens, users can have a big picture of initial task arrangement and collaborators’ progress and working styles, e.g., P77 reads slowly while P67 reads fast.

Figure 4.6b shows P65 has already extracted nuggets from two documents while other panelists are still working on initial documents. DocuMap also provides him with information about what documents have not been looked at yet from which P65 will select one of them to continue.

By the moment shown in Figure 4.6c, All documents have been touched. However, DocuMap shows that there is one document without any extracted nuggets. Also, according to the Parallel Sets plot, there are few extracted nuggets belong to
“safety”. Therefore, panelists decide to review their work and attempt to examine whether something is omitted. On the other hand, P65 finds that there is a place within a document that receives a lot of attention but has only one nugget extracted. He thinks it is because the content is relevant but confusing (difficult to extract) and may need another look. He thus continues by examining that part carefully, as shown in Figure 4.6d.

4.3 User Study

The goal of this user study is to observe how participants use NuggetLens to collaboratively seeking information, interpret indicators revealed in the formative study, and develop an awareness of activities. This helps us understand the different means participants approach a complete set of information, the criteria they use for decision making, and how the versatile features provided by NuggetLens cater to these needs. These understandings lead to some design implications.

4.3.1 Study Setup

This user study is an extension of the previous formative study (see Section 3.3). Our research team recruited 33 undergraduate students as the participants. These students major in Information Sciences and Technology at Penn State University and they were enrolled in visual analytics course Spring 2016. At the time of study, they were in the middle of the semester and were considered equipped with basic visual analytic abilities. The user study includes two sessions with almost the same setting. There are 22 persons involved in the first session and 11 in the second. The second session took longer time than the first since it included more themes and thus increases the complexity of the issue. The issue we chose for this user study is “Collegiate Housing Overlay (CHO)”, as introduced in section 3.3.1.
4.3.2 Procedure

The participants are organized into 6 groups. Each group contains either 5 or 6 participants. For each group, a short introduction and a training session were conducted before performing tasks. In the first session, the instructor introduced the community issue review in general and particularly described the nugget extraction task in terms of its concepts, process, expected outcome and role in the context of CIR. Then CHO was introduced, as well as the themes that help organize the issue.

In the training session, the instructor (the lecturer of the course) firstly explained to the participants the user interface, visual design and supported functions of NuggetLens. After that, the participants were given about 10 minutes to interact with NuggetLens using sample data with the help of the instructor.

Then, participants organized in groups were asked to read documents and extract information nuggets. Although they were collocated in the same room, they were required to use a chat room if they need to communicate (as they were in a distributed situation). During the process, we logged participants’ behaviors through observation, and uses of visualization (open/close actions). We also analyzed recorded chat messages in order to find out how NuggetLens support their activity awareness and group coordination.

During the working session, two researchers observed participants’ activities and tried to identify interesting actions, interactions, and behavior. Their observations could help us confirm the assumptions based on which NuggetLens was designed, and allow participants to reflect on their work in the later interview phase.

4.3.3 Data Analysis

We ended up with 408 chat messages and hundreds of records of participants’ interactions with our system with the 6 teams that participated in our study. The chat messages were coded by me following the coding scheme adapted from another
study that performs content analysis on communications within collaborative information seeking tasks [González-Ibáñez et al., 2013]. The coding scheme includes five main categories - task coordination (TC), task content (TN), task social (TS), non-task (NT), and non-codable (NC). Within each group, we synchronized the chat messages and the interaction with NuggetLens for each participant in order to demonstrate how various views of NuggetLens provide explicit evidence to which participants can refer for communication. While chat messages and system logs provided valuable insight regarding the usage of NuggetLens, they can only provide partial information which is not reliable enough. Therefore, we also conducted post-task interview that serves as a complementary method to verify the usage of NuggetLens by providing more detailed explanations to interpret the participants’ intentions behind actions, e.g., whether they are trying to minimize the cost of moving between documents, or looking for opportunities with more nuggets potentially (see Appendix A for the detailed questionnaire).

4.3.4 Results

User interactions with NuggetLens recorded are shown in Figure 4.7. The chart displays the amount of time (in seconds) spent by each participant on interacting with NuggetLens over time. 4 out of 22 participants in the first session and 1 out of 11 participants in the second did not use NuggetLens at all. According to the log data, participants who had records of using NuggetLens referred to NuggetLens all the time. As the task moved towards completion, they tend to spend more time with NuggetLens.

As only open and close events (a user opens or closes the NuggetLens window) were logged due to system limitation, chat messages serve as a complementary source for analyzing the interaction with NuggetLens. As chat logs suggest, the need for coordination motivated them most to refer to NuggetLens. Near the end of the time, they would like to ensure no important information is missing by review
each others’ work with the help of NuggetLens. The complexity of data by this moment and the pressure of finalizing their outcomes encouraged them to spend more time and review more carefully.

In order to confirm the hypotheses of evaluation metrics (ST1 - ST4), we also conducted a survey that asks their opinions whether they agree or disagree on the above hypotheses. Their opinions were elicited using 5-point Likert scales (1=strongly disagree; 5=strongly agree). The ratings by the participants are shown in Figure 4.8.

Overall, participants are positive on the usefulness of these indicators. The statement ST1 received an average rating of 4.1/5.0, indicating that nugget coverage was believed to be a strong indicator of completeness. As suggested by ST2, a combination of nugget coverage and view frequency could also serve as an indicator. Interestingly, according to our logged data in the user study, before all group members came to a consensus regarding task completeness, the intensity of activities
decrease apparently. However, participants did not consider it as a very strong indicator. The distributions of themes are considered as a weaker indicator with a mean score of 3.8/5.0.

As to the interview, generally, participants’ opinions were very positive. DocuMap is considered to be most useful since it provides information about the density and coverage of nuggets within each document and serves as strong evidence for evaluating task progress and locating potential unfinished work.

“One of the best places to look is at the left-hand side in relation to the screen. We notice the density and amount of coverage completed in each individual document. If there is a document that seems to have less coverage, we can assess that the team isn’t quite done and will have to go back and check the untouched / barely touched documents.”

The Parallel Sets plot is prompt in answering uncertainty of task progress by providing information scent about which (parts of) documents need more attention and effort.
“when you click on a column, you can see how many people have worked on the specific topic. This also allows you to see what topics need more attention”

Showing panelists’ latest activities using salient icons can prompt users to quickly evaluate group task progress by examining each individual’s statuses. This could help build common ground and thus facilitate participants to coordinate their work faster and more efficiently.

“The little people figures [are useful] because they stand out and can easily count to see if all my team members are finished. It provides a good platform to streamline teamwork.”

By visualizing information evolution and peer activities and allowing users to interact with the data visually, NuggetLens enables a user to easily share their work with other group members.

“The ability to collaborate and share work with members on my team made the entire process much easier. Our collaboration was expedited due to the ease of use of the tool.”

The user study confirmed the indicators (D1 - D4) participants adopt for performing collaborative information seeking drawn from the formative study, and demonstrated the usefulness of NuggetLens as a visual analytics support for performing nugget extraction collaboratively in CIR. However, a more structured quantitative study with real CIR panel is necessary to evaluate the impact the tool has. As we understand more about this process and have more sophisticated designs, we are able to statistically compare the task completion time and accuracy in various settings.
4.4 Summary and Discussion

Currently, CIR is difficult because of the messiness of input data and the complexity of socially embedded community issues requires human intelligence rather than any automatic approaches to analyze documents and produce statements. We made the first step towards a better understanding of information foraging activities happening in the context of the CIR process.

However, the selection of human subjects in this work brought certain limitations and we could not achieve a realistic simulation. Although we tried our best to model the work conditions, since the participated students were not stakeholders in the issue, their incentives were different from those of relevant citizens.

There are several directions we can take as the next step. First of all, participants were asked to come up with a set of themes initially. The set of themes is important as a guideline for seeking information nuggets and to organize collected information nuggets. Therefore, these themes should be generated by a more reliable mechanism, rather than being generated quickly when participants have little knowledge of a given document collection. Second, although we assumed users are able to recognize information nuggets, it turned out to be challenging for them to do so effectively. Therefore, technological support, e.g., nugget recommendation, might be useful to support nugget recognition. Third, this work concentrates on the foraging of information while the use of information is also important. The extracted nuggets are organized as a list that can be categorized by themes. More flexible ways of structuring and representing them can potentially improve further use.

The next chapter addresses the first issue of themes with an approach that can enable participants to have a structural understanding of a given document collection quickly with an interface powered by interactive topic modeling.
Chapter 5  |  Designing Interactive Topic Modeling Interface for Knowing about an Information Landscape

Information foragers engage in sensemaking over the course of information seeking. However, sensemaking before formally starting foraging information is especially critical because it helps people construct an initial mental representation of interrelated patches of information relevant to achieving a certain goal based on available data. The usage pattern of NuggetLens shown in Figure 4.7 suggests that a rough depiction of the information space could be useful to cold start information foraging at the beginning of information seeking when there are not sufficient traces of collaborative information foragers. Inspired by information foraging theory [Pirolli and Card, 1999], especially the Scatter/Gather browsing technique [Cutting et al., 2017, Pirolli et al., 1996], this chapter discusses the role interactive topic modeling can play in schema induction by helping users to learn about an information landscape. The induced schema could serve as the initial set of themes for boosting nugget extraction. In particular, an interactive topic modeling tool called Fast Information Landscape Analyzer (FILA) was designed to support making sense of petitions, serving as a pre-step for information seeking.

Although CIR is the background of this dissertation, we start with analyzing
petitions rather than community issues here. There are some similarities between community issues and petitions, for example, the role of both of them is connecting ordinary citizens and decision makers through better information expression and delivery. But they also have certain differences, for example, the topic of each petition is more obvious and the petition data set is more diversified, while articles about a community issue are usually more concentrated and the sources and types of those articles are more comprehensive. In addition, there have been some studies on the application of topic modeling techniques in petition analysis [Hagen et al., 2015a, 2016]. As a result, in this chapter, we start with petitions as an application scenario to introduce the design and implementation of FILA. Later in the chapter, we will demonstrate experimentally that the approach can be applied to community issues as well.

5.1 Related Work

5.1.1 Characterizing Petition Data and Information Needs

Petitions are requests towards authority about an issue and they are usually signed by many supporters. The dataset used in this study comes from the White House petition platform “We the People” and is considered as a unique source for understanding citizens’ policy concerns and preferences [Paul Hitlin, 2016b]. This section focuses on understanding the specific challenges of analyzing online petition data.

Citizen involvement in democratic deliberative processes is necessary to create, assess, and prioritize policy proposals for the public good. Although the importance of civic engagement is widely recognized, many existing forms of public participation, such as social voting and e-rulemaking, are initiated and/or facilitated by decision makers, professional facilitators, or researchers, other than citizens themselves [Hagen et al., 2016]. On the other hand, petitioning aims to forward
citizen’ policy suggestions and requests to relevant authorities and they are able to represent the collective agreement by gathering a number of signatures [Fischer and Forester, 1993]. With the popularity of online technology, e-petitioning is becoming more common and many online petitioning platforms emerge, including both government-sponsored electronic petitioning systems (e.g., “We the People”) and private organizations (e.g., Moveon.org, change.org, gopetition.com, avaaz.org). In addition to the United States, governments of England, Germany, Scotland, Australia, and Norway are taking advantage of online petitioning systems to facilitate decision-making process [Hagen et al., 2016] as well. Compared with traditional petitioning, online petitioning platforms are more accessible, effective, and direct, and thus narrow the distance between citizens and decision-makers: technology enables citizens to directly express their opinions and ideas to authority, and social media venues further spread their voices and concerns. As a result, nowadays the most popular way of participating in political activity online is signing petitions [Aaron Smith, 2013, Dutton et al., 2005].

However, due to the popularity of e-petitioning, a large amount of unstructured textual data is created that are difficult to analyze: the document collection is too large that requires an effective way of structuring and indexing; an effective mechanism is desired to read, make sense of, and summarize these large texts. Valuable insights can be gained if we can effectively analyze these data, and many useful applications can be created if we can effectively take advantage of these data.

Many computational tools have been developed to deal with “information overload” problem. For examples, [Koch et al., 2014] visualizes large and complex documents by representing them at different levels of abstractions, which supports an in-depth visual analysis and exploration of the hierarchy of documents. Similarly, [Dou et al., 2013] enables exploring and analyzing large text corpora by representing document collection using hierarchically topics learned from the corpus. These techniques represent texts at varying granularity and support hierar-
chical analysis. [Aggarwal and Zhai, 2012], on the other hand, utilizes document clustering algorithms to derive important information from the data. [Lee et al., 2012a] also uses topic model-based clustering technique to help make sense of large document collection. In e-Government field, “mini-public” is proposed as a solution to “information overload” problem by using a citizen panel to analyze data relevant to a pending issue (sometimes with technological support) and presenting the general public with a crystallized version of the data [Cai, 2017, Sun and Cai, 2017]. In short, topic modeling, clustering, and visualization techniques are widely used (sometimes together) to support the analysis and sensemaking of big unstructured textual data by automatically uncovering latent patterns.

This work adopts topic modeling to support the analysis and sensemaking of petition data as the discovered topics can offer a structural summary of petitions through soft clustering. Topics, serve as a kind of information scent that provide a succinct representation of the content “whose trail leads to information of interest” [Pirolli, 1997]. This is especially the case when users are presented with documents without any trails at the beginning of information foraging tasks in CIR. With the help of topics, petitions can be represented as clusters of multi-dimensional data points for supporting the cold-start of information seeking, topically similar petitions can be recommended once an individual petition is specified, an overview that summarizes petition space becomes possible for better navigation, topically duplicated petitions can be discovered, hidden topics can emerge, outlier petitions can get some attention, and more. In terms of implementations, different from techniques such as information extraction approach NER whose performance is domain specific while also requires time-consuming manual annotation [Hagen et al., 2015a], topic modeling is unsupervised or semi-supervised that requires little manual work. These features make topic modeling a “perfect fit” for supporting the browsing task in information foraging.

Insights gained by practicing interactive topic modeling on the petition data,
on the one hand, can be used to inform the design of a visual analytic system that supports making sense of community issue based on topics; on the other hand, will benefit the topic modeling community in terms of the design of interactive topic modeling systems for non-experts users.

5.1.2 Information Visualization Approaches for Sensemaking

Many studies on supporting the sensemaking of large documents have been done using visualization powered by techniques other than topic modeling techniques. While techniques are different, the visualization design approaches can provide valuable insights. Some of these studies are reviewed before proposing interactive topic modeling as a solution for making sense of petitions for information seeking in IBKC.

Decision-making often involves making sense of data because of lacking knowledge. Sensemaking could help find patterns in unstructured data and might result in a more accessible representation of the data. Sensemaking can be conducted in either top-down or bottom-up manner. The top-down approach utilizes representation to organize the process of sensemaking [Russell et al., 1993] while the bottom-up approach derives structure from data [Bellinger et al., 2004]. Sensemaking sometimes takes place as an iterative process that involves both approaches [Pirolli, 2005, Pirolli and Card, 2005].

A number of visualization methods and system prototypes have been developed to support the sensemaking of collection of documents [Bradel et al., 2013, Brehmer et al., 2014, Luo et al., 2012, MacEachren et al., 2011, Proulx et al., 2006, Stasko et al., 2008, Yang et al., 2010]. Some of them support document understanding visually with various computational techniques (especially automatic text analysis technique) based on exploiting document content, e.g., topic modeling [Chuang et al., 2012, Lee et al., 2012b], word clouds [Heimerl et al., 2014, Viégas and Wattenberg,
2008, Wattenberg and Viégas, 2008], and syntactic parsing [Muralidharan et al., 2013]. Another class of visual analytics tools utilizes the embedded semantic structure to aid investigative analysis of document collection (e.g., entity-relation models [Cao et al., 2010, Görg et al., 2013, Uren et al., 2006]).

5.1.3 Interactive Topic Modeling

The topic-based analysis of petitions is expected to benefit both citizens and decision makers who are not topic modeling experts. Therefore, there should be an approach that enables non-expert users to incorporate their knowledge into the topic modeling loop. Nevertheless, currently interactive topic modeling techniques are designed mostly for machine learning experts [Lee et al., 2017]. This section starts with a review of existing topic modeling techniques, followed by interactive topic modeling tools that attempt to increase human involvement (mainly experts). By analyzing the advantages and disadvantages of these tools and approaches, we have come up with some design implications for developing interactive topic modeling for non-expert users.

5.1.3.1 Topic Modeling Techniques

One of the earliest topic models is LSI [Landauer et al., 1998], which (also known as Latent Semantic Analysis, LSA) derives latent topics by performing a matrix decomposition technique called SVD (singular value decomposition) on the term-document matrix. A probabilistic variant of LSI known as PLSA (probabilistic latent semantic analysis) was developed later to achieve better performance [Hofmann, 1999]. Non-negative Matrix Factorization (NMF), like PLSA, is also an instance of multinomial PCA and can be used for topic modeling [Gaussier and Goutte, 2005]. LDA (Latent Dirichlet allocation), as a variance of PLSA, was then developed with an assumption that topic distribution has a Dirichlet prior [Blei et al., 2003]. LDA a generalization of the PLSA model when a uniform Dirichlet prior is over the latent
topics while PLSA is a regularized maximum likelihood variant of LDA [Girolami and Kabán, 2003]. LDA involves priors and thus performs better on documents with short texts because it can give a reasonable guess to avoid data overfitting.

LDA has a number of variants. For example, by exploring the relations between topics, hierarchical topic models can be generated [Bakalov et al., 2012, Griffiths et al., 2004, Kataria et al., 2011, Li and McCallum, 2006, Mao et al., 2012, Nguyen et al., 2013, Petinot et al., 2011] to reveal the relations among topics. In another example, based on the fact that LDA is performed on unlabeled data and learned topics are sometimes not compatible with the ground truth, some researchers revised LDA in a supervised learning fashion [Lacoste-Julien et al., 2009, McAuliffe and Blei, 2008, Ramage et al., 2009b, Zhu et al., 2012] to incorporate known labels into the topic modeling process. Researchers also extend LDA by taking into account the attributes of documents and words (e.g., order), such as [Blei and Lafferty, 2006, Chang and Blei, 2010, Mimno and McCallum, 2012, Rosen-Zvi et al., 2004, Wallach, 2006].

There are also topic modeling techniques other than LDA, for example, Correlation Explanation (CorEx). CorEx is developed based on the theory of Correlation Explanation [Gallagher et al., 2017, Ver Steeg and Galstyan, 2014] in information science. CorEx represents the latent topics in a document collection in a way that maximizes the informativeness. Compared with LDA, CorEx can achieve comparable performance in terms of topic coherence for documents of short length. Most importantly, CorEx provides anchoring methods that can help incorporate human knowledge into the topic modeling process.

Topic modeling techniques have been applied in many applications: for instances, analyzing topic evolution [Malik et al., 2013, Wei et al., 2010]; recommending televisions [Pyo et al., 2015], scientific articles [Wang and Blei, 2011], or users of similar interests [Pennacchiotti and Gurumurthy, 2011]; understanding and summarizing online petitions [Hagen et al., 2015b]; characterizing social network
posts [Bernstein et al., 2010, Ramage et al., 2010, Resnik et al., 2015]; performing biological data clustering, classification, and feature extraction [Liu et al., 2016]; and conducting aspect mining [Mukherjee and Liu, 2012].

5.1.3.2 Human-in-the-loop Topic Modeling

Researchers have been developing tools that combine topic modeling techniques and interactive visualization to help users explore and analyze large collections of documents. For example, to make large collections of text more accessible, a visual exploratory text analytic system Tiara [Wei et al., 2010] derives time-sensitive keywords from topic-based document summarization to illustrate the content change over time. It also provides various interaction methods that allow users to visually make sense of the change of topics. Similarly, TopicFlow applies topic modeling techniques to produce an overview of twitter stream by visualizing the evolution of latent Twitter topics [Malik et al., 2013]. Such topic-based visualization allows users to analyze large time-sensitive social media data sets more effectively. HierarchicalTopics [Dou et al., 2013], different from the above two, supports the exploration of large document collections, especially the discovery of document structure, by uncovering the hierarchical structure of underlying topics. Topics help represent large collections of text through clustering of topically similar documents. For instance, TopicNets [Gretarsson et al., 2012] used a web-based interactive visual interface to enables users to discover topics of increasing granularity through an informed selection of relevant subsets of documents.

Topic modeling is not an out-of-the-box tool and usually requires careful selection and tuning of parameters. The qualities of discovered latent topics by LDA and other topic modeling algorithms vary greatly. As topic models are mostly unsupervised, users can either accept or leave the results. Many topics are considered of poor quality because: (1) they often confuse two or more themes into one topic; (2) they often pick up two different topics that are (nearly) duplicates for human;
and (3) nonsense topics [Hu et al., 2014], (4) topics with too many generic words (e.g., “people, like, mr”) [Boyd-Graber et al., 2014], (5) topics with disparate or poorly connected words [Mimno et al., 2011], (6) topics misaligned with human interpretation [Chuang et al., 2013], (7) irrelevant topics [Ramage et al., 2009a], (8) missing associations between topics and documents [Daumé, 2009], and (9) multiple similar topics [Boyd-Graber et al., 2014]. The presence of poor-quality topics has been cited as the primary obstacle to the acceptance of statistical topic models outside of the machine learning community [Mimno et al., 2011]. The root of these problems lies in the fact that the objective function that topic models optimize does not always correlate well with human judgments of topic quality [Chang et al., 2009]. Due to these problems, the use of topic models to analyze domain-specific texts often requires manual validation of the latent topics to ensure that they are meaningful [Hall et al., 2008].

To address the problem of poor-quality topics, a variety of human-in-the-loop methods have been proposed to allow users to incrementally refine a topic model by incorporating human knowledge [Andrzejewski et al., 2009, Hoque and Carenini, 2016, Hu et al., 2014, Lee et al., 2012a]. These methods typically involve the use of direct manipulation of topic models to diagnose topic qualities and improve them through operations such as adding or removing words in topics, adjusting the weights of words within topics, splitting generic topics, and merging similar topics [Hoque and Carenini, 2016]. For example, ITM [Hu et al., 2014] allows users to add (limited in corpus vocabulary), emphasize, and ignore words within topics.

While these operations can be supported by direct manipulation and algorithmic extensions, it is also challenging to diagnose the quality concerns of machine-discovered topics, and in assessing if a refinement strategy results in topic improvement. Therefore, visualization techniques are widely adopted in interactive topic modeling. Topic-based clustering of documents has been demonstrated as a useful visualization technique that helps identify related subsets of documents during
exploration of a large collection of documents. For example, UTOPIAN [Choo et al., 2013] allows users to adjust the weights of words within topics, merge and split topics, and create new topics in an interface that visualizes topics as clusters of documents, represented as scattered dots. In the same manner iVisClustering [Lee et al., 2012a] allows users manually create or remove topics, merge or split topics, and reassign documents to another topic, with the help of visually exploring topic-document associations in a scatter plot. Other visualization techniques have been applied to provide a summary of large data sets. Topic Browser [Chaney and Blei, 2012] uses a tabular visualization technique to assist assessing term orders within each topic, creates a navigator of the documents, allowing users to explore the hidden structure that a topic model discovers. ConVisIT, a topic modeling system for asynchronous conversations that revises the model on the fly on the basis of users’ feedback, visualizes the topic hierarchy as a tree-like graph. Termite [Chuang et al., 2012] focuses on supporting effective evaluation of term distributions associated with LDA topics by visualizing relationships between entities of topic models as matrices. Different from applying existing visualization techniques, TopicLens develops a creative lens that allows users to dynamically explore a subset of document collection by recomputing a topic model for the subset in real time and thus helps uncover the finer-grained topical structure of large text corpus [Kim et al., 2017].

In addition to visually-aided topic manipulation, human knowledge can also been incorporated into topic models in the form of specifying must-link/cannot-link constraints (whether certain words should be in one topic) [Andrzejewski et al., 2009], human-defined semantic knowledge [Chemudugunta et al., 2008], or anchoring words (force one or more words have to be attached to a topic) [Gallagher et al., 2017, Lund et al., 2017].
5.1.3.3 Disconnection Between Implementation and User Perception

While many tools and techniques have been proved to be promising in supporting human-in-the-loop analysis of large texts, most of them are developed for users who are equipped with topic modeling experience. In another word, these approached are designed to address the problems of configuration complexity and topic model quality while users who lack knowledge about topic modeling implementation may have difficulty interacting with the systems [Lee et al., 2017].

One exception is [Hu et al., 2014], who applies must-link and cannot-link constraints to implement the operations of adding words and removing words. However, this work lacks descriptions about the design rationales and thus fails to provide implications for designing similar tools for non-experts. UTOPIAN [Choo et al., 2013] and iVisClustering [Lee et al., 2012a], on the other hand, support modifying word weights in topics, but directly manipulating these numbers is dangerous. Non-expert users have little knowledge of how the underlying model works and do not understand the semantic meaning of the operations, thus the results of taking certain actions are unpredictable or make no sense to them.

5.1.4 Interactive Topic Modeling for Non-expert Users

Compared with the interface that is built on top of a static topic model, studies show that users prefer an interface that allows users to dynamically refine underlying topic models. [Hoque and Carennini, 2015] found that the capabilities of splitting and merging of topics are favored by users during the analysis of online conversations. Another study by [Hu et al., 2014] also pointed out that enabling users to improve topic models interactively can better engage non-expert users.

In order to find out how non-expert users perceive topic models and what operations should be available to non-expert users, [Lee et al., 2017] firstly conducted an interview study to analyze user understanding of topic models by asking them...
to evaluate topics and propose topic refinement operations, and then conducted another crowd-sourced experiment to find out user preference of topic refinement operations by asking participants to improve topic qualities. The interfaces used in this work are implemented without underlying topic models updated and are thus flexible enough to avoid the constraints of actual topic modeling implementations. The results reveal the expected topic refinement operations and their usage patterns (ordered by overall frequency): remove words, remove documents, change word order, add words, merge words, split topic.

However, this work has certain limitations when taking into account the actual implementation. In terms of the topic refinement operations, first of all, the added words should come from the original corpus [Hu et al., 2014], otherwise, the words may be helpful as labels, but cannot be used for model updates. Similarly, merging words inside one topic does not contribute to refining the underlying model except making the topic itself more interpretable. Changing word order, on the one hand, is difficult to implement as most of the current implementations cannot enforce such constraint. On the other hand, only top n (10 <= n <= 20) topic words are shown in most interactive topic modeling interfaces while a topic contains hundreds or thousands of words, the influence of changing the word order is subtle. Similar to changing word order, removing a document from a collection of thousands or hundreds of thousands of documents has a relatively subtle influence to updating the underlying topic model; allowing users to remove (filter out) a subset of documents could be a better feature to support [Gretarsson et al., 2012, Kim et al., 2017].

As to the ordering of preference of primary refinement operations expected by non-expert users, the usage patterns of these operations actually depend on specific tasks and topic modeling implementations. Proper preprocessing can help reduce the frequency of some of the topic refinement operations [Schofield et al.]. For example, merging words with the same root can be avoided by either stemming or lemmatization. Connecting words that make up phrases can be handled by
n-gram technique before feeding into the model. Carefully identifying and removing stopwords can also benefit the training of topic models and increase the quality of discovered topics [Schofield et al., 2017].

The initial number of topics, or in another word, the ratio of topics to documents, may influence the usage pattern: if the number of topics is relatively small, more splittings are needed as single topics are likely to contain multiple irrelevant themes; if the number is large, coherent topics may be scattered and thus requires more topic merging operations. The attributes of document collection can also make a difference: if individual documents are more topically concentrated, users are less likely to remove documents.

[Lee et al., 2017] provides meaningful insights for supporting topic model refinements for non-expert users, and also emphasize the importance of revisiting their findings using a fully functional interactive topics modeling system. It is foreseeable that, in selecting and supporting topic refinement operations, a balance needs to be made between non-expert user expectation and topic model implementation, and the characteristics of tasks and data sets have to be taken into account as well. It is worth noting that for statistical topic models such as LDA, seemingly small changes by end-users may have unpredictable results. Therefore, the ability to reproduce the results, and preserve topic patterns after refinements is important for evaluating the effect of refining topic models.

The representation of document collection is also critical for non-expert users to make sense of documents [Lee et al., 2017]. Many current interactive topic modeling tools only make topic words (represented in different forms) presented to users while improved topics may not be consistent with the document collection, and thus cannot be used to help to make sense of large texts. Studies of sensemaking show that the most popular (thus low learning curve) way of organizing content is to group content as small clusters and attach a label to each cluster [Choo et al., 2013, Fisher et al., 2012, Lee et al., 2012b]. Therefore, in this work documents are
represented as topically clustered groups. Note that the granularity of documents depends on the segmentation of text content.

Several measures have been proposed to evaluate the quality of topic models. Perplexity is probably the most widely used way, which measures the log-likelihood of a held-out test text set [Blei et al., 2003]. However, studies show that perplexity does not correlate well with human judgments of topic qualities [Chang et al., 2009]; sometimes they are even counter-correlated. As a result, researchers also developed measures of topic coherence based on pointwise mutual information (UCI coherence) [Newman et al., 2010], normalized pointwise mutual information (NPMI coherence) [Bouma, 2009], asymmetrical confirmation measure between top word pairs (UMass coherence) [Mimno et al., 2011], or context vectors of every topic top word (CV coherence) [Aletras and Stevenson, 2013]. Evaluations conducted using all publicly available topic relevance data indicate that CV coherence has the best performance in terms of interpretability [Röder et al., 2015]. Still, topic coherence is not a perfect measure of user perception of topic quality. For example, Users who are given topics of high coherence tend to decrease the quality of topics through refinement operations [Lee et al., 2017].

Interactive systems are sensitive to latency. Users need to receive feedbacks, either intermediate results or final outcomes, from a system in a reasonable time to continue interactions. LDA training is time-consuming and a retraining of an LDA model in real time is almost impossible. In [Lee et al., 2012a], only the inference step of LDA algorithm is performed without updating the model itself. [Hu et al., 2014], on the other hand, designs a new topic modeling algorithm for the interaction purpose, making the latency of updating models in time ranging from 5s to 50s.

The reproducibility of topic modeling results is important for comparing information patches before and after updating models. In order to make results reproducible, a fixed random seed should be used and initiated every time wherever a random number need (e.g., LDA, t-SNE). As interactively refining topics is not a
linear process, by combing this feature with the ability to undo a topic refinement operation [Lee et al., 2017], users are able to go back and forth during the topic refinement process, rather than having to continue from the current state.

[Gretarsson et al., 2012] provides another direction in addition to focusing on manipulating topics themselves: iteratively selecting a subset of documents based on topical representations of the documents can help concentrate on documents of interests (more coherent) and thus produce topics that are more consistent with the selected documents. This method is consistent with information foraging theory, especially the getter/scatter browsing technique [Pirolli et al., 1996]: generated topics provide summaries of document clusters (gather) and a new topic model will be updated based on the selected documents (scatter).

To summary, an ideal interactive topic modeling system designed for non-expert users are expected to support topic refinement operations include at least: add, remove, and adjust words, merge and split topics, and filtering documents. Multiple means should be provided to retrieve documents to provide a rich context for topic sensemaking. The system should provide an easy-to-use interface that supports intuitive and straightforward interactions: e.g., drag-and-drop of topics words. Visualization of documents should be provided as well as context for topics and better topics can intern facilitate document sensemaking. Users should be allowed to retrieve previous topic modeling results and start from there if the current result is unsatisfying. Although it is doubtful whether non-expert users are able to improve topic models, they should gain a better understanding of the topics through interactions.

5.2 Design Considerations

Towards supporting making sense of large government data, this work focuses on understanding the specific challenges of designing interactive topic modeling tools
for ordinary citizens in the context of analyzing a document collection.

We gained insight by actually practicing interactive topic modeling on the petition data we collected from the White House online petition website “We the People”. This data set is considered a unique source for understanding citizens’ policy concerns and preferences [Hale et al., 2013]. The insight gained from this practice is used to inform the design of a visual analytic system that supports topic model diagnostics, refinement, and evaluation. We reflect the use of visual analytic methods to enable users to interactively make sense of topics.

To designing interactive topic modeling for non-expert users to explore online petition data, the selection of topic model type and the design of interface should be considered together. The user interface for ordinary people is expected to be designed in a way that users do not have to worry about the underlying implementation and the interacting with visual elements is intuitive and straightforward. As the literature reviewed in the previous section demonstrates, existing interactive topic modeling either only applies the inference step of generative topic modeling [Lee et al., 2012a], develops new topic modeling techniques specifically for interactive topic modeling [Choo et al., 2013, Hu et al., 2014], or designs interfaces without implementations [Lee et al., 2017]. Many of them even ask users to directly manipulate topic weights, of which the semantic meaning varies greatly from model to model. However, I found that a new anchoring technique for topic modeling, which can add a constraint that one or more words have to be attached to a topic, could be very useful for interacting with topic models [Gallagher et al., 2017, Lund et al., 2017]. For example, in addition to its potential of supporting topic and topic word refinement operations, correlation explanation-based topic modeling provides competitive performance compared with LDA while runs much faster.

Because the system is designed for ordinal people who lack topic modeling experience, we want to only expose things that have to be taken care of by the users while hiding other things related to concrete implementations, e.g., preprocessing
Figure 5.1: Topic analysis using topic coherence. The context vector-based topic coherence of fitting the online petition data is shown for CorEx models, where the number of topics varies from 10 to 185 of texts, the select of topics, and topic model parameters. As the result, things invisible to users have to be carefully configured. For example, to provide a proper initial number of topics k for users to start with, we refer to existing literature on conducting vanilla LDA topic modeling on similar dataset [Hagen et al., 2016] and also experimented different k on our dataset. While varying K consistently improves coherence, K values of 20 to 30 present relatively high coherence score, in the meantime, it makes it possible for users to manually make judgments of topic qualities. As a result, we select 20 as the initial number of topics.

One motivation of merging words in a topic found in [Lee et al., 2017] is that the words share the same root. Both stemming and lemmatization can solve this problem by converting related word forms to a common canonical representative form [Bird and Loper, 2004]. In order to achieve the same goal, stemming removes derivational affixes of words through a heuristic process while lemmatization removes inflectional endings of words by referring to a vocabulary (dictionary). A more
detailed explanation of comparing these two techniques can be found in NLTK documentation. In our case, lemmatization is preferable since the results will be presented to end-users to interpret.

It has been demonstrated in many interactive topic modeling tools that a representation of documents, either an overview visualization of document collection or showing individual document text, is useful to provide context for refining topics. Nevertheless, the selection of specific visualization technique, as well as how to integrate topic view, document collection view, and individual document view to support iterative topic refinements process, is still undetermined.

A commonly used technique for visualizing topically represented documents is t-SNE (t-distributed stochastic neighbor embedding) [Choo et al., 2013, Lacoste-Julien et al., 2009, Zhu et al., 2012]. Applying topic modeling to document collection can be viewed as representing each document as a vector of which each vector component describes how important the corresponding topic is in the document. t-SNE is a technique that can reduce the multi-dimensional documents to a 2-dimensional space so that users can perceive. There is also another dimensional reduction technique, such as principal component analysis (PCA). PCA usually performs worse than t-SNE since it can only capture linear correlations among variables but runs faster. In reduced 2-dimensional vector space, topically similar documents are placed adjacently. However, cluster boundaries may not be easily identified as petitions can be multi-themed [Paul Hitlin, 2016b].

As identified in [Lee et al., 2017], assessing the quality of topics, the coherence of retrieved documents and the consistency of topics and documents are equally important. Therefore, the interface should integrate the views of refining topics, an overview of document collections, and individual documents, and enable users to follow an iterative topic refinement process. For example, presented with an initial view of topics and document space, a user can select a subset of documents in overview view based on visual patterns and examine them in details in individual
document view by carefully reading, and conduct topic refinement accordingly.

Data Preprocessing is critical for topic modeling as well. Removing documents of low-quality (a document is considered of low quality if the weight of the dominant topics is below a certain threshold) [Lee et al., 2012a] and carefully creating stopword list (e.g., in addition to standard stopwords, remove rare words and too frequent words) [Schofield et al., 2017] help improve topic quality. Suggestions for merging words to connect phrases [Lee et al., 2017] can also be avoided through n-gram technique. In this work, as one of the goals of IBKC is about the completeness, all documents are kept in the initial round. We manually created a stopword list with the help of word counts. A bigram is used to solve most of the phrase problems (“white” and “house” will appear like “white_house”).

Computational measures of topic model quality could be useful in human-in-the-loop topic modeling. On the one hand, topic coherence can help validate the outcomes of topic refinements. On the other hand, it can provide evidence to direct users to the best place to focus. For example, topics with high coherence can be kept while ones with low scores can be removed. Topics with average coherence may have the potential to be improved [Lee et al., 2017].

5.3 FILA: System Implementation

Following the principles described in the previous section, an interface that enables non-expert users to interpret, assess, and refine topics in context was designed and implemented. This section illustrates technique details and design rationales of the interface, as well as demonstrations of supported interactions.

5.3.1 Data Preprocessing

As topic models treat documents as “bag-of-words”, the first step of preparation before model training is tokenization, which splits each petition into a set of words.
As words may have various forms, lemmatization is then applied to transform them into a common base form. Compared with the stemming technique that shares a similar goal, lemmatization takes advantage of vocabulary analysis and thus can produce the dictionary form of words that users can interpret. Bigrams are also used here for performance purpose [Wang and Manning, 2012]. Finally, stopwords are removed from the texts, as well as the overly common terms that appear frequently (top 50), to avoid possible discrimination. The resulting corpus contains 11,189 unique terms.

5.3.2 Interface Design

Figure 5.2 shows the user interface of interacting with petition documents and topic words. This system has two functional areas. The lower part is a topic-word visualization that supports direct manipulation of words-to-topics correlation.

The upper part is designed for exploring the topic quality from the perspective of how the petitions (documents) are clustered according to the space defined by the topics. The points cloud map provides a visual overview of the petition space where topically similar petitions are positioned adjacently. It is generated using t-SNE (t-distributed stochastic neighbor embedding), a technique that is widely used for visualizing topic modeling results [Choo et al., 2013] by reducing the high-dimensional petition data to a 2-D vector space that human can perceive easily. Due to its nature of being non-deterministic, t-SNE usually transforms a high-dimensional data point to a different 2-D vector. However, the relationships between the data points will remain almost the same. An example of visualized petitions is shown in the Figure 5.2.

Each petition is assigned to one cluster based on its most salient topic and is color-coded correspondingly. Users can apply filters and highlighters on topics to manipulate the petition overview map. Highlighting enables users to review petitions in context while filtering allows users to focus on the petitions of interest.
Figure 5.2: The user interface for exploring a large document collection based on interactive topic modeling
When hovering over a document point, a pop-up window displays the title, body, and topics of the document. In the meantime, the topic distribution (in terms of weights) of the selected document is visualized as a bar chart. By clicking a topic label, its topic-words distribution is visualized as color-coded bars.

Topic refinement view is placed in the lower left part. Each topic is represented as a block. Topic words are visualized as labels with bars represent their weights. Each topic is assigned a different color, and the same color scheme is used in the overview view and individual document view. Users can reorder words inside a topic for better interpretability or move a word to another more appropriate topic through drag-and-drop operations. Buttons available in this view allow users to apply changes, roll back to a previous state, or export underlying data for other usages (e.g., visualizing topically represented petitions using Tableau that supports richer interactions).

Document reading view, placed in the lower right part of the interface, allows users to take a closer look at the selected documents. Initially, all documents are selected and listed in the scrollable table of contents. Users can select a subset of documents via, either performing a “lasso” selection in the overview view or clicking a topic block to retrieve documents with highest affinity scores. For each document, the title, the rationales, the assigned topics, and the signature count (and threshold) are shown. Keywords are color-coded if they contribute to the topics attached to the document.

At the back-end of the system, we choose Correlation Explanation (CorEx) [Steeg and Galstyan, 2014] as the topic modeling algorithm to perform interactive topic curation. Built on the theory of Correlation Explanation [Ver Steeg and Galstyan, 2014] in information science, CorEx strives to represent the substrate information in a document collection that maximizes the informativeness of the data. Due to its fast training time and capability of supporting anchoring, CorEx can be easily tailored to incorporate human imposed correlations or constraints for semi-
supervised topic modeling, making it an ideal choice for supporting interactive topic modeling [Gallagher et al., 2016]. Using CorEx, users can anchor multiple words to one topic, anchor one word to multiple topics, or any other creative combination of anchors in order to discover topics that do not naturally emerge. By leveraging CorEx’s capability of topic seeding through anchor words in our system, human analysts can incorporate their knowledge and insights into the process of refining topic models. Furthermore, the running time of CorEx topic modeling is acceptable, e.g., it takes about 15 seconds to generate 20 topics with 50 iterations of training on petition dataset. The number of iterations is set to 50 here because our experiments show that topic model results will converge in most cases with this number of iterations.

5.3.3 Initial Topic Discovery

The first step is to use topic modeling algorithm with random seeds to run an unsupervised discovery of topics. The user must specify how many topics are expected to be produced, with the understanding that different numbers of topics can be chosen to analyze the petitions data on different levels of granularity and it is likely to generate a different set of topics [Hagen et al., 2015b, Paul Hitlin, 2016a].

After initial unsupervised topic modeling with CorEx, users assess the topic model and conduct diagnostic analysis on topics. In particular, users will inspect topics, both individually and as a group, to evaluate their qualities by examining topic words with the help of interface components. Depending on the quality and potential of a topic, users may choose to keep, improve, or discard it.

5.3.4 Supported Operations

Topic refinement is achieved through manipulating topics-word representations at the bottom part of Figure 5.2. We included an anchoring mechanism to be coupled
with CorEx models. It allows users to anchor one or more words to one topic, anchor one word to multiple topics and anchor one or more words to some topics while not others. With this anchoring mechanism, topic revision interactions are supported by operations such as *splitting a topic*, *merging by joining*, and *merging by absorbing* (following [Hoque and Carenini, 2016]). More complicated operations can be achieved through a combination of the above basic operations. For example, investigating more fine-grained topics can be accomplished by splitting topics iteratively. In addition to functions that support topic manipulation, the system also provides functions that enable users to take control of the analysis process. Below is a list of these operations and features.

### 5.3.4.1 Split a Topic

If a topic is considered be “bad” if it contains two or more meaningful themes. A solution could be to split the topic into two or more topics. To do so, the user can check the topic he/she intends to split and then click the “split” button. Before applying the operation, the user is provided with the option to configure the number of resulting topics. Once confirming, the underlying model training will re-run under the new constraint that only the selected topic is decomposed while the others remain the same in terms of word allocation. Updated results will be generated and visualized.

In the backend, splitting a topic into \( n \) topics involves training a *word2vec* model to produce word embeddings [Mikolov et al., 2013]. The resulting model is used to calculate the semantical similarity between words. After that, a similarity matrix of the words within this topic is produced, and spectral clustering is applied to the matrix to categorize the words into \( n \) clusters. The \( n \) clusters of words are encoded into the previous model as anchor words and will produce \( n \) new topics to replace the original one.
5.3.4.2 Merge Topics

If several topics are judged to have something common in their semantic meaning, they can be merged into one topic. This is accomplished by selecting these topics and then clicking “apply” button. The system automatically applies the constraint that words assigned to the topics to be merged have to appear in the resulting topic. Accordingly, the underlying model will be updated and the visualization will be re-rendered. In the backend, the words that appeared in the two topics are now anchored under the same one.

5.3.4.3 Move Topic Words

If one or more words in a topic are considered intruders and fit better with a different topic, the user can re-allocate topic words through drag-and-drop operations. Specifically, a user can select a word that is considered allocated incorrectly and move it to a more related topic. After reallocation of words is done, the petition view will update to reflect the modification. In the back end, Merging topics by absorbing is basically a reallocation process where selected words in one topic are anchored to the other one and a new model is trained. The rest of the topic-word assignments remain the same through anchoring as well.

5.3.4.4 Set the Number of Topics

In the beginning, the number of topics is automatically calculated with the optimization goal of maximizing topic coherence. Although the metric is considered to be consistent with human perceptions generally, it is possible that users are dissatisfied with the number and feel that there should be fewer, or more, topics. In this case, the users are given enough flexibility to set the number of themes themselves, and the topic model will be regenerated accordingly.
5.3.4.5 Navigate through Topic Models

Interactive topic modeling is not a linear process: if a user is not satisfied with an improvement, the user should be allowed to roll back the system to previous states and start from there. However, according to our knowledge, currently, no interactive topic modeling system can support the function of returning to the previous step. Our system uses a specialized data structure to store all the model generated. With the help of fixing random seed, the system guarantees that the same state will be reached when applying the same operations. Sometimes users are not satisfied with the entire topic refinement process. The function of starting over is a shortcut that allows users to start topic refinement again from the beginning.

5.3.5 Usage Scenarios of Topic Refinement

We practiced topic curation process on the online petition dataset to experience how well our system supports topic diagnostics and refinement. Firstly, we run the CorEx topic modeling and generated 20 topics. A fixed random seed was used to make sure the same results can be reproduced. Table 5.1 shows 5 samples out of 20 produced from a topic model. The initial result from the CorEx topic modeling reveals interesting topic clusters from the data set. In the provided samples, topic 0 mainly talks about “disease”, topic 4 generally discusses “economy”, topic 5 describes “election”, and topic 16 represents “law enforcement”. The bottom part of the table shows the results after applying certain topic revision operations.

5.3.5.1 Move Intruder Words

By examining Table 5.1, we find that topic 4 contains a word “health” that is clearly different from other words (see Figure 5.3). We also find that some petitions related to health but has nothing to do with “economy” are assigned to this topic during the petition exploration phase. One example petition is “place mental health as a
Table 5.1: Selected topics (#topics = 20)

<table>
<thead>
<tr>
<th>id</th>
<th>topic words (top 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>disease, patient, cancer, treatment, doctor, cure, disorder, medication, pain, awareness, symptom, illness, medicine, diagnosis, disability</td>
</tr>
<tr>
<td>4</td>
<td>health, economy, tax, cost, benefit, increase, company, money, market, pay, healthcare, fund, research, dollar, debt</td>
</tr>
<tr>
<td>5</td>
<td>election, investigation, vote, voter, candidate, hillary_clinton, voting, campaign, department_justice, fbi, ballot, office, corruption, violation, democrat</td>
</tr>
<tr>
<td>6</td>
<td>internet, consumer, energy, information, technology, provider, service, device, car, access, fuel, safety, standard, road, vehicle</td>
</tr>
<tr>
<td>16</td>
<td>officer, police, law_enforcement, evidence, police_officer, county, aircraft, judge, governor_chris, killing, conviction, department, scene, cat, chief</td>
</tr>
<tr>
<td>0'</td>
<td>health, treatment, disease, condition, patient, doctor, cancer, awareness, pain, illness, medicine, disability, disorder, cure, medication</td>
</tr>
<tr>
<td>4'</td>
<td>money, benefit, company, pay, economy, business, cost, fund, tax, industry, dollar, budget, study, market, increase</td>
</tr>
<tr>
<td>6.1</td>
<td>service, information, com, access, standard, technology, internet, consumer, provider, content, http, privacy, https_facebook, internet_service, customer</td>
</tr>
<tr>
<td>6.2</td>
<td>safety, vehicle, energy, car, device, accident, fuel, road, aviation, forest, traffic, emission, faa, air, carbon</td>
</tr>
<tr>
<td>5+16</td>
<td>investigation, vote, election, officer, police, law_enforcement, campaign, candidate, corruption, voter</td>
</tr>
</tbody>
</table>

required course in junior high and middle schools”. In order to correct this topic assignment, we performed topic refinement by moving the intruder word “health” from topic 4 to topic 0. The re-generated topic words are shown in Table 5.1 as topic 0’ and topic 4’.

In order to assess if such a strategy of refining topics has led to a better outcome, we rendered the petition clusters in relation to the new topic definition and the
result is shown in Figure 5.4. From this figure, we can clearly see how topic groups are isolated and cut. Compared with Figure 5.2, outliers are nicely scattered apart and small clusters of outliers disappear. Such result suggests that the change of topic model by moving “health” from topic 4 to topic 0 is a good move. This claim is further confirmed by a calculated metric of topic coherence based on word context vectors [Aletras and Stevenson, 2013]. This metric has been demonstrated to have the highest correlation with the interpretability of topics [Röder et al., 2015]. The topic coherence of topic 4 is increased from 0.453 to 0.555 after removing the word intruder, and the overall topic coherence is increased from 0.431 to 0.443.
Figure 5.4: A comparison of visualized petitions before and after moving words between topic 0 and topic 4

5.3.5.2 Split a Multi-theme Topic

Observations show that the distribution of petitions of topic 6 is scattered in the reduced-dimensional space: there are several small clusters of petitions. By
sampling some of them for detailed inspection of petition contents, we found that some semantically irrelevant petitions are placed adjacently in the visualization, e.g., “Prevent the FCC from ruining the Internet” and “Put a fee on carbon-based fuels and return revenue to households”, the former is about Internet and information technology, while the latter is related to energy. This finding can also be validated by examining topic words of topic 6: “internet”, “information”, and “technology” are clearly incoherent with “energy”, “fuel”, and “safety”. Therefore, we believe topic 6 is of low quality since it contains several sub-topics and needs to be diluted.

To address the quality concerns of topic 6, we split topic 6 into two topics (by clicking on Topic 6 and choose "Split" button). The modified version of the topic model is shown in Table 5.1 as 6-1 and 6-2 and Figure 5.5 as 6 and 7. The figure
Figure 5.6: A comparison of visualized petitions before and after splitting topic 6 shows that the weights of the first several topic words are increased, indicating that these words can better represent the topics. It is also apparent from Figure 5.6 that the distributions of petitions for topic 6 and topic 7 become more focused, indicating that the petitions documents within the same cluster are more topically homogeneous. After the new topic model applied, the above example petitions are allocated to the correct topics respectively, resulting in an increase of overall coherence value from 0.431 to 0.441. Specifically, the original topic 6 has an
individual coherence score of 0.341, while the scores of newly produced topic 6 and topic 7 are 0.594 and 0.419 respectively.

5.3.5.3 Merge Semantically Similar Topics

If the number of topics is set to a large number, CorEX algorithm will generate topics in a finer granularity of topics. This could create situations where words that contribute to a single theme end up in separate topics. Under such circumstance, a merging operation is necessary to make sure that petitions of similar topics are grouped together. In order to demonstrate this situation, we trained another topic model by setting the number of topics as 50 (relatively large) and the topic words are shown in Figure 5.7a. By looking at the topic words, topic 1 and topic 7 both describe “healthcare” but appear to be different topics.

The topic words after merging these two topics are shown in Figure 5.7b. Petitions of these two topics are now grouped into one cluster as well. Subsequently, these petitions can be processed and analyzed as a whole, e.g., summarized and forwarded to the Department of Health and Human Services.

Merging topics is also useful when a small number of topics is used. Referring to the before-mentioned topic model of 20 topics, we found that topic 5 contains words “investigation” and “justice” that may be related to topic 16. Therefore, we performed a merging by joining on these two topics and it leads to a more general topic denoted as 5+16. Although the coherence value of merging the two topics remains almost the same, it is noteworthy that a new word “corruption” is prioritized as it could serve as a bridge to connect two topics represented as “election” and “law enforcement” (e.g., a petition titled “Arrest and prosecute officials who tried to suppress the vote in the 2012 election”), showing that merging topics has the potential of revealing latent relationship among them.

Topics that are difficult to interpret may still exist even after several iterations of topic refinements. On the other hand, some petitions are complicated in that
they have multiple equally important aspects and even people have difficulty in identifying the most representative one. For those documents that are related to “bad” topics and can not be fixed at this round of analysis, the system can collect them into a subset of data to be fed into the next round of analysis.

5.3.6 Interacting with FILA

This section describes the means of interacting with FILA beyond topic refinement operations. Although NuggetLens is not explicitly designed for improving topic quality, it is necessary for constantly diagnosing topics as well as assessing the impact of topic revisions, since these interactions provide different lenses for examining a
Topics are considered of poor quality if: (1) they confuse two or more themes into one topic; (2) they pick up two different topics that are (nearly) duplicates for human; and (3) exist nonsense topics [Hu et al., 2014], (4) topics with too many generic words (e.g., “people, like, mr”) [Boyd-Graber et al., 2014], (5) topics with disparate or poorly connected words [Mimno et al., 2011], (6) topics misaligned with human interpretation [Chuang et al., 2013], (7) irrelevant topics [Ramage et al., 2009a], (8) missing associations between topics and documents [Daumé, 2009], and (9) multiple similar topics [Boyd-Graber et al., 2014] (a replica of what has been discussed in section 5.1.3.2). These problems can be grouped into problems
of words within a topic, problems of multiple topics, and disconnections between topics and documents. As such, we designed the interface shown in Figure 5.2 that support the diagnose of topic quality from the above perspectives.

5.3.6.1 Inspect a Single Topic

Users can evaluate topics by looking at the coherence of the component words and their relative weights (see the bars next to words) on a topic. Topics are also color-coded in the visualization window. Clicking on the legend of a topic results in all the petitions with sufficient weights on that topic being highlighted (while other petitions are dimmed). These functions allow users to explore the patterns of how petitions of the same topic clustered. A good topic tends to create a cluster of petitions that are less mixed with petitions.

5.3.6.2 Visually Investigate Clusters of Documents

Users can evaluate one or more topics together by observing semantic relations to spatially close or remote topics, and by looking at the spatial relationships (e.g., overlapping clusters, adjacent clusters, or non-intersecting clusters) between petitions of the two topics. Applying filters and producing fewer topics on the figure helps reduce visual clutters.

These clusters of documents, which are visually represented by topics, can also be used as a tool for document retrieval. Instead of working on one topic at a time, a user could also make sense of a set of semantically similar documents. A user can specify the information space by means of highlighting and selecting a particular set of documents, and then identify documents of interests using a lasso selection tool. This feature is expected to be extremely useful when relevant topic words are of poor quality.
5.3.6.3 Drill-Down in Context of Topics

Making sense of concrete documents is usually the next step for document retrieval. After a set of documents are selected, they are listed as a table of contents. For petitions, each simply displays the title and body. For community issues that require document segmentation, a set of keywords are extracted and shown for each paragraph based on their relevance to the dominant topics of the paragraph since usually multiple documents share the same title. The full body of each paragraph is also appended to avoid a lack of context.

In addition, the topic words within a document are colored based on their belonged topics. Hovering over a word shows a tooltip that explains its belonged topics. This feature is useful to connect documents to topics by explaining why they are considered relevant.

5.4 User Study

I designed and conducted a user study in order to learn about whether interactive topic modeling can help people quickly understand the information landscape. In addition to answering the above questions, I am also interested in investigating the usefulness and usability of FILA, the kinds of strategies users employ and how they influence the results. Insights gained from the process could be used to help derive better design solutions for information landscape representations.

Going beyond the latest efforts on investigating non-experts’ behavior during interacting with topic models [Lee et al., 2017], we implemented an interface with a topic model that can be dynamically updated (as opposed to a single static topic model) and supports the majority of desired topic refinement operations. As the goal of the designed task in this study is different from the ones in previous interactive topic modeling literature (generating a summary rather than improving topic quality), we did not expect topic quality can be improved. As a result, we
conducted the study under two conditions of static and dynamic topic models. In addition, according to previous studies, we anticipated that the results tend to be a group and label structure [Fisher et al., 2012].

It is worth-noticing that FILA is designed to help users quickly make sense of an information landscape, rather than improving topic quality. Existing tools have been designed to automatically generate best topic models or a most diverse information subset (with a certain optimization goal/metric). In fact, FILA is designed to include non-expert users in model training loop (no more “take it or leave it” [Hu et al., 2014]), enabling them to incorporate their own knowledge. FILA is aimed to provide a rough structure (schema) of a given document collection collaboratively, and such “schema induction” process [Fisher et al., 2012] can help frame information seeking in terms of schematization.

5.4.1 Methods

Video capture techniques were employed in this study to explore how users summarize the given document collection with an interface powered by interactive topic modeling. Surveys, interviews, and video coding were used to understand the factors that influence the sensemaking of the information landscape.

5.4.1.1 Datasets

We used two types of datasets for our study: ordinances, expert opinions, and news articles related to community issues; and online petitions.

The petition dataset used in this study was crawled from “We the People” (WtP) online petitioning platform and is the same one described in the previous sections on FILA feature demonstration. It is a relatively ideal case for topic modeling as the content is more issue-focused, the lengths of petitions are relatively short and equal (thus does not require segmentation), and there is plenty number of petitions for model training.
One community issue is called “Collegiate Housing Overlay” (CHO). It is a new zoning amendment that increases the amount of housing available for students and may lead to the construction of taller buildings with more functionalities in State College downtown. We collected 45 documents (5 ordinances, 13 expert interviews, 27 news articles from statecollege.com, centre daily times, and collegian) and there are 664 sections after segmentation. Paragraph segmentation was applied in community issue cases since original documents tend to be comprehensive and are less topic-modeling friendly.

Another community issue called “Tax Inflationary Indexing” is a proposal that increases taxes regularly and incrementally to compensate for the inflationary increases. Similarly, we applied paragraph segmentation and it ended up with 459 documents. To better manage the study, we split the WtP datasets into 2 and denote them as WtP1 and WtP2 respectively. Finally, we had 4 datasets. A statistical description of these datasets is provided in the table.

<table>
<thead>
<tr>
<th>Issues</th>
<th>Document Count</th>
<th>Reading Hours (200 wpm)</th>
<th>Avg. Length ± Std.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHO</td>
<td>663</td>
<td>2.33</td>
<td>42.48 ± 61.42</td>
</tr>
<tr>
<td>Tax</td>
<td>459</td>
<td>1.19</td>
<td>34.74 ± 68.80</td>
</tr>
<tr>
<td>WtP1</td>
<td>2141</td>
<td>18.67</td>
<td>104.61 ± 32.11</td>
</tr>
<tr>
<td>WtP2</td>
<td>1839</td>
<td>16.24</td>
<td>105.94 ± 33.36</td>
</tr>
</tbody>
</table>

5.4.1.2 Participants

Subjects came from a range of backgrounds and experiences. All of them are female and are between 21 and 30 years old. Two of them are Data Science major undergraduate students from the College of Engineering, while one is a doctoral student from the College of Information Science and Technology.

All the participants are well-educated and highly skilled at computer use. They are not local citizens and are thus not knowledgeable about local community issues.
Two datasets we use come from WtP petitions and could have been politically polarizing, but politics is not considered as an issue as information summary task does not require subjective opinions.

There are totally 4 datasets. For 2 out of the 3 participants, each of them completed the task on all 4 datasets. For each of these data sets, the conditions for the 2 participants to perform the task are different (one under static, another under dynamic). Each participant completed tasks under two dynamic and two static conditions. Due to conditional constraints, the third participant completed only one session under dynamic condition using the fourth dataset. Eventually, we got 9 sets of study results, among which 5 are dynamic conditions and are static ones.

5.4.1.3 Scenario Description and Task Design

Our specific task questions were chosen for several reasons. Making sense of information landscape is the step before foraging information (IF theory). However, it is difficult to evaluate the schema acquisition of users on a document collection directly. To make the study manageable, an ideal task should produce its own independent result. Previous studies about schema induction indicate that seeking information for learning about a topic ends up with an outcome in the form of multiple terms and a list of detailed descriptions for each term [Fisher et al., 2012]. Inspired by that, the task asks participants to generate a terse summary of a set of documents. In another word, we can evaluate the use of schema instead of assessing schema acquisition directly.

Local citizens want to provide their opinions and suggestions to a [Community Issue]. In order to do that, they need to learn more about this issue. There is a set of documents related to this issue that can help. However, most local citizens do not have time to read them. You, as a representative, will make sense of these documents and succinctly summarize the contents. The summary should enable other citizens to
learn about the available information related to this issue without having to read these documents.

Unfortunately, the amount of documents is too many, making it impossible for individuals to read them one by one [you have a time limit of 60 min]. Luckily, we have a tool called FILA that can help you make sense of them faster and more effectively.

In short, you need to summarize the contents of the provided documents with the help of FILA.

For online petition tasks, the first paragraph is modified a little bit:

Decision makers want to learn about what issues citizens care about, what their direct policy suggestions are, and what things they are petitioning for? In order to do that, they need to learn more about the petitions available on their online petitioning platform. However, decision-makers and analysts are busy and do not have time to read them. You, as a representative, will make sense of these petition documents and succinctly summarize the contents. The summary should enable decision makers to learn about the available information out there without having to read these petition documents.

This task essentially asks participants to summarize the provided document collection under the constraints of limited time, individual setting, and tools available. It should be noted that the task does not require distinguishing the relevant from the irrelevant; an ideal result should be a succinct representation of the given document collection.

5.4.1.4 Procedure

All participants engaged in either dynamic or static version of FILA. The two versions are different depending on whether the underlying topic model can be
updated. We denote the participants as P1, P2, and P3; the datasets as DS1 (Tax), DS2 (CHO), DS3 (WtP1), and DS4 (WtP2); the versions as dynamic (D) or static (S). Although a time limit is specified in the scenario description, the time limit was not enforced - participants were allowed to end the task as long as they believed they are able to form a good summary. The study ended up with 8 sessions, among which 4 are static and 4 are dynamic. After each session, we survey subjects about their perceived usefulness of the tool and interviewed them about their strategies and behaviors (see Appendix B for the detailed questionnaire). Total session time varied between 36 and 97 min. All participants performed tasks in a laboratory environment with a 27’ iMac. All on-screen actions were recorded with screencast software (QuickTime Player [Mac]).

The Dynamic Condition and the Static Condition

In the Dynamic Condition, participants are able to take full advantage of FILA in terms of interactive topic modeling. While in the Static Condition, participants were allowed to update the underlying topic model (with update model buttons disabled). In another word, although users can still refine topics in certain ways (like in human touch), they have to stick with the original topical representations of documents.

P1 and P2 were each assigned two different dynamic tasks and two static tasks with 4 different datasets. The task arrangement guarantees that each task is performed by 2 participants under dynamic and static conditions respectively. Later on, we recruited another participant P3 to work on a randomly selected dataset under the dynamic condition for the purpose of better correlation analysis.

Survey

After each session, participants are asked to complete a detailed survey about the summary they generated and their experience in the task. The survey aims to collect participants’ subjective perception of cognitive load, evaluation of topic quality, perceived usefulness of the tool, and the likelihood they are going to generate
similar outcomes in terms of content and structure. In addition to providing a rating score ranging from 1 to 5, they were also asked to provide more detailed explanations about their ratings.

**Interview and Retrospective Probing**

At the completion of each session, we conducted a semi-structured interview with retrospective probing, consisting of a set of standard questions as well as questions specific to each participant’s behaviors. We asked subjects about their difficulties and challenges they encountered during the task; we discussed their usage of the interface we provided; we tried to understand the tactics and strategies they employ to elicit the practices and norms surrounding the technologies used in this context.

**5.4.1.5 Data Analysis**

Participants’ on-screen interactions with FILA were recorded. Content analysis was conducted by identifying special characteristics of video narratives. Specifically, I coded the following actions users performed: moving topics words around or combine them for creating a new topic, merging homogeneous topics and splitting heterogeneous topics, changing topic labels, selecting or highlighting colored points in the visualized representation of document collection, retrieving a set of documents through a lasso select tool, and hovering over feature words that contribute to a topic, etc. These fine-grained activities are organized under three major views: topic view where topic refinement operations are performed, visualization view where interacting with document representations happened, and document view where participants can read the textual content of a particular document. The time participants spend on each of these views were also calculated based on their mouse trajectory. Based on the coded behavior, as well as the evaluation of the produced results, we analyzed the strategies participants employed to achieve better results.
5.4.1.6 Scoring Performance

The quality of summary is evaluated in terms of the information coverage of a small set of extracted results with respect to the original large set. The metric used in evaluating information coverage is from [Ma and Wei, 2012] and it has two major advantages compared to other major information coverage metrics: it is based on information entropy theory and thus more interpretable; it captures and preserves information structure in addition to information content.

The metric contains two aspects: content coverage essentially measures the textual similarity between the extracted small set and the original large set; structure coverage measures the information distribution of the extracted set with respect to the original set while the structure is modeled using information entropy. Such a metric setting enables us to computationally evaluate the generated summaries from the perspectives of both content and structure.

5.4.2 Results

Our results focus on the process (as opposed to product) of interactive topic modeling that we observed. Since the dynamic condition differs very little from the static one and the frequency of updating topic models in the dynamic condition is relatively low, we do not intend to do a full comparison of performance between the two conditions. Instead, we focus on the outcomes of various strategies and additional cognitive benefits in the dynamic condition.

5.4.2.1 Performance Results

we saw performance differences between conditions; namely both content and structure coverage were higher during dynamic sessions than static sessions, while the redundancy was lower in dynamic sessions. It is worth mentioning that these participants achieved this result with less time in the dynamic condition compared
to the static condition despite that they have to spend more efforts on refining topics and interpreting the updated models.

Table 5.3: Performance Comparison

<table>
<thead>
<tr>
<th></th>
<th>Content Coverage</th>
<th>Structure Coverage</th>
<th>Content Redundancy</th>
<th>Time Spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>0.1426 ± 0.0812</td>
<td>0.8144 ± 0.0763</td>
<td>0.3569 ± 0.0614</td>
<td>74.5 ± 17.87</td>
</tr>
<tr>
<td>Dynamic</td>
<td>0.1430 ± 0.0517</td>
<td>0.8278 ± 0.0807</td>
<td>0.3364 ± 0.0992</td>
<td>60.25 ± 20.03</td>
</tr>
<tr>
<td>Compare</td>
<td>+0.2%</td>
<td>+1.6%</td>
<td>-5.7%</td>
<td>-19.13%</td>
</tr>
</tbody>
</table>

It is initially expected that dynamic interactive topic modeling has advantages over the static one since it provides the flexibility that enables participants to incorporate human knowledge into the process of refining topic models. Such flexibility allows participants to examine the document collection from various angles in a more iterative and manageable way and thus can improve the efficiency of sensemaking, leading to more comprehensive summaries.

Table 5.4: Time Spent Comparison

<table>
<thead>
<tr>
<th></th>
<th>Topic View</th>
<th>Visualization View</th>
<th>Document View</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>6.5 ± 6.1 (8.72%)</td>
<td>4.65 ± 3.12 (6.24%)</td>
<td>63.3 ± 15.37 (84.97%)</td>
<td>74.5 ± 17.87</td>
</tr>
<tr>
<td>Dynamic</td>
<td>10.25 ± 8.61 (17.01%)</td>
<td>6.93 ± 2.67 (11.50%)</td>
<td>43 ± 10.56 (71.37%)</td>
<td>60.25 ± 20.03</td>
</tr>
<tr>
<td>Compare</td>
<td>+57.69%</td>
<td>+49.03%</td>
<td>-32.07%</td>
<td>-19.13%</td>
</tr>
</tbody>
</table>

However, turns out to be very little improvement (with information coverage improved a little bit and redundancy reduced). It can be partly explained by that participants did not stop until they felt they have made sense of all the key points have been collected (same stop criteria). On the other hand, although the dynamic group spends more time on topic view and visualization view, they spent dramatically less time on reading documents, so is the total time.

5.4.2.2 Behavioral Patterns and Strategies

We analyzed the videos in order to identify behavioral patterns and associated strategies participants employed to construct summaries. Four general categories
of behavior emerged from our analysis and each participant employed at least two of them:

**S1** interpret topic words;

**S2** select a subset of documents to show per topic;

**S3** select a subset of documents using lasso selection;

**S4** refine topics using topic refinement features.

Among them, S1 was employed as the primary tactic for making sense of topics (8 out of 9 sessions). In S1, participants attempt to interpret topics based on topic words alone. Topics that can be interpreted by participants are assigned meaningful labels. We anticipated topics of good quality may not require additional complementary information, but the result showed that S1 is participants’ first step after which they always adopt some of the other three strategies for the supplementary purpose. For topics of good quality, participants would like to confirm that they do not misinterpret those topics.

Participants tend to read documents associated with a certain topic if it is difficult to understand it based on the topic words alone. One tactic is to select those documents simply by topic: for a topic, either select all associated documents or select some of them that appear to be more clustered (S2, 7 out of 9 sessions). In addition, participants can also select documents using lasso selection when multiple topics are checked if they appear to be relevant (S3, 4 out of 9 sessions). After a subset of documents is retrieved, participants will choose to read some of them selectively or all of them depending on workload.

Of course, refining topics was common in the dynamic conditions (S4, 5 out of 5 sessions). When participants were restricted to only topic refinement operations at the interface level, they chose not to use them at all as they knew that the change would not actually take effect. In the dynamic condition, participants resort to the strategy of refining topics if other strategies did not work out.
We did not map those strategies on the timeline since it is difficult to tell where the boundaries of applied strategies are. At the highest level, the major strategy used by participants is divide-and-conquer. Depending on the approach by which the process is decomposed, we categorize the general strategy into tactics of topic-by-topic, phase-by-phase (a phase indicates that a user is working on a specific area of the interface, e.g., topic view), and hybrid. For example, a participant can choose to work purely on one topic until the topic is fully elaborated (topic-by-topic), or make sense of all topics from an overview perspective iteratively (phase-by-phase). Therefore, the above analysis is based on whether a strategy is employed in a session.

5.4.2.3 Subjective Feedbacks

Below is a statistical description of the collected survey ratings. It was expected that the differences of ratings between two conditions were subtle as the dynamic version of FILA differs from the static one only in the feature of dynamic model updating, and the rating questions are not directly related to the difference.

<table>
<thead>
<tr>
<th>Rating Items</th>
<th>Static</th>
<th>Dynamic</th>
<th>Row means (± Std. Dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Effort</td>
<td>3.00 ± 0.71</td>
<td>3.00 ± 0.71</td>
<td>3.00 ± 0.71</td>
</tr>
<tr>
<td>Topic Quality</td>
<td>3.50 ± 0.50</td>
<td>3.625 ± 0.41</td>
<td>3.56 ± 0.46</td>
</tr>
<tr>
<td>Helpfulness of FILA</td>
<td>4.00 ± 0.00</td>
<td>3.75 ± 0.43</td>
<td>3.88 ± 0.33</td>
</tr>
<tr>
<td>Likelihood of Creating a Similar Summary Structure</td>
<td>4.00 ± 0.71</td>
<td>3.75 ± 0.83</td>
<td>3.88 ± 0.78</td>
</tr>
<tr>
<td>Likelihood of Creating a Similar Summary Content</td>
<td>2.75 ± 0.83</td>
<td>3.00 ± 0.71</td>
<td>2.88 ± 0.78</td>
</tr>
</tbody>
</table>

Overall, all participants were capable of using FILA with an acceptable cognitive load. They were able to master FILA very quickly (with a 10 min training session) and make sense of a given document collection by taking the most advantage of the features provided by FILA, reflecting a reasonable learnability of the tool.

The perceived quality of the initially given topics was above average. The quality
was assessed based on the coherence and clarity of topic words, the consistency between the topics and relevant documents, and the representativeness of cluster-based visualization. The scores were given since all participants agreed that at least half of the given topics were ready for direct use, while the rest of them require additional actions or should be discarded.

Most participants believed FILA is useful in terms of summarizing document collections in a more efficient manner. Participants benefited most from reading topic words (4) and changing topic labels (3). These two features combined enable participants to better comprehend the meanings of unsupervised-generated topics and organize them. Some participants (2) also mentioned the lasso selection function available in the visualization view is a convenient and effective way of retrieving a cluster of topically similar documents.

In terms of the topical representation of a given document collection, participants agreed that they were likely to create a summary with similar structure, i.e., organizing the content by different aspects of the data set in a hierarchical structure. However, all participants pointed out that topic words themselves are far from enough; sometimes they are difficult to interpret or even misleading. It prompts them to read the original text and connect them to the relevant topic.

5.4.3 Discussion

In this study, we observed several interesting patterns and also found something did not work as expected. We discussed a few of the findings collected from the survey and interview, as well as the limitation of this study.

5.4.3.1 Topic Quality Improvement

In each dynamic condition, we also asked participants to provided their assessments of the quality of topics before and after tasks. It was based on an assumption that non-expert users are able to improve topic quality, at least subjectively. However,
only one session out of four ended up with a perceived improvement on topic quality; two sessions remained unchanged; and one became worse. The major complaint about topic quality include the confusion of topic words, clustered documents do not seem relevant, and the disconnection between topics and retrieve documents.

Due to the problems with topic quality and the fact that participants failed to improve topic quality with interactive topic modeling features, we further asked participants whether they trust the given representations of documents and how likely they are going to use the tool for quick information summary. Both P1 and P2 chose to trust the computational approach. The reason provided by P1 is that more effort does not necessarily lead to better results: “I think both human and machine perform similarly in terms of validity, but artificial work would consume more time and effort”. The limitation of time makes it impossible to read all documents manually and difficult to review and evaluate the topic quality carefully, thus P2 chose to directly use the given topics. P2 also appreciated the flexibility of dynamically updating underlying topic models: “60 min is too short for carefully examining topic quality and consistency. I appreciate the flexibility of interacting with models directly”.

5.4.3.2 Topic Refinements

Although we found that the number of topic refinement operations is correlated with the coverage of summaries, they were not frequently used (Mean: 5.60). P2 pointed out that she was not interested in making the topic quality better as better topics are not necessarily aligned with coming up with better summaries. Therefore, she chose to perform the task with the topics given in the beginning until the task can no longer be continued. In another word, topic refinement operations serve as an approach to unblock the obstruction. P1 also emphasize the topic refinement operations, including the dynamic features, are good to have because these features make her feel that she takes control of the modeling process.
Topic refinement operations are not heavily used could also result from that they are not working as expected. For example, P2 anticipated that splitting a topic would lead to a split of the associated documents. However, our implementation of interactive topic modeling functions in a way that each iteration generates a new model on the entire dataset, while the way P2 expected is to iteratively run topic modeling on a subset of the entire dataset. The problem of “subsetting” in our case is that it may generate a set of topics at various levels of abstractions and thus difficult to organize.

5.4.3.3 Document Collection Visualization

Documents are topically represented and each document is a vector of \( m \) dimensions where \( m \) is the number of topics specified. In order to represent the high-dimensional data on the screen, we have to do dimension reduction. As a result, although their relative positions remain constant and semantically similar documents are located closer, the semantic meaning of document vectors is lost and it makes it difficult to interpret the visualization of the document collection. In particular, low-quality documents that contain multiple topics but none is important, could be randomly assigned to one of the topics, making it even difficult to interpret the document clusters and retrieve relevant documents.

As mentioned by both participants, the visualization is a convenient way to examine the quality of topics (based on the quality of clustering), identity similar topics (closer in the visualization), and retrieve relevant documents, but it requires more training to understand and master the feature. Specifically, they suggested that “low quality” documents should be removed (either by pre-processing or manually).
5.4.3.4 Limitations

Our results should be qualified by the limitations inherent in our study design. First, we collected data from only three users, each of whom had considerably different reading capabilities, interaction styles, and task behaviors. Our analysis is based on their 9 search sessions.

Second, although our task questions were drawn from two types of datasets, each has two, the findings may be directly generalized to other domains, as the nature of documents, e.g., the length, complexity, and comprehensiveness of documents could play a role in performance results. Therefore, we may expect to see different social strategies and behavior patterns for datasets of various characteristics.

Thirdly, in our broader context, using interactive topic modeling for making sense of information landscape is a schema induction process that helps users induce a structured information space for further nugget extraction in information foraging tasks [Gick and Holyoak, 1983, Kittur et al., 2013]. To make this study more manageable (no dependency), the goal of the task is changed to generate a summary of a given document collection. Further study is needed to integrate all the steps of supporting CIR.

5.5 Summary and Discussion

An effective information foraging requires us to explore the resources available before actually exploiting them [Pirolli et al., 1996]. Such “browsing” can allow us to have a basic understanding of the potential information space and thus help formulate specific actions to take. By adopting this browsing technique (Scatter/Gather), users are equipped with automatically summarized contents, clustered documents, and a navigation tool. The interactive topic modeling technique introduced in this section is able to support all of these requirements: documents are assigned to their dominant topics and visualized for interactive retrieval.
Our goal is exploratory and formative: to assess how users interact with an interface powered by an interactive topic modeling system. According to our knowledge, there are currently no systems that are able to support this set of operations due to implementation limitations, such as “latency of model updates and the instability of the model across iterations” [Lee et al., 2017]. Thanks to CorEx, its feature of supporting anchoring constraints make it possible for non-expert users to interact with complicated topic models in a more intuitive, stable, and timely way. There may be other appropriate topic modeling methods, visualization techniques, and interface designs available, but the focus of this case study is on investigating user behavior using a real working system.

The system implementation of FILA and a comparative study demonstrated its capabilities of enabling users to efficiently understand the structure and nature of a given document collection. The acquired initial schema can help better organize the next phase of nugget extraction.
Chapter 6  |  Conclusions

Online civic engagement, especially online forums for discussing and reviewing community-level issues, are often beneficial for online participants, online facilitators, decision makers, and community members. In the IBKC process, such as CIR developed in this work, a small panel of citizens, serving as the mini public, forage all the important information about an issue, make sense of the extracted data, compact the data and present them in an accessible and affordable way. Understanding the challenges and difficulties faced by the participants during the process can provide implications for designing tools to support information foraging in online civic engagement and thus further encourage public participation.

Although the benefits of using a mini-public to engage citizens in decision-making processes are clear, only a limited number of studies consider incorporating information and communication technologies to optimize the process. In this dissertation, we explored the information foraging behavior in the nugget extraction phase of IBKC and designed a visual analytical tool to support activity awareness. The results show that visualization is useful in supporting collaboration in information foraging in the context of CIR. The results also point out the need for automatic text analysis to provide summaries of documents as a schema induction process in order to cold-start information foraging. Complementing the study, we further explored how interactive topic modeling, with the help of visualization, can be used for non-experts to make general sense of a large collection of government data.
Scenario analyses and a comparative study have been carried out to demonstrate the usefulness of the developed tool.

6.1 Research Contributions

While the scope of this research is defined by the overlap between civic engagement, information visualization, and text analysis techniques, this work focuses on the system design to support collaborative nugget extraction in IBKC. Specifically, I position my research in the field of information science by mapping the IBKC phenomenon to what we already know about information seeking.

The first contribution of this research is the integrated conceptual framework for describing the information seeking phenomena in a local CIR context. My conceptual framework is built on top of existing literature for theorization, modeling and supporting information seeking. I conceptualize IBKC from several interrelated dimensions that distinguish it from other information seeking phenomena, i.e. problem context, goal, data, participants, schema, and information behaviors, to understand its unique requirements of collaboration among participants, an evaluation metric of completeness, and information behaviors other than searching. I build on these dimensions to understand the information seeking behaviors in IBKC. My model is able to account for how messy data are transformed into accessible knowledge in the context of the local CIR through a collaborative process.

Beyond the theoretical contribution, this conceptual model has also shown its value in guiding the design of tools for supporting collaborative nugget extraction in my study: it helps us to identify key design issues to support collaborative nugget extraction activities. By conceptualizing the nugget extraction phenomena in IBKC as a collaborative information foraging task with the purpose of achieving optimal information coverage, the design issues for nugget extraction support can be organized from the two aspects of process and products. The former focuses
on enabling individual information foragers to be aware of each other’s activities, while the latter provides support for group members to make sense of intermediate products during the process.

The second major contribution of this study is the visual analytical approach for supporting collaborative nugget extraction. Compared with existing methods for supporting information seeking, my visual analytical emphasizes the active role of the real-time representations of activities in terms of both process and products. While each individual still needs to go through the cognitive processes to develop individual awareness, and use this to coordinate group work, assess task processes, make decisions and take actions; the computer system takes the responsibility to maintain a collective picture of the whole collaborative activity, and promote activity awareness by representing the knowledge in real-time and in an accessible way that enables participants to better extract information nuggets.

Following these design principles, the nugget extraction supporting approach is built based on two categories of components: (1) an illustrative representation of the information landscape that reflects patterns of reading behavior and nugget extraction actions of different group members, and (2) computational representations that provide a statistical description of activities and relationships between constructs. I utilized the design and implementation process of NuggetLens as a vehicle to demonstrate the advantages of my approach in handling the increased level of complexity introduced by the IBKC setting.

The obstacles we found during the practice of supporting nugget extraction led us to propose an interface that enables users to quickly acquire an understanding of an information landscape by providing a structural summary. As a result, information forgers can prepare themselves for nugget extraction with a structural information landscape in mind. Comparing existing methods for supporting information summary, our interactive topic modeling approach can be useful for non-expert users and it allows users to interact directly with the underlying
computational models in an intuitive way.

Figure 6.1: Positioning the major contributions of this dissertation in IBKC framework.

Figure 6.1 shows where the major contributions of this work are positioned in IBKC framework. In addition to a comprehensive description of IBKC, this work analyzes the characteristics of the first phase of the IBKC process and accordingly designs a tool to support the phase in terms of promoting group awareness. It also provides an approach to deal with the preparation of IBKC that happens even before nugget extraction.

6.2 Limitations and Future Work

IBKC is a complex process that involves multiple phases and all kinds of information behaviors and phenomena. Based on the conceptual framework proposed in this study, it is apparent that there are several directions the research can take. This
dissertation concentrates on the nugget extraction phase and leaves the rest for future work.

In addition, this work is still in its preliminary stages and thus I investigate system design issues of supporting collaborative nugget extraction using an exploratory and qualitative methodology based on a limited number of participants. More quantitative methods on a larger scale should be used for data collection and analysis in the future to investigate all kinds of factors and determine their main effects on the process.
Bibliography


171


Paul Hitlin. ‘we the people’: Five years of online petitions. Technical report, Pew Research Center, 2016b.


A. Schofield, M. Magnusson, L. Thompson, and D. Mimno. Understanding text pre-processing for latent dirichlet allocation.


B. Shneiderman. Copernican challenges face those who suggest that collaboration, not computation are the driving energy for socio-technical systems that characterize web 2.0. Science, 319:1349–1350, 2008.


Appendix A: The Questionnaire for Evaluating the Design of TopicLens
Questionnaire

Please submit feedback regarding the NuggetLens tool.

* Required

1. Your name

2. NuggetLens supports the exploration of activities through overview filter and detail-on-demand. *
   (Refer to Screen 1)
   Mark only one oval.

   1 2 3 4 5

   Strongly Disagree  ○ ○ ○ ○ ○ ○ Strongly Agree

3. NuggetLens makes a good use of color coding schemes for understanding activities. *
   (hint) Effective use of color can enable pre-attentive processing.
   Mark only one oval.

   1 2 3 4 5

   Strongly Disagree  ○ ○ ○ ○ ○ ○ Strongly Agree

Screen 1
4. Please provide justification or additional comments to your answers about your assessment of NuggetLens.

________________________________________________________________________________________

________________________________________________________________________________________

________________________________________________________________________________________

________________________________________________________________________________________

5. If you want to assess your team’s completeness, which part(s) in the NuggetLens do you find helpful? Why?
   Please be as comprehensive as you can.

________________________________________________________________________________________
Please open NuggetLens now and answer the following questions

6. If you want to assess your own progress towards complete, which part(s) in the NuggetLens do you find helpful? Why?
   Please be as comprehensive as you can.

   ______________________________________
   ______________________________________
   ______________________________________
   ______________________________________
   ______________________________________
   ______________________________________

7. What are the (other) reasons that motivate you to refer to NuggetLens?

   ______________________________________
   ______________________________________
   ______________________________________
   ______________________________________
   ______________________________________
   ______________________________________
8. If you want to know the places in documents that could potentially contain important information, which part(s) in the NuggetLens can indicate such places?

You and your team members may have skipped/skipped through some sections, and you need to ensure you didn’t leave behind anything important in those texts. Please be as comprehensive as you can.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Please open NuggetLens and give your opinion to the following statements

9. If nuggets are intensively extracted (as shown in the figure below), the user feels more comfortable with their thoroughness in reading the document.

*Mark only one oval.

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Strongly agree</th>
</tr>
</thead>
</table>
10. For a part of a document, if there is no nugget extracted, but this part has been frequently read (as shown in the figure below), then it is more likely to be complete. *

*Mark only one oval.*

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly disagree</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11. If the frequency of the panel’s nugget extraction decreases (as shown in the figure below), the panel’s work is more likely to be complete. *

*Mark only one oval.*

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly disagree</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
12. If a theme is widely distributed across all documents (as shown in the figure below), then the task is more likely to be complete. *

Mark only one oval.
Appendix B: The Task Description and Instructions, and the Questionnaire for Assessing the Design of QILA

Scenario

Local citizens want to provide their opinions and suggestions to a [Community Issue]. In order to do that, they need to learn more about this issue. There is a set of documents related to this issue that can help. However, most local citizens do not have time to read them. You, as a representative, will make sense of these documents and summarize the important points. The summary should enable other citizens to learn about the available information related to this issue without having to read these documents.

Unfortunately, the amount of documents is too many, making it impossible for individuals to read them one by one [you have a time limit of 30 min]. Luckily, we have a tool called NuggetLens that can help you make sense of them faster and more effectively.

In short, you need to summarize the contents of the provided documents with the help of NuggetLens.

Note

You are not expected to read all the documents one by one. Ideally, you only
need to read a small set of them that you believe (with the help of NuggetLens) to be representative and informative, and you can confidently come up with a summary.

**Interface**

Topic Section: this is how the machine comprehend the document collection by decomposing it into several aspects (topics), words in each topic slot help describe the corresponding aspect.

Visualization Section: documents are positioned based on their semantic contents, and colored based on their primary topic. Topically similar documents are positioned closer. You can highlight, select documents. You can also lasso select documents. Selected documents will show in the Document Section.

Document Section: the table of content of selected/all documents. Clicking each title yields to detailed contents (with topic representations). Documents are sometimes paragraphs in articles, and full articles will be provided as context.

**Questions** (provide explanations if necessary)

**Glossary** (Word definition)

Topic quality: the clarity, coherence, and interpretability of topic words with each topic, and how orthogonal, independent, and separate between topics

Navigate/Index: Such as the table of contents, to seek, retrieve and organize documents

Structure and Content: Structure: various aspects of document collection; 2D hierarchical, irrelevant to the content Content: real contents that articulate each aspect

**Questions**

1. Rating of Cognitive Effort for sensemaking (Effort-consuming:5; easy: 1)

   - How easy you make sense of the documents?

   - How easy you interact with all kinds of elements (buttons, visualizations, topic words)?
2. Self-rated Topic Quality (2 scores, before and after for dynamic model)

- Coherence: To what degree these topics are homogeneous?
- Clarity: Does the topic lead to any confusion or misunderstanding?
- Consistency with documents: After you read some of the documents, do you still comprehend the topic the same way?
- Visual representations: Does the quality reflect on the visualization?

3. Helpfulness of TopicLens (very helpful:5; not helpful: 1)

- What features are most useful?
- What features are less useful?
- What features are desired but not available?

4. Likelihood of Creating a Similar Document Organization (very likely:5; not likely: 1)

- Do you consider using a similar way (clusters of documents structured by topics) to organize topics? If not, how do you organize the summary in your mind?

5. Likelihood of Creating a Similar Topic Content (very likely:5; not likely: 1)

- Do the contents of topic words reflect the majority of the summary contents in your mind? If not, to what degree it covers?

Summary: summarize the document collection in your own way. You can use this space to collect materials, craft sentences, and organize contents.
Vita

Feng Sun

Education

• Ph.D., Informatics, Pennsylvania State University, PA, USA, Aug. 2014 - Present

• M.S., Electrical and Computer Engineering, University of Florida, FL, USA, Aug. 2012 - May 2014

• B.E., Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Jiangsu, China, Sep. 2008 - June 2012

Experience

• Graduate Research/Teaching Assistant, College of Information Sciences and Technology, Pennsylvania State University, PA, USA, Aug. 2014 - Present

• Software Engineer Intern, Facebook, CA, USA, May 2017 - Aug. 2017

Selected Publications


• Guoray Cai and Feng Sun (2018). Interactive Visualization for Topic Model Curation. 2nd ACM IUI Workshop on Exploratory Search and Interactive Data Analytics, Tokyo, Japan.

• Guoray Cai, Feng Sun, and Jessica Kropczynski (2017). Crystallizing Local Political Knowledge for Informed Public Participation. 9th IFIP WG 8.5 International Conference on Electronic Participation, St. Petersburg, Russia.

• Lida Huang, Guoray Cai, Hongyong Yuan, Yan Wang, Qing Deng, and Feng Sun (2018). Modeling threats of Mass Incidents Using Scenario-based Bayesian Network Reasoning. 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA. (Best Paper Award)