In-Situ Sensemaking Support Systems

Nathan Hahn

CMU-HCII-20-109

September 2020

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Aniket Kittur (Chair)
Brad Myers
Adam Perer
Jaime Teevan

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was sponsored by the National Science Foundation (IIP-1701005, IIS-1149797), Bosch Global and Google. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Information Systems, Human Computer Interaction, Information Seeking, Sensemaking

Abstract

In-Situ Sensemaking Support Systems

The internet has become the de-facto information source for individuals — from finding a new cookie recipe, to learning how a transistor works. Tools for helping users find answers to their questions have grown tremendously in response; a user can get an answer in their search results for when the next Pirates game is, or get a list of facts about their favorite movie actress. However, for many questions, such as buying a car, there isn't one right answer — the best choice for an individual depends on their particular set of circumstances and personal preference. For these situations, it's up to the user to make sense of the answer space: accumulate what options are available, what the differences and features of these options are, and eventually choose between them.

This process, sensemaking, is a highly iterative and cyclical process, where information is constantly being found, incorporated, restructured, summarized, and generalized. As users continue to collect new information, they need to adjust and restructure their existing information, while incorporating the key known points of the new information into their understanding of the problem. This constant adjustment of both the data and structure surrounding the data puts a significant mental burden on users, and often requires them to resort to external means to track and manage this information. In the context of online sensemaking, this can be done in notepads, tabs, word processors, spreadsheets, kanban boards, or even emails. As users proceed to move along with their data in this process, they need to manually update and transfer data between these tools, which might often be more trouble than its worth.

In this work, I explore how integrated, in-situ sensemaking tools allow users to manage, structure and evaluate their sensemaking data more effectively. I posit that by reducing the interaction costs for users to externalize their mental models in four key stages of the sensemaking process: seeking, triage, structuring, and evaluation, users can more adequately juggle the large amount of information required to perform effective sensemaking. First, I explored and developed a workflow for performing sensemaking using crowdworkers with the Knowledge Accelerator System. Crowdworkers, using this workflow, were able to answer complex questions while spending less than 5 minutes on a single task, suggesting a lightweight scaffolding that could be adapted for use in individual sensemaking. This led to the creation of a sensemaking framework that connects the process of sensemaking with the cognitive processes and tools users leverage for their sensemaking tasks.

Using this framework, I then went on to develop three systems for supporting individual sensemaking: Bento, Distil and Meta. These systems were designed to

provide in-situ support for one or more of the four phases of seeking, triage, structuring, and evaluation. Through these tools, I was able to identify several key strategies for creating effective online sensemaking tools. These strategies suggest several promising directions for the future of integrated, in-situ browser sensemaking tools.

Acknowledgments

As I completed my junior year of my undergraduate degree, I would have never pictured myself completing a PhD. Luckily, I was fortunate enough to have taken my advisor, Niki Kittur's, class on sensemaking, and the rest was history. Without his constant support and guidance, I would not be where I am now. His energy, optimism and patience have allowed me to push forward through tough times and rejection on my journey as a researcher.

None of this work would have been as robust and complete without my partner in crime, Joseph Chang. He has served as a constant sounding board for my ideas, problems, frustrations, and been there throughout every single one of my research projects. His continuous friendship and support gave me necessary stability during constant uncertainty, and look forward to what we might continue to do in the future. My committee, Adam Perer, Brad Myers and Jaime Teevan, have given excellent guidance on not only this thesis work, but also many fun, novel and unique research projects. I would also like to thank my other Microsoft mentor, Shamsi Iqbal, for her continued support and guidance on not only research, but life as a researcher.

I've had the privilege of being able to work with two wonderful UX and design professionals, Julina Coupland and Brad Breneisen, who have not only made my tools significantly easier to look at and use, but also support the work I've done through numerous user interviews and evaluation. My lab mates over the years, Jeff Rzeszotarski, Felicia Ng, Michael Liu, Hyeonsu Khang and Andrew Kuznetsov, have consistently provided clear feedback and a lot of help in building Ikea furniture for the lab. I've also had the chance to work with a large number of undergraduate RAs, whose designs and creativity inspired and refined many of the systems I've built.

I've been extremely fortunate to have made many friends along the way who have helped both inspire me and ground me throughout my time at CMU. I would especially like to thank Alexandra To, Cole Gleason, Laura Licari, Michael Madaio, Rushil Khurana, Qian Yang, Joseph Seering and Judy Oden Choi for helping to make my life outside of work just as fulfilling (and filled with D&D). The aid of the wonderful staff at CMU, particularly Queenie Kravitz and Diana Rotondo, have helped to keep me on track and ensure I'm able to complete my PhD work without overburdening myself.

Lastly, I would like to thank my parents, David and Nancy, and my siblings, Laurent and Stewart, who have been there for me during some of the most difficult times. Their love and support have given me those necessary breaks to de-stress, as well as the motivation to continue my work.

Contents

Chapter 1

Introduction

People are increasingly relying on web-based information sources to make sense of unfamiliar domains. With the ever increasing breadth, depth and diversity of online information sources, individuals are using the internet for a wide range of research tasks, such as understanding medical diagnosis [50], performing in-depth product comparisons [185], or creating an itinerary for an upcoming trip [33]. While some online information tasks have simple, uncomplicated answers, such as the score of the Steelers football game, others, such as planning a week long vacation, can vary significantly based on situational factors (time of year, age of individuals on the trip, etc.) [43, 150]. Estimates suggest that up to 33% of the time spent online [117, 151, 189], or, as of 2009, around 24 billion hours per year in the US alone, are spent doing this type of aggregation and synthesis [13]. In order to understand how the information they encounter online interacts with their personal situation, users engage in the process of **sensemaking**. Through this process, users collect and develop an information landscape around a particular topic, allowing them to make decisions, answer questions, or generate hypotheses [192].

In its simplest and most general form, sensemaking is the process of combining existing and current information with situation specific parameters to achieve some sort of goal [57]. As researchers have begun to investigate this process in greater detail [128, 181, 192, 244], they have uncovered the unique role that structure development plays in this process. As users come to understand a particular information space, they develop a structure, or representation of the space [192], which they then utilize to seek out additional information and further refine their understanding, or they eventually leverage for their goals. This structure is developed and utilized in both a bottom-up, as well as a top-down manner, with users typically going through several cycles of gathering, summarization and refinement before they have a finalized structure [181, 192, 244]. In Pirolli and Card's notional model [182], they divide up the process into two additional loops - an "information foraging" loop, where individuals are collecting sources of information to exploit, and then the "sensemaking" loop, where problem structuring, evidentiary reasoning and decision making take place. These two processes are tightly coupled together, and one drives the action in the other. As noted in Klein's data frame sensemaking model [128], there is a constant push and pull between trying to fit data found into a particular structure, as well as adjusting and refining that structure to appropriately contain the data.

For example, imagine an individual shopping for a camera for the first time. Based

on their previous interactions with cameras of their friends, they can understand that there are different sizes of cameras, cameras with different levels of zoom, as well as different output quality, and price levels. Because they plan on traveling a lot with the camera, they'd prefer for it to not be too bulky, but still take excellent pictures. This starting frame, or structure, is then utilized for driving their initial search – they might look for "best cameras for travel". As they come across potentially good models they might write them down in a document for further exploration or price comparison across different websites, along with notes from the sources saying why those are good cameras. As they begin to dig deeper into the data, they discover that quality is more complicated – there are cameras that perform well in low light, have better color accuracy, or are higher resolution. Additionally, there are often trade-offs between camera size and zoom level, as well as price and customizability. Using this additional knowledge, they can update their frame / structure, and then revise their searching strategy to look for a "full frame compact camera," as that would fit their quality and size constraints the best. They then have to go back to their existing list of cameras and update them, saying which ones are full frame, which ones perform well in low light, etc. or find additional data to fill those gaps. This process will continue until the searcher feels like they have enough information to make a decision and purchase a camera, or they run out of time or cognitive resources to complete the task with.

1.1 Online Sensemaking

To perform this process online, a user might use a number of different tools, interfaces, and information intermediaries to help them. First, they would most likely use a search engine to look up different cameras, or answer questions about camera features. They might then collect different useful, or potentially useful pieces of information as a collection of tabs, screenshots from pages, excerpts, or links. As they start to assemble a working list of cameras they want to consider, they then leverage a document or a spreadsheet to organize their findings into. Lastly, as they come across incongruent information, missing information, etc. they insert placeholders into their artifact, or generate a list of todo items that they need to look out for as they continue their searching process. Uniquely, due to trends in interface usability, enhanced methods for input, computing power, and device ubiquity almost all of this process can, and might, occur on a single computing device [165, 210, 211]. Previously, due to the rigidity of user interfaces and substantial amount of available information, users might resort to physical media such as books, pen and paper or post-it notes to perform a significant portion of their sensemaking process [6, 181, 192, 212].

The ability to perform the sensemaking process in its entirety on a computing device offers a unique opportunity to take advantage of some key features of modern computing: ubiquity of information access, easy distribution / collaboration, fast transformations into new formats for interpretation, and high speed data processing. While the community has begun to explore some of these facets through collaborative sensemaking interfaces [121, 126, 164, 166, 167, 176, 177, 227], enhanced foraging interfaces [45, 51, 134, 218], and unique interfaces for presentation and organization [95, 111, 142], these tools often focus on one portion of the process, be it foraging or

structuring data.

However, as noted above, sensemaking requires a constant shift between top-down and bottom-up processing. Users need to simultaneously draw conclusions from new data, while applying their personal understanding of a space to eliminate extraneous information and focus on the data relevant to their needs. This continuous back and forth between these two modes of processing offers a challenge for computing systems, as they simultaneously are expected to provide some computation benefit through aggregation (bottom-up) and/or filtering (top-down) while still allowing users, who are adding their own perspective and interpretation, to adjust the outcomes in response to their needs. Because tools are often designed to only support one of these two modes of processing, users are often left with a scenario where they're juggling multiple tools: tabs to track what their searching, a document to allow for the integration of personal preferences, and a table to support evaluation and aggregation. They constantly need to transfer information between these tools in order to gain the cognitive support they require, or in many cases they just forgo using extra tools due to the costs involved.

This creates a unique challenge for sensemaking tools: they need to support both aspects of bottom-up and top-down processing, while allowing users to constantly tweak and add additional data based on their existing experience. In essence, tools need to be able to "keep up" with the internal cognitive model of a user, while also allowing users to manage and manipulate ever increasing amounts of information through tracking, filtering and aggregating data. Because of the nature of the sensemaking process, there are a few key challenges that need to be considered when constructing these tools:

- Low cost means to input and remove data due to it becoming invalid during the process
- Means for users to record, track and externalize their personal preferences and perceptions
- The means for users to easily pickup where they left off, as this process can span multiple sessions

Ideally, a system could support users as they work their way through this process, providing ways for them to manipulate their data as they move through the sensemaking process, instead of requiring them to maintain their data in multiple tools. In this thesis work, through a few different systems, I explore how sensemaking can transform from a process where the onus is on the user to manage and keep track of each data point, to one where computation can take some of the burden off users. By utilizing optimized workflows and computation, users can worry less about tracking, filtering and aggregating the data points that feed into their decisions, and focus on interpretation and outcomes. I posit that by reducing the interaction costs for users to externalize their mental models in four key stages of the sensemaking process: seeking, triage, structuring, and evaluation, users can more adequately juggle the large amount of information required to perform effective sensemaking. In this thesis, I present three systems that support lightweight externalizations of user's sensemaking models through lower interaction costs for tracking data points, structuring incoming data, and organizing data into visualizations for evaluation.

1.2 Modeling the Process

In the first part of this thesis, I built a crowdsourcing system the Knowldge Accelerator (KA), to explore how the sensemaking process could be deconstructed in a way that crowdworkers could complete it with microtasks **(Chapter 3)**. Through this system, I was able to develop a workflow that captured the core activities necessary for online sensemaking: finding sources of information (seeking), extracting information (triage), clustering that information into groups (structuring) and then collating and developing that into a coherent article (evaluation). While the KA system demonstrated a unique way to perform sensemaking with crowdworkers, its workflow also had the potential to be adapted for individual use. KA provided mechanisms for workers to externalize what was "good", so they could pass it off to workers later on in the process. If these mechanisms could be streamlined for individuals performing sensemaking, it could not only reduce their mental burden, but also provide opportunities for computation to step in and automate parts of the process, similar to the benefits of microproductivity [213].

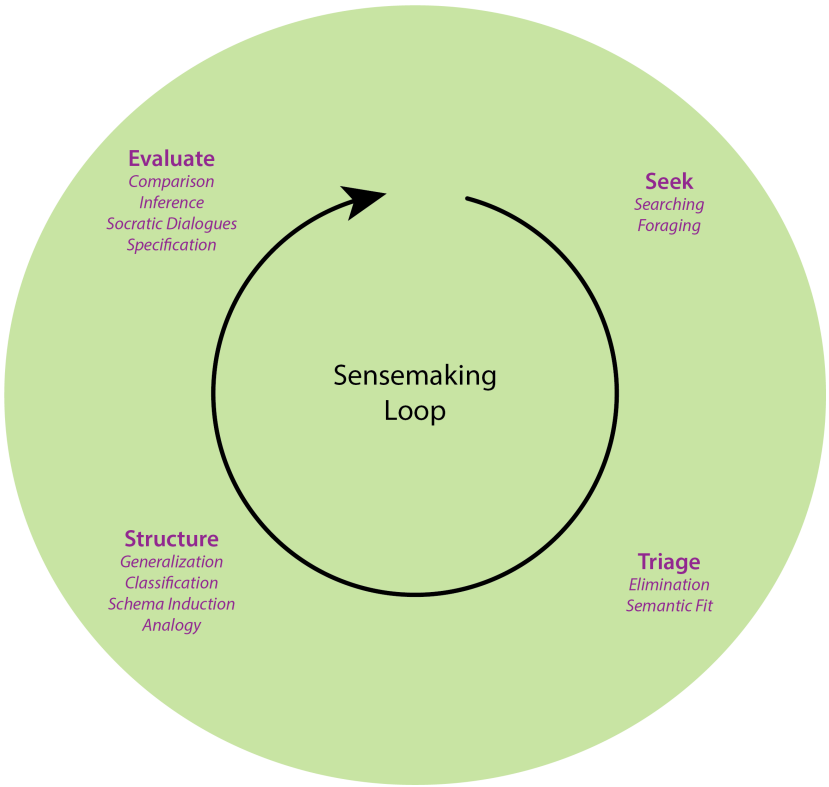


Figure 1.1: A synthesized sensemaking framework based on the KA workflow and existing models

Using this workflow as a guidepost, along with several existing sensemaking models [128, 182, 192, 244], I developed a synthesized framework that seeks to connect the different cognitive needs of the sensemaking process [244] with the typical workflow of a user performing this process in a digital setting (Figure 1.1). The framework simplifies previous models and, unlike previous models, highlights and specifies the necessary cognitive processes that need to be supported when designing tools and interventions

to support the online sensemaking process. In this framework, I highlight four key phases of the sensemaking process: seeking, triage, structuring, and evaluation, and tie them to the cognitive mechanisms occurring in each stage, such as foraging, elimination, classification, and inference. The framework encodes the transitions between bottom-up and top-down processing, with "seek" and "structure" corresponding to bottom-up processing and "triage" and "evaluate" connecting to top-down processing (Figure 1.2). More details regarding the development of this workflow, along with the existing models that drove its development are found in Chapter 2.

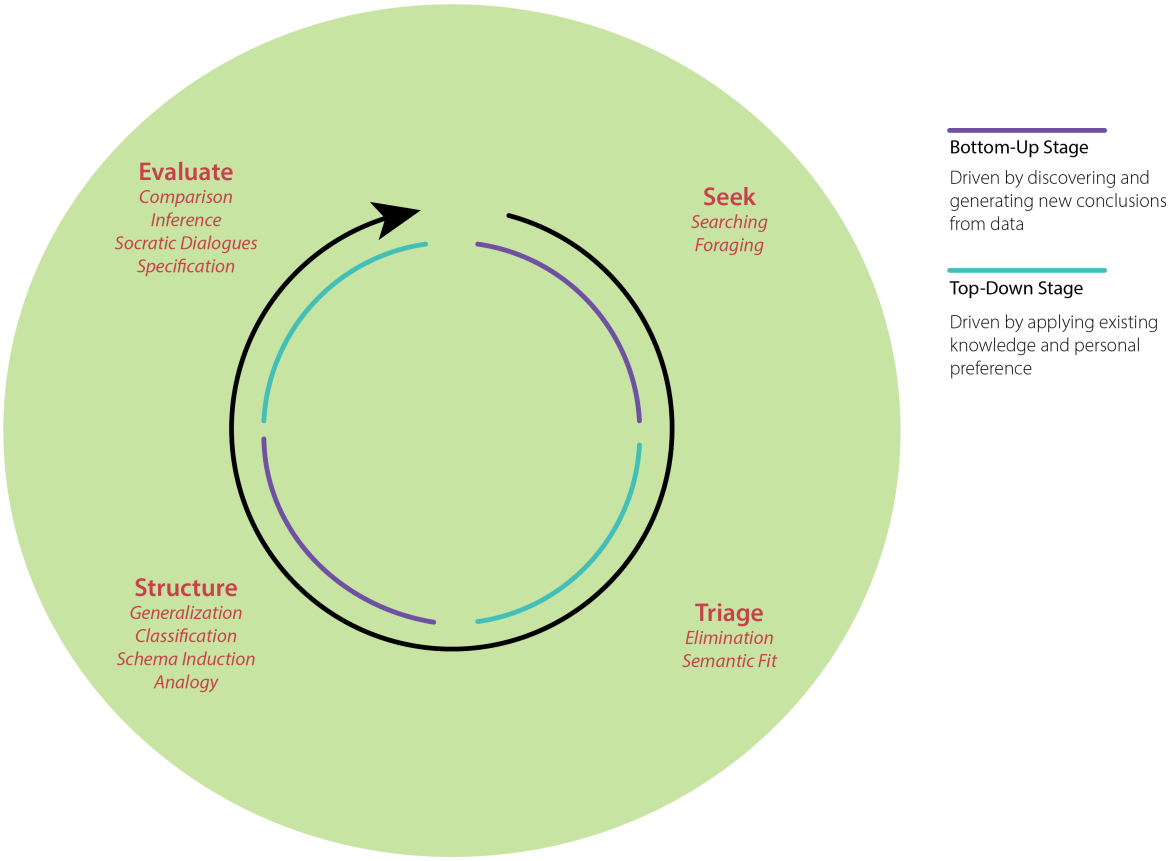


Figure 1.2: The synthesized sensemaking framework highlighting the alternating bottom-up and top-down phases

From this framework and the KA workflow, I then began to design tools that could effectively support one of more of the phases through refined interactions. The three systems I present as part of this thesis work: Bento, Distil and Meta, each sought to streamline and reduce the cognitive burden on users for a progressively larger portion of the sensemaking process (Figure 1.3).

1.3 Overview

Starting with the first two activities in the process, seeking and triage, I developed Bento: a tool for mobile tab management **(Chapter 4)**. Bento was designed to better support the cognitive activity occurring during this phase: moving from one or more

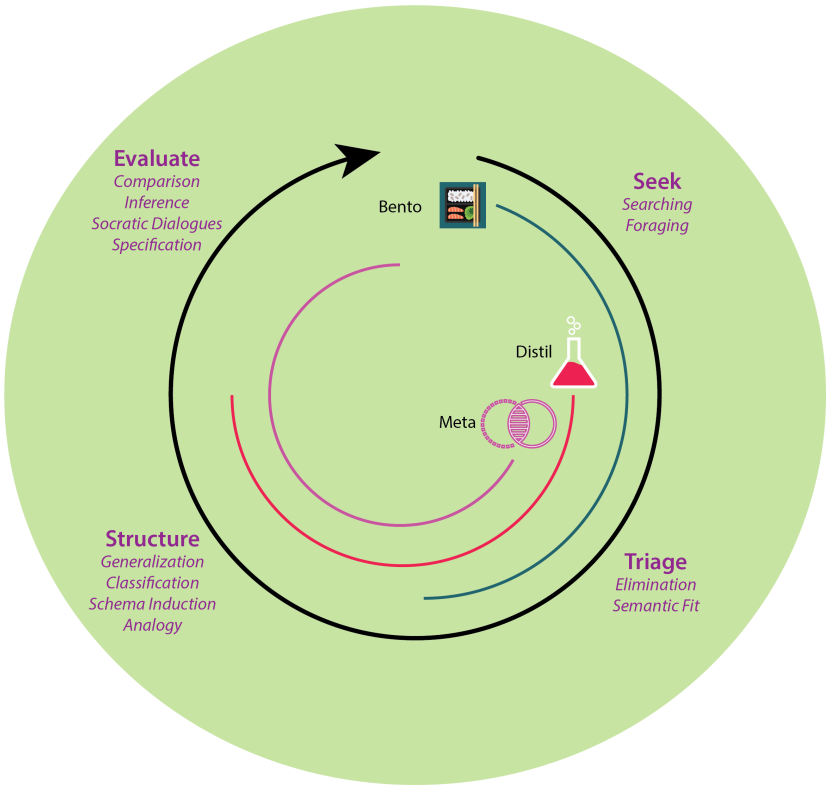


Figure 1.3: A breakdown of the developed systems and which phases of the synthesized framework they cover

unknowns to a series of possible information sources to a curated final set of sources. By creating a "sensemaking workspace", similar to a task-based desktop or workspace [65, 222], users can record their list of current unknowns by creating persistent queries that live along with their workspace. These queries are then executed as searches, through which users can triage the results through trashing, starring, and gaining progress indicators on the list of results. Through these features Bento was able to make users feel more organized and resume their sensemaking activities more rapidly compared to a traditional web browser.

While Bento gave users the ability to triage information at the source level, users are unable to effectively break out, further filter, and categorize information from a source. Throughout the sensemaking process, users have varying levels of uncertainty [41], where early on users are just trying to gain a sense of the information space, and its features, versus later on where specific detailed information becomes critical to decision making. Therefore, tools need to be able to support a fluid transitions between collecting and structuring high level, overview information and smaller, detail oriented information. To tackle the first part of this issue, I developed a toolkit, which allows users to use a variety of selection interactions to annotate and extract information ranging from an entire page, to a specific word. This toolkit is then utilized in, Distil **(Chapter 5)**, which aims to tackle the second issue of managing information in an ever-changing landscape. In Distil, users are able to leverage interactive "smart categories", where they can define auto-updating categorizations that automatically pull in relevant existing

and new information. These smart categories allow users to more efficiently perform the processes of classification and schema induction on their collected data through streamlined categorization. With Distil, users were able to quickly create and adjust their categorizations, using them to both more deeply explore the dataset as well as organize it.

Through Bento and Distil, I was able to explore and tackle issues surrounding seeking, triage and structuring. However, neither system assists users with the final part of the sensemaking process: evaluating information. During online sensemaking, users are comparing the data they've collected from different sources, and trying to match it with their personal needs. In sensemaking senarios such as travel and product shopping, these are often discrete options such as a product or a place. Often times, different sources can have different opinions, leaving it up to the end user to determine who to trust and listen too. To support users dealing with these decisions amount a wealth of information, I developed Meta **(Chapter 6)**. Meta takes a user's open set of tabs, extracts the options highlighted by any articles, and collates them. The Meta interface then provides a number of different views, including a brand popularity table, sorting and filtering tools, and a trustworthiness scorecard to help users figure out what sources agree on, and what some of the more promising options are. Meta works as a seamless, integrated system, where a user can go from unstructured open tabs, to a set of structured interfaces that enable evaluation.

Through these systems, I am able to demonstrate how lowering interaction costs during the sensemaking phases of seeking, triage, structuring and evaluation gives users the ability to manage, organize, and draw conclusions from the data they've collected quicker and more thoroughly (Table 1.1). In the final portion of this dissertation **(Chapter 7)** I summarize the different innovations and contributions from the systems, and look at how they could be integrated and further extended in future systems.

System	Sensemaking Stages	Contribution
Knowledge Accelerator (KA)	Seek Triage Structure Evaluate	Developed a workflow to accomplish sensemaking with microtasks Created novel design patterns to deal with worker motivation while completing these tasks
Bento	Seek Triage	Designed and implemented a mobile browser that supports improved task management and triage during sensemaking Demonstrated that the browser allows users to track and resume their sensemaking activities more easily compared to traditional browsers
Distil	Triage Structure	Developed keyword-based "smart categories" that allow users to more fluidly apply structure to information collected during sensemaking with less manual intervention
Meta	Triage Structure Evaluate	Created an entity-driven tool for aggregating and comparing opinions from multiple sources while shopping Demonstrated the tool allows users to gain a more accurate understanding of a product landscape Explored how such a tool impacts user's perceptions of source trust compared to manual, user-driven aggregation.

Table 1.1: A summary of the different systems discussed in this thesis. For each system, the stages of the sensemaking process they were designed to provide improved support for is listed, along with the primary contributions of the systems.

Chapter 2

Background

First to provide some context surrounding online sensemaking tasks and current user tools, I will provide a high level overview of information seeking and sensemaking behavioral models. I will then connect these to the synthesized model created in this document, and use this to discuss some of the previous sensemaking support systems.

2.1 Sensemaking Models

Sensemaking is generally considered to be an iterative process where a user is building up an understanding of an information space in order to achieve a goal [57, 192]. Theories and related empirical work point out that unlike simple factual information finding (e.g., what is the weather, when was someone born), for complex sensemaking tasks like shopping or making health decisions, finding relevant information sources is only the first step in the search process [192, 231]. Users must also perform additional synthesizing to produce an actual understanding. A number of models of sensemaking have been proposed, including Russell et al.'s cost structure view [192], Dervin's sensemaking methodology [56], Klein et al.'s data-frame model [127], organizational process views [78], organizational adaptation views [53, 161], the notional model by Pirolli and Card [182], and the comprehensive model by Zhang et al. [244].

In this work, I develop a simplified synthesized senesemaking model to explain the necessary primitives of sensemaking support (Figure 1.1). This model draws primarily from Pirolli and Card's notional model [182] and Zhang's comprehensive model [244]. The notional model, developed through cognitive task analysis, defines ten processes and six representations of the sensemaking process (Figure 2.1). These representations are presented in a waterfall where a user utilizes bottom-up processes to move to a higher level representation of the data, and top-down processes to evaluate and fill gaps in the representation. Generally, this process moves from information, to schema, to insight, and finally product, with two large loops of activity:

- An information foraging loop where the sensemaker is finding, filtering, reading and extracting information
- A sensemaking loop where the sensemaker is iteratively building and refining a mental model that best explains the data

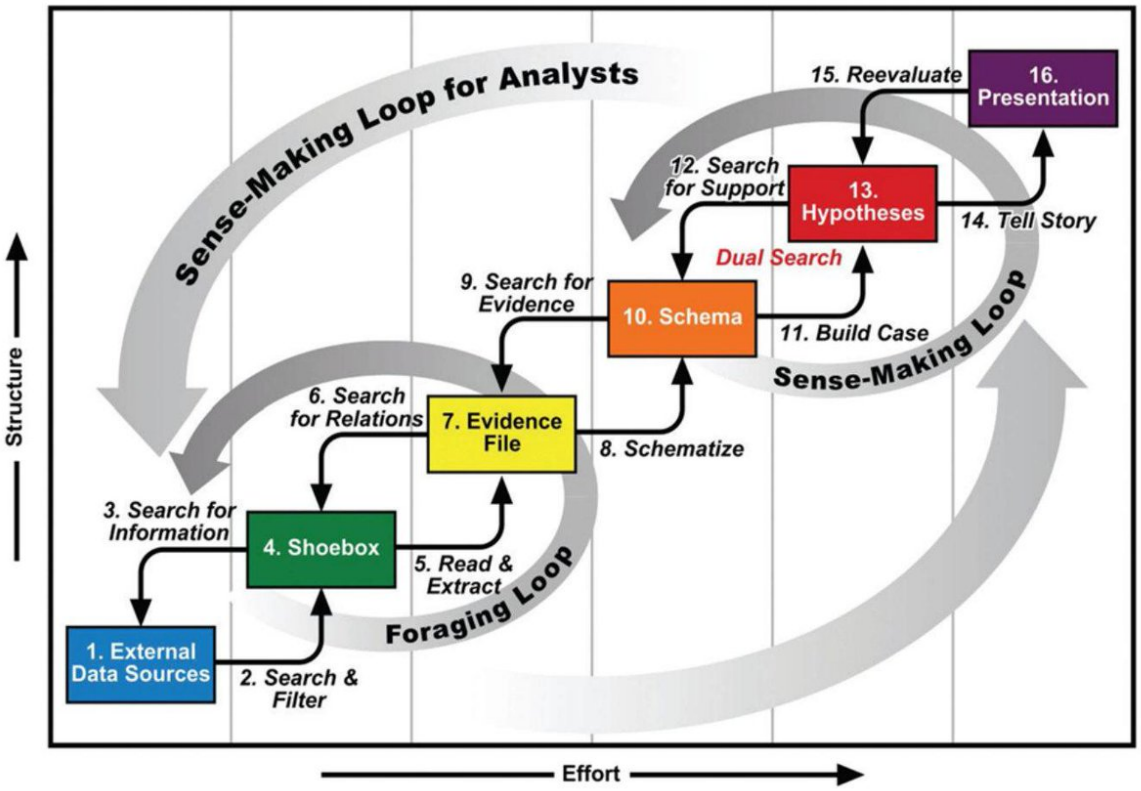


Figure 2.1: Pirolli and Card's notional model of sensemaking

Zhang et al. augments this model by suggesting specific ways in which the structure might be adjusted over time or in one loop, as well as the role that external representations play in the schema generation process (Figure 2.2) [244]. Additionally, Zhang defines a set of top-down and bottom-up cognitive processes from the literature in reasoning, reading comprehension and learning that operate on the structure and data loops (Figure 2.3).

The model highlights four primary activities occurring during sensemaking, listing out several cognitive processes that might be performed in a particular stage. The four activities map directly to Pirolli and Card's notional model stages:

- Seeking: Search & Filter
- Triage: Read & Extract
- Structure: Schematize
- Evaluate: Build Case & Tell Story

The cognitive processes are taken from Zhang's list of top-down and bottom-up mechanisms, which map loosely to the set proposed by the notional sensemaking model. The top-down mechanisms are spread between the "Evaluate" and "Triage" actions, as these are the stages where users are leveraging their structure or model to identity, filter, and fill information gaps. The bottom-up mechanisms are contained in the "Structure" process, as this is where users take the residual information from their triage process that isn't able to be appropriately fit to adjust and update their structure (Figure 1.2).

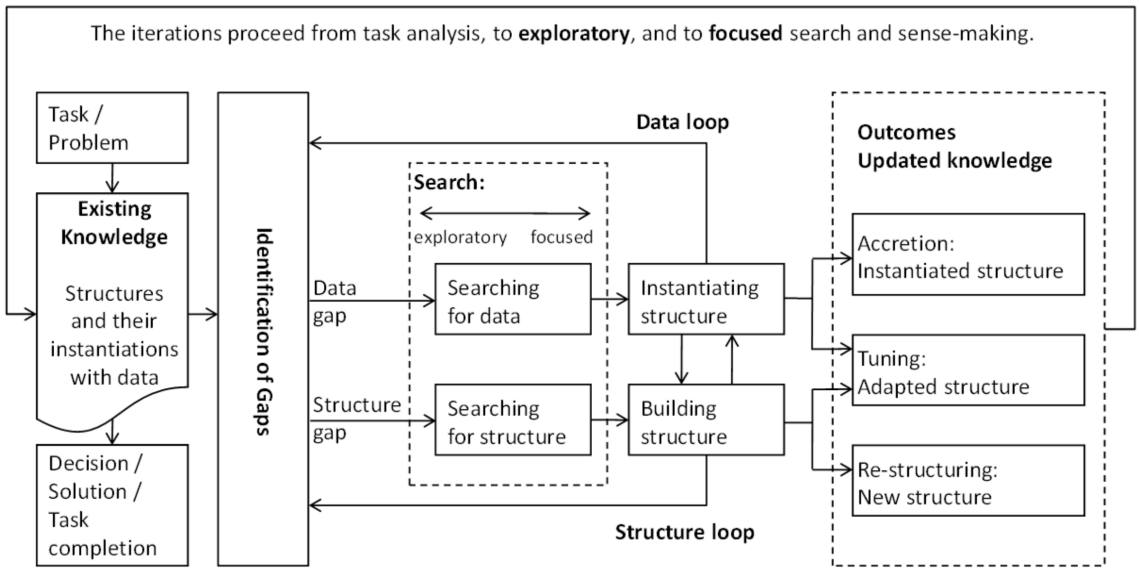


Figure 2.2: Zhang's comprehensive model of sensemaking

Inductive (data-driven, bottom-up) mechanisms	**Structure-driven (logic-driven, top-down) mechanisms**

• *Key item extraction*
Identifying key words/concepts (Kavale, 1980).
• *Comparison*
Comparing a concept to other concepts (Kavale, 1980).
  • *Similarity*
  Recognizing common features or attributes shared by concepts (Vosniadou & Ortony, 1989).
  • *Differentiation or discrimination*
  Recognizing different features of concepts (Chi, 1992; Vosniadou & Brewer, 1987).
• *Analogy and metaphor*
Analogical reasoning: concepts that share common features or belong to common categories may exhibit other common characteristics (Toulmin et al., 1979; Vosniadou & Ortony, 1989).
• *Classification*
Relating a concept to a broader conceptual category (Kavale, 1980) and grouping of sufficiently alike concepts.
• *Schema induction*
Discovering regularities in the co-occurrence of certain phenomena (Rumelhart & Norman, 1981; Vosniadou & Brewer, 1987).
• *Generalization*
Making claims about groups based on a sufficiently representative sample (Chi, 1992; Toulmin et al., 1979).
	• *Definition*
	Defining different aspects of a concept, such as purpose, function, and use (Kavale, 1980).
	• *Specification*
	Specifying conditions or requirements of a problem or task (Vosniadou & Brewer, 1987).
	• *Explanation-based mechanisms*
	Reasoning from cause: examining the causal connections of two phenomena (Toulmin et al., 1979).
	• *Elimination*
	Eliminating structures or facts that do not meet certain criteria in certain attributes (Kavale, 1980).
	• *Semantic fit*
	Examining the reasonableness with which a concept appears to fit a certain schema slot as it relates to the meaning of the knowledge structure as a whole (Kavale, 1980).
	• *Socratic dialogues*
	Critical dialogues to facilitate awareness of inconsistencies in the current schema. Recognition of anomalies can play an important role in initiating schema restructuring (Vosniadou & Brewer, 1987).
	• *Inference*
	Drawing a conclusion or making a logical judgment on the basis of circumstantial evidence and prior conclusions (Johnson-Laird, 1999).

Figure 2.3: Zhang's list of cognitive processes

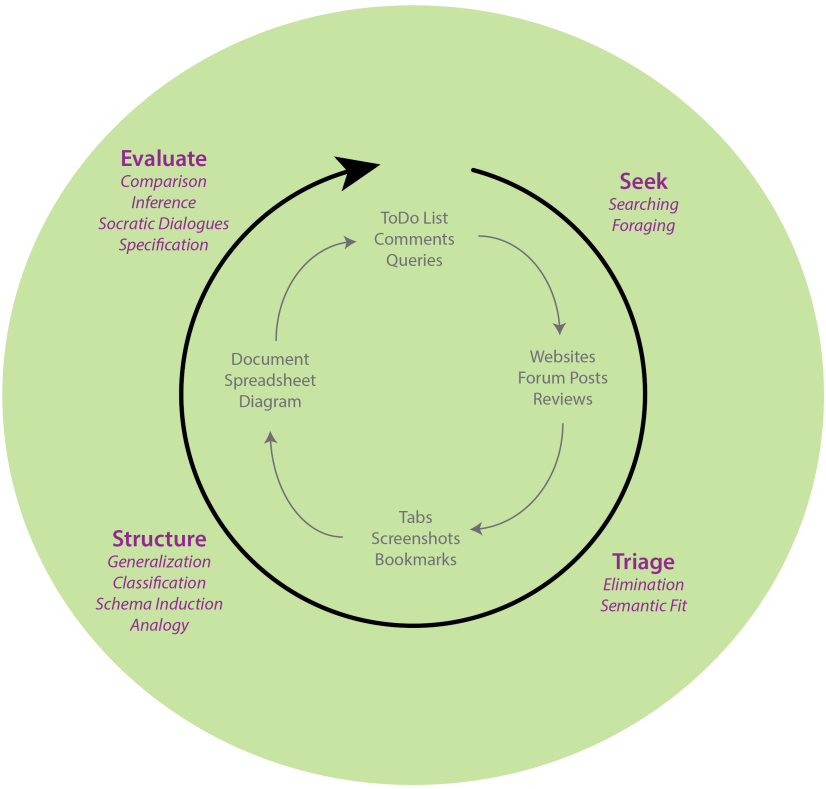


Figure 2.4: The synthesized framework with existing digital artifacts and tools outlined

In this document, I utilize this model as a way to highlight the contributions of the systems I've developed (Figure 1.3). However, beyond this it serves as a way to map specific end-user tools and interventions to the sensemaking process, for example we can use it to explore tools a modern user might use in their online process (Figure 2.4). In this example, Websites, forums, and search results are the online external sources of information that feed into the sensemaking process [13, 32, 232]. Tabs, clippings, and bookmarks serve as the evidentiary intermediary [42, 165, 210], and documents, spreadsheets, and diagrams are the tools through which users realize their models [123, 142]. Finally, ToDos, inline comments, and queued queries serve to represent current unknowns or residuals which have not been fully investigated yet [66, 220]. More generally, this framework serves to highlight the key points where interventions could be developed to better support end users, and what core cognitive processes would need to be supported in those interventions.

2.2 Sensemaking Systems

Over the past two decades, researches have developed a number of individual sensemaking support tools, typically designed to support one of the four activities of seeking, triaging, structuring, and evaluation. Below I give a brief overview of a selection of these systems.

2.2.1 Seeking Tools

Some of the earliest tools designed to support online sensemaking revolve around assisting users with finding and managing information sources. These search support tools provide a wide range of support, from improving users ability to find the document they're looking for, to managing and revisiting their collected information sources. One of the earliest tools, Scatter/Gather [51], utilizes document clustering to help users refine their document collection. Through a cycle of clustering and selection, users can achieve the right level of granularity in the documents they need to answer a particular question. Apolo [45] uses a similar technique with belief propagation, where users select individual documents of interest, instead of clusters, to drive further exploratory search in the domain. Intentstreams [15] utilizes an evolving set of keywords, rather than documents, to build out a stream of results that match more specific or tangential queries an individual might want to explore. Faceted Search [135] provides filters to end-users based on common, intrinsic properties of search results. These filters can also provide exploratory searchers with a broad understanding of some of the important features and dimensions in a particular information space. Lastly, DataShift [175] utilizes crowdworkers to augment the search process for queries involving non-traditional search media (such as images) and vague / unusual queries.

Two other tools, SearchBar [164] and Sensemaker [19] enhance the revisitation and refinding experience. Sensemaker introduces collections of search results from one or more sources. End users can continue to build out or further constrain these collection by issuing additional queries. SearchBar, on the other hand, persists user's queries and results over multiple sessions. When a user resumes a sensemaking task after an extended period, they can use these persisted queries and results to resume their sensemaking activities.

2.2.2 Triage Tools

Once users have a set of information sources they are working with, they then proceed to the process of triage or "active reading" [165]. During this process, users filter out irrelevant documents, read, and then markup and consume relevant documents as they build out their understanding of the space. TRIST [111] focuses on document-level triage, using techniques such as clustering, trend-analysis, and entity linking so a sensemaker can quickly focus in on relevant items. Other tools, such as VarifocalRead [130] provide an enhanced reading experience for large documents through three different zoom-level views. Lastly, InkSeine [98] and LiquidText [211] improve the document annotation experience through flexible markup, extraction, and summarization.

2.2.3 Structuring Tools

Users often need to reconstruct the information they've found in useful format for display, consumption and sharing. Tools, such as the Visual Knowledge Builder (VBK) [201] and IdeaMache [142] utilize a free-form canvas (similar to a desktop) where users can position either whole portions or sections of their documents. They can then attach category groupings / labels onto those documents, giving them the capability to

visual structure their information. Other tools further extend this desktop metaphor by adding features such as piles [148] or performing automatic topical clustering [11]. Lastly Hearst et al.'s tool [95] further enforces the cluster-based paradigm by having user assign one or more topics to a particular document. These can then be viewed in a "group view" as well as a "table view". While these tools offer a number of techniques for users to apply hierarchy and organization to their free-form sensemaking data, they are largely based on manual techniques that can take a significant amount of time and effort for a user to implement.

2.2.4 Evaluation Tools

Lastly, users then need to make decisions based on found information. Researchers have build a number of interactive interfaces that aim to support decision making under multi-criteria and multi-option scenarios, such as faceted interfaces [93, 196], table-based decision support systems [48, 144, 184, 206], and visualization systems [195]. While these approaches allow users to evaluate and narrow the scope of their investigation by viewing information subsets or visualizing trade-offs, generally these approaches rely on highly structured pre-compiled metadata or require users to manually clip evidence for each source.

Chapter 3

The Knowledge Accelerator: Distributing Sensemaking [1]

To gain a better understanding of the global constraints surrounding the sensemaking process, and what a reasonable workflow might be like for an online sensemaking task, we gave ourselves a "grand challenge" of distributing the process across crowd workers with microtasks. The resulting system, the Knowledge Accelerator (KA), through this constraint of microtasks, allowed us to explore the global bounds and restrictions of the sensemaking process. Additionally, it also uncovered potential areas where individuals could receive computation assistance: source collection / management, clipping, structuring, and the development of an artifact. In the KA system, using a variety of computational and workflow techniques, we were able to utilize these workflow primitives that allowed workers to produce summarized answers that were better than the top Google search results for 11 different topics on the dimension of comprehensiveness, confidence, helpfulness, trustworthiness, understandability, and writing. Through the development of the prototype system, we were able to refine several different design patterns that can help to support not only crowdwork, but also individuals users working with large collections of data. In this next section, I provide some additional context surrounding the challenges of crowd work systems, details of the KA system, how we evaluated it, and the resulting implications for both future sensemaking work as well as crowd work.

3.1 The Trouble with Microtasks

Microtasks offer an interesting alternative to conventional tasking, providing a way for workers to complete usable work in context free, bite-sized pieces. Because microtasks are quick to perform, they allow people to work without having to set aside large blocks of time and while mobile [20, 105, 169, 219]. Additionally, due to their limited context, they are easy to share with others and thus commonly used within the context of crowdsourcing [23, 46, 49, 136]. By decomposing and distributing the cognitive work of

[1] Portions of this chapter previous appeared as Hahn, N., Chang, J., Kim, J. E., & Kittur, A. (2016, May). The Knowledge Accelerator: Big picture thinking in small pieces. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 2258-2270).

an individual, crowdsourcing can provide a larger pool of resources more quickly and with lower transaction costs than through traditional work.

However, much work in the real world is not amenable to crowdsourcing because of the difficulty in decomposing tasks into small, independent units. As noted by many researchers [23, 125, 143, 146], decomposing tasks – ranging from writing an article to creating an animated film – often results in pieces that have complex dependencies on each other. Take for example the goal of writing an article that synthesizes information on the web about a given topic (e.g., growing better tomatoes). Coming up with a coherent and comprehensive set of topics (e.g., soil, sunlight, watering, pruning) is challenging without a global view of the data. The need for coherence extends throughout the fractal nature of the article: each section, paragraph, and sentence must have a proper transition and flow. Supporting such work requires having a big picture view of different pieces at different scales and ensuring they all fit together.

Accomplishing big picture thinking through microtasks is challenging because it means that each person can only have a limited view of the bigger picture. As a result, many of the applications of crowdsourcing have been limited to simple tasks such as image labeling where each piece can be decomposed and processed independently. Those approaches that do crowdsource tasks requiring big picture thinking — such as volunteer communities such as Wikipedia, open source software, or paid crowd work approaches such as flash teams [186] or Turkomatic [136] — have relied on a heavily invested contributor such as a moderator or an experienced contributor to maintain the big picture. For example, in Wikipedia, a large proportion of the work is done by a small group of heavily invested editors [126], and the quality of an article is critically dependent on there being a small number of core editors who create and maintain a big picture structure for more peripheral members to contribute effectively [121].

In this chapter, we explore how a computational system, the Knowledge Accelerator, can scaffold an emerging interdependent, big picture view entirely through small contributions of individuals, each of whom sees only a part of the whole. Through a development of a working software system and an evaluation across a variety of topics, we were able to create a set of design patterns which can aid in the development of future systems dealing with the issue of a large global context.

3.2 Related Work

In the development of the KA system, we drew heavily from previous sensemaking models in the development of the distributed workflow [53, 56, 78, 127, 161, 181, 192, 227]. Generally, the models agree that sensemaking is a dynamic and iterative process involving searching for information; filtering that information based on a user's goals and context; inducing a schema or structure from the information; and applying the schema to take action (e.g., writing a report, making a presentation).

A number of systems have been developed aimed at supporting these stages of sensemaking for an individual user [19, 57, 58, 140, 150, 175] or a group of users working together [121, 126, 167, 176, 177, 227]. However, prior research has focused almost exclusively on situations of integrated sensemaking in which individuals (even in groups) are heavily engaged in the entire sensemaking process. Instead, we sought to distribute

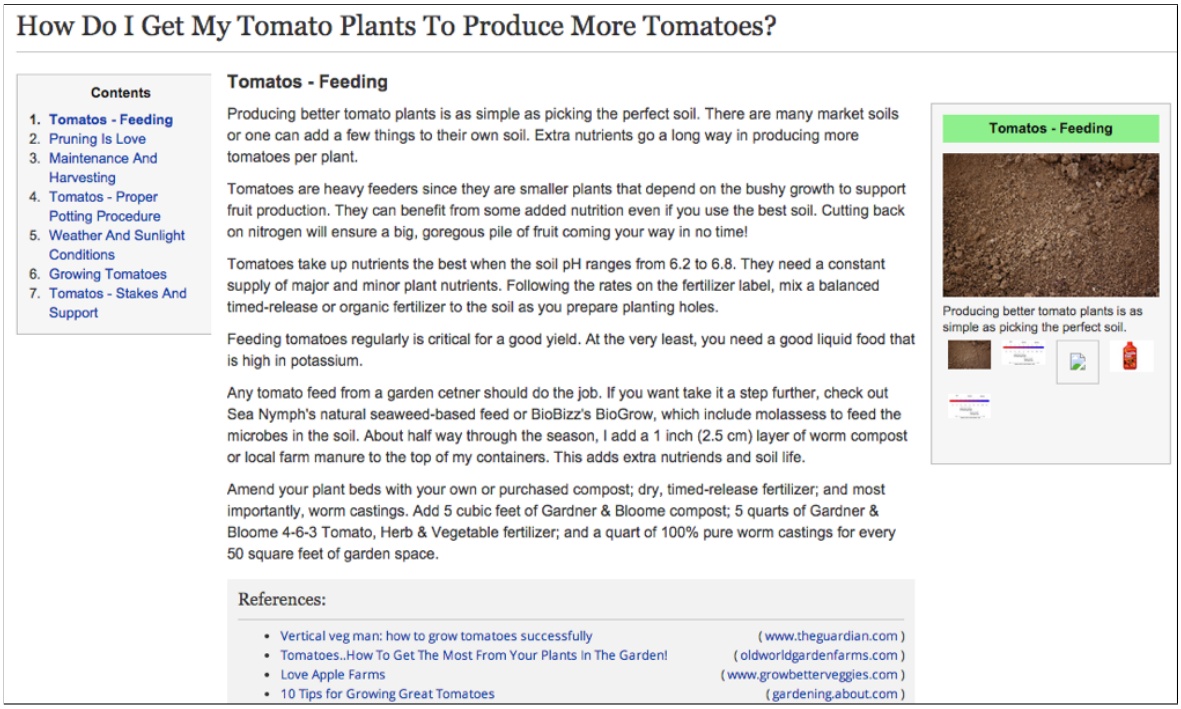


Figure 3.1: The final output of the Knowledge Accelerator system.

the information synthesis process across many different individuals, each of whom may see only a limited view of the process.

3.2.1 Crowdwork: Complex Cognition and Workflow

While most crowdsourcing approaches have focused on simple and/or independent tasks, there is a growing interest in crowdsourcing tasks that tap into complex and higher-order cognition [122]. Many of these fall into the class of decomposing cognitive processing in a structured way such that many workers can contribute [8, 23, 28, 118, 120, 125, 136, 138, 139, 143]. Our work builds on this foundation by incorporating adaptive crowd workflows (e.g., TurKit, JabberWocky, CrowdWeaver), crowd-driven task generation (e.g, CrowdForge, Turkomatic), combining the outputs from decomposed tasks to create a global understanding (e.g., Cascade, Crowd Synthesis) and a multi-stage crowd quality control process in which crowds can both generate new versions of output as well as vote on it (e.g., CrowdForge, Soylent, TurKit). However, we go beyond previous work in aiming to support a coherent big picture view while avoiding individual bottlenecks. Doing this is significantly more challenging than the tasks decomposed in prior research, requiring a search for structure during the sampling process, a reliance on novices to function with more context than they enter the task with, and a tight interdependence between each subtask such that any failures could negatively impact the value of the entire artifact.

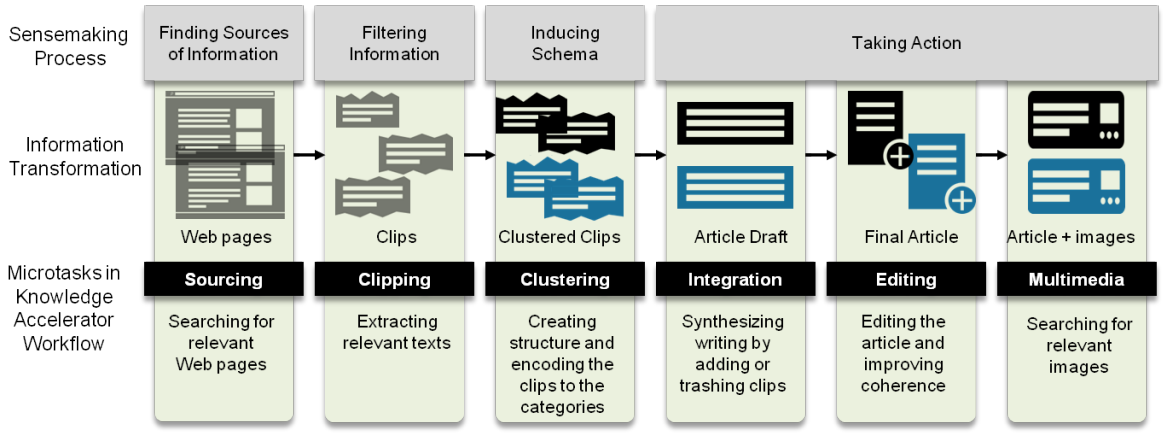


Figure 3.2: The process of the Knowledge Accelerator (KA), from start to finish

3.2.2 Computational Information Synthesis

Finally, some purely computational approaches have been explored for supporting information synthesis. For example, Question Answering (QA) research addresses the methods and systems that automatically answer questions posted by human in natural language. Complex, interactive QA (ciQA) has been introduced at TREC 2006 and 2007 in addition to factoid and list QA [54]. However, automated QA approaches (and their crowd-based variants [25]) focus on answering short, factual questions instead of the complex sensemaking processes we are interested in, where users build up rich mental landscapes of information. Another approach is multi-document summarization [21, 80, 149, 154], which aims to use computational techniques to extract of information from multiple texts written for the same topic using feature based [84], cluster based [108], graph based [69] and knowledge based methods [89]. However, such approaches have limitations in dealing with complex yet short and sparse data like that encountered on the web, and do not yet engage in the complex synthesis humans perform, which results in cohesive and coherent output.

3.3 System Architecture

Broadly, there are two hard problems involved in crowdsourcing information synthesis: learning a good structure for the article based on sampling information from different online sources, and developing a coherent digest given that structure. In this section, we discuss how the Knowledge Accelerator system addresses each of these problems in turn.

3.3.1 Inducing Structure

How can a crowd learn a good structure for an article on an arbitrary topic? Previous crowd approaches such as CrowdForge or CrowdWeaver [120, 125] required workers to decide on a structure before collecting information on each of the topics. However, these approaches fail when the structure must be learned from the data. For example,

few workers will know what the subtopics should be for fixing a Playstation's blinking light or for dealing with arthritis; instead, the appropriate structure should emerge from the data. A single individual making sense of a topic often engages in an iterative process of sampling data and building a structure; however, to reduce the latency of having multiple cycles we explore an alternate approach in which the crowd samples a large amount of data in parallel, then leverage a novel hybrid crowd-machine approach that clusters information into topics without requiring any one worker to see the whole picture.

Finding Sources

To search for and filter high quality information sources we asked five workers to each provide the top five web pages relevant to the target question. We found these numbers to work well in practice; future work using optimization approaches [112] could potentially set these dynamically. To ensure high quality responses, for each source we asked workers to report the search term they used and provide a small text clip as "evidence" showing why the source is helpful. This approach appeared to be successful in encouraging workers to find high quality sources: workers made on average 2 different queries ($\sigma = 0.3$), and their more commonly cited sources covered more categories of the structure with fewer sources than choosing sources using standard information retrieval approaches (i.e., using the MMR diversity-based re-ranking algorithm to reorder the sources gathered from the crowdworkers [34]). Sources cited by at least two workers were sent to the filtering stage.

Filtering Information

To filter relevant information snippets from each source, workers were presented with one web page and asked to highlight and save at least five pieces of information that would be helpful for answering the question using an interface similar to that described in [124] (Figure 3.3). One challenge we encountered was that each page could contain a variable amount of useful information, with some long pages having more snippets than a single worker would extract. To spread out worker coverage on long pages, we showed workers sections that had been highlighted by previous workers and asked them to first look for unhighlighted areas when choosing clips. This preference for novelty and surfacing prior workers' effort allowed us to engage multiple workers for tasks with an unknown amount of relevant information in a more efficient way than simply letting loose many independent workers who would overly focus on the beginning of the page, or having some workers start at the beginning and others at the end [23]. To focus more effort on potentially rich sources the system dispatches two workers to each source with an additional two workers for every two additional citations a source received.

Initially we had workers provide labels to categorize each clip, which we planned to use to develop a structure for the article. However, the lack of context of the bigger picture made these labels poorly suited for inducing a good structure. For example, in Figure 3.4 the top box shows the category structure induced from labels generated during clipping, while the middle and bottom boxes show the structure induced from the subsequent clustering phase and from a gold standard developed by two independent

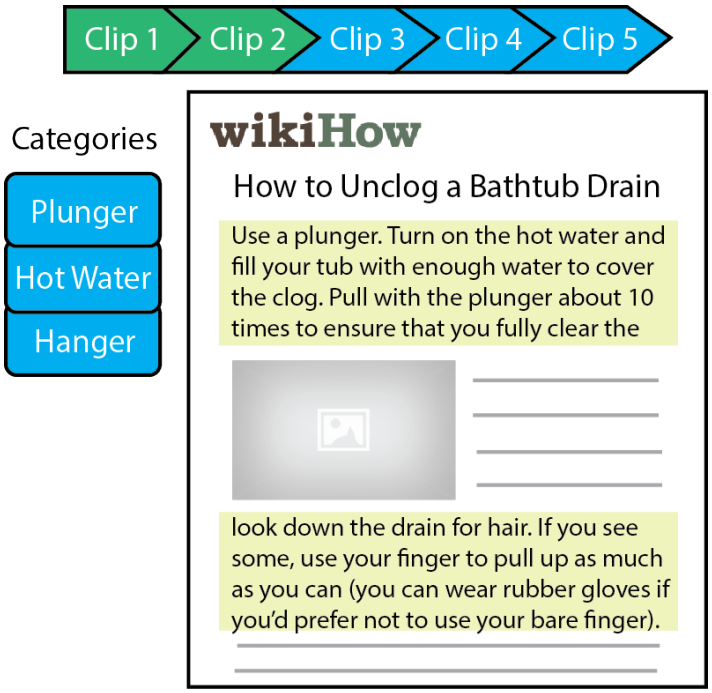


Figure 3.3: Workers extract 5 different pieces of relevant information from pages and give each of them a label

annotators with access to all clips and sources, respectively. Categories induced from the clipping labels poorly match the gold standard, and include categories with very different abstraction levels (e.g., *Use Drano Max Gel* vs *tips*). This motivated the development of the subsequent clustering phase.

```
categories induced during clipping:
Boil Water, use hot water, Plunger, try a snake, How to Remove drain stopper,
bleach, Use Drano Max Gel, baking soda, drain, tips to unclog, problem, tools,
research, internet research, ..., etc.

---

categories induced after clipping:
Hot Water, Plunge, Plunger, Snake the Drain, Remove
the Drain Cover, Drain Cleaner, Remove Hair Clusters.

---

annotator categories:
Hot Water, Plunger, Plumbing Snake, Remove Cover, Chemicals, Bent Wire Hanger,
Call a Plumber, Shop Vacuum.
```

Figure 3.4: Categories induced from different stages for Q1: *How do I unclog my bathtub drain?*

Clustering

Inducing categories in unstructured collections of text typically requires understanding the global context in order to identify categories that are representative of the information

distribution and at appropriate levels of abstraction. The problem of inducing structure without any single worker having a full global context is a particularly challenging problem, and although we describe a basic solution to the problem here for reasons of space and scope, we present a more sophisticated distributed approach in [44] that further generalizes the problem to other domains.

Our approach takes advantage of the fact that many real world datasets have long-tailed distributions, where a few categories make up the bulk of the head of the distribution and many categories with few instances make up the tail. The intuition behind our approach is that first, the crowd can act as a guide to identify the large categories in the head of the distribution, with their judgments training a classifier to categorize the easy cases with high confidence. After automated classification, the crowd can again be used for "clean up", covering the low-confidence edge cases in the tail of the distribution. This also has the added benefit of easily breaking up the larger question context into sub-contexts for easier consumption in the later parts of the system.

In the first phase, we use workers to label a number of representative categories and leverage those labels to identify meaningful features for an automated classifier. One critical challenge is that workers need to obtain a sense of the distribution of the data without seeing it all. To accomplish this we developed a design we call open-ended set sampling in which workers are presented with four random clips as seeds, and are asked to replace them repeatedly with another random clip until they can determine that the four seed clips belong to meaningfully different categories. Therefore, not only do they have to read the information present in the initial seed clips, but they also need to sample multiple times to understand what "different topics" mean for this dataset. In doing so they are randomly shown new clips, which means they are more likely to encounter categories with probability matching the distribution of topics in the data (i.e., higher probability of encountering larger categories).

After workers pick the seeds, we ask them to highlight discriminative keywords in each of the seed clips which are used to query for similar clips from the full dataset, which the workers then label as as *similar* or *different*. With the keyword highlights and the labels created by the workers, we use an SVM classifier and hierarchical clustering to cluster the high confidence portion of the dataset, sending the uncertain instances to Phase 2.

In the second phase, we employ crowdworkers to clean up the output of the classifier, by presenting them the existing clusters on the left of the screen, and the remaining clips on the right. The workers are first familiarized with the clusters by asking them to review the clips in each cluster and give it a short description. They then categorize the remaining clips into existing clusters or create new clusters if no existing cluster is relevant. These categorization judgments are used to refine the hierarchical clustering model.

3.3.2 Developing a Coherent Article

In this section we describe a set of processes which take as input a set of topics and clips for each topic and output a coherent Wikipedia-like article. There are two core challenges in doing this: first, creating coherence within a topic (e.g., consolidating

redundant information); and second, creating coherence between topics (e.g., maintaining consistency across sections).

Integration

Within a single topic, there may be many clips which all contain substantively identical information (e.g., the ideal pH level of soil for growing tomatoes); one goal is to reduce this redundancy so that the final article only describes this information once. At the same time, we recognize the value to seeing that multiple sources all say the same thing; thus, we would like to keep track of all the sources that mention a particular chunk of information. Furthermore, tracking source provenance allows the user to drill back to the original information source in case it is described inaccurately or in a biased way.

To accomplish this we developed an interface in which workers were presented with 5 random clips of information for a given subtopic and asked to integrate that information into a shared text pad. Specifically, they were asked to write the gist of the clip in their own words and transfer the provenance of the clip as a footnote. Missing footnotes triggered a verification check.

Initially, we just instructed individuals to cluster similar items together and insert only the footnote for redundant information. However, we noticed that workers were reluctant to change what they perceived as another worker's contributions, consistent with the social blocking found in Andre et al. [17]. This developed into a larger challenge: How could we get workers to gain an understanding of what was in the existing shared pad and feel comfortable modifying it? We introduced a technique we call evaluate then act that requires individuals to read what others have already put into the integrated answer before they are allowed to make a decision about the clip. Our final interface prompts workers to provide specific line numbers corresponding to existing information relevant to their clip, or to explicitly mark their clip as new information or trash. Compared to a version of the system without this structure, significantly more clips were inserted into the middle of the pad to align better to their given section (13% more, $t(24) = 2.568$, $p < 0.05$) or excluded (11% more, $t(24) = 4.592$, $p < 0.01$) when workers were asked to evaluate before acting.

Editing

We also noticed that coherence needed to be managed not only within topics, but between topics as well. A number of between topic inconsistencies became apparent during the development process, ranging from formatting to structuring to prose. For example, some topics would be organized with bullet points versus paragraphs, and some in the second person point of view versus third person. Previous crowdsourcing approaches have trouble dealing with cross-topic consistency because reading even a single topic can take significant time, let alone reading and editing across all topics. For example, CrowdForge's [125] approach simply concatenates topics into an article without any attempt at maintaining global coherence. This approach can succeed if the topics and structure either do not require consistency or if they are extremely well specified beforehand: in CrowdForge and CrowdWeaver defining a science article "template" with clear sections such as *what is the problem, what the researchers did*,

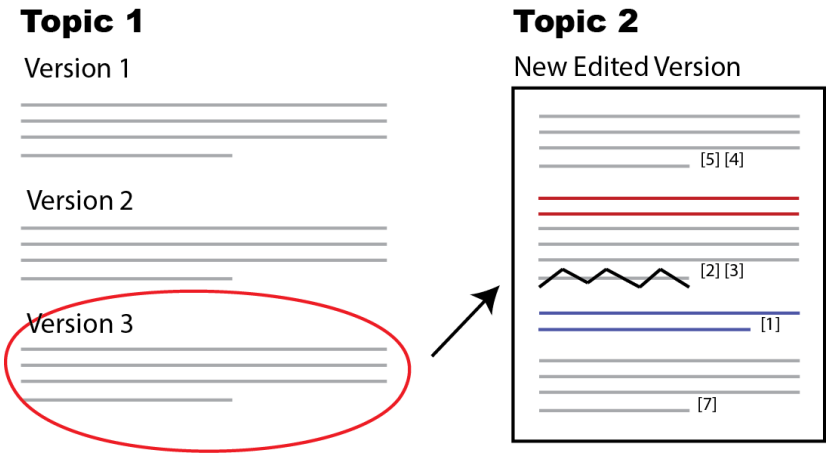


Figure 3.5: Editing users the 'vote-then-edit' pattern to promote consistency and motivate workers

accomplishes this effectively in a similar manner to core editors specifying a structure in Wikipedia that peripheral members then fill in [121]. However, in the general case such well-defined and pre-specified templates are not always available.

To address this we introduced a new pattern which we call vote-then-edit (Figure 3.5). This pattern asks workers to first review and vote on and choose the "best" version of a subtopic created by previous workers, while simultaneously getting a sense for commonalities in style, grammatical choices, and organization. They proceed to edit a new subtopic (phase one) or improve on the item they voted on (phase two). In the second case, we expected workers would more carefully select the best version to reduce their future workload, as well as be more motivated to fix issues in it because they had a choice in what they wanted to do.

We used the vote-then-edit pattern in an interleaved "horizontal" and "vertical" workflow. The horizontal phase uses the refined and edited versions of a subtopic section as a "model" for improving the rough output from the integration phase for another subtopic section. Specifically, three workers vote on which of three versions of an edited subtopic section is the best and then edit a different subtopic subsection using their answer from voting as a model. Their resulting edited output is sent to the vertical phase, in which three workers vote on which of those versions is the best, and are then asked to further improve this now with all of the other subtopic paragraphs presented to them, to ensure the current subtopic has good flow with the other sections. The output from these workers is used in a new horizontal phase, and the cycle continues. The intuition here is that the horizontal phase provides only a single section as a model since there is substantive editing work remaining that requires relatively limited context, while the vertical phase provides all sections because the primary editing work remaining is ensuring consistency across sections. Splitting editing into two interleaved phases with different context-work tradeoffs appeared to be more effective than an older editing approach with a single phase. When we compared the evaluation ratings for the older editing to the interleaved vote-then-edit approach for two questions (Q1 and Q2 in Table 3.1 respectively), the newer answers were found to be significantly more understandable ($\bar{x} = 0.457$, $p < 0.01$) and helpful ($\bar{x} = 0.373$, $p < 0.05$), suggesting this

design pattern helped to create more coherent output.

Multimedia

Images and video can help the reader skim and digest information quickly, as well as provide rich information such as diagrams, instructions, and how-to examples. In our system we enable multimedia from diverse sources to be tied to information blocks, which we define as sections of text demarcated by footnotes. Informally, information blocks correspond to units of information, such as steps in a how-to, or statements or evidence. This has the benefit of ensuring that the images found are specific to pieces of information found in the answer, rather than just being general to the subtopic. For the version of KA described here we did not employ redundancy or voting in the multimedia stage as we did not encounter quality issues; however, since multimedia enrichment is not a particularly interdependent task existing known quality control approaches such as redundancy and voting [122] would likely be sufficient for a production system.

3.4 Design Patterns

As mentioned in the above task descriptions, during our iterations on each stage we ended up introducing several design patterns that improved the output. Each phase had its own distinctive challenges, yet they still suffered from some of the core challenges highlighted by previous work: motivation, quality-control, and context [122]. Our design patterns served to guide our final system design and add to the set of crowd patterns introduced by previous research [23, 28, 122, 125, 136, 137, 143]. They may be particularly relevant for challenges involving complex interdependent tasks requiring global context for workers seeing only local views.

3.4.1 Context before Action

One of the biggest challenges in crowdsourcing a complex, interdependent task such as information synthesis is providing workers with sufficient global context to perform well despite them having only a local view. Previous researchers have suggested a variety of useful patterns related to this goal, including making the cost of spurious answers as high as valid ones [119], identifying and surfacing specific sub-task dependencies [136, 186], unified worker interfaces [243] and re-representing tasks in simplified forms [16, 120]. We contribute a set of patterns adding to this literature, specifically focusing on a key tradeoff: given a limited amount of time and effort for an individual worker, how can we provide workers with global context (i.e., investing in their ability to make better decisions) but also engage them in actual production work? Too much invested time providing context reduces the amount of time available for improved task performance.

Open-ended Set Sampling. One challenge with large datasets is giving workers a sense of the distribution of the data despite their observing only subsets of it. This pattern involves a comparison task in which workers are asked to sample random items from the data in order to create a set of non-matching items, as seen in the first step of clustering. A key design factor in this pattern is having a good set function

that provides a driver for open-ended sampling and also a stopping point (e.g., when a worker's familiarity with the distribution gives them a sense that their four seeds represent substantively different topics in the dataset).

Evaluate then Act. In order to get workers to understand the context provided to them, we designed evaluation mechanisms at the beginning of their main task that would allow them to get acquainted with the output from previous workers. This helped workers understand how previous workers processed the information provided to them, improving consistency of the output on parallel tasks, and reducing repeated information. This pattern was leveraged in a number of tasks: clustering, integration, and editing. In the integration phase, we additionally used the evaluation phase to signal to workers that removing others' work was acceptable and expected, showing that it could be useful in socializing workers into desired procedural practices as well as providing them with context.

3.4.2 Tasks of Least Resistance: Leveraging Worker Choice

Since workers were mostly dealing with dense textual information on a topic they were likely unfamiliar with, we wanted to ensure they were sufficiently motivated. Therefore, we developed a pattern that doubled as both a quality control measure, we well as an incentive for workers. The "task of least resistance" pattern requires that the same crowd worker be involved in two stages of the task, a first stage in which they choose what to work on from a number of alternatives (e.g., voting) and a second stage in which they themselves benefit from their choice in terms of having to do less work, easier work, or being able to submit a higher quality output. The intuition is that to minimize their later work workers will choose a foundation that requires the least amount of work possible; i.e., they will choose the "task of least resistance". This act of choosing is intended to also provide workers with a sense of agency and purpose, which has been shown to increase task performance [39, 188]. This choice also has the potential to increase task performance through workers trying to avoid cognitive dissonance: since workers have themselves presumably chosen the best quality work to start, poor quality final output could reflect on their own worth [226]. This has a trade off of potentially making tasks longer, more complicated, and more expensive, however the benefit is a higher quality output.

3.5 Implementation

The main portion of the application was built using Ruby on Rails and integrated with Amazon's Mechanical Turk through the Turkee ruby gem [110]. The Ruby on Rails application served as the primary user interface for both the question asker, crowd worker, as well as the answer viewer. A question posed to the system would start the workflow, beginning with source finding. For each stage, after a certain set of conditions were met (number of sources, clips, completed clustering, etc.), the next task in the workflow was automatically started. This allowed the system to run through the entire process with minimal intervention.

The clipping task utilized Readability's parser API to simplify the appearance of

the sources provided during the sourcing phase. This allowed workers to view a cleaner interface in which to clip from, and it also removed some technical limitations involved with clipping from pages that might be multi-paged (readability combines these into one long document) or featured heavy javascript functionality that would interfere with the clipper tool.

For the first phase of the structure induction tasks, the TfIdfSimilarity ruby gem is used for searching clips similar to the seed clips [155]. LIBSVM is used for combining the crowd judgments and cluster a large portion of the dataset [40]. For the integration and editing tasks, we utilized the Etherpad-lite text pad library [216] to allow workers to simultaneously work on the same output.

3.6 Evaluation

To evaluate the usefulness and coherence of the system's output we compared it to sources an individual might use if they were to complete this task without the KA system. This would most likely involve the use a search engine such as Google to gather information and use existing information sources to learn about the topic. Therefore, as an evaluation, we had a separate set of crowd workers perform a pairwise comparison of the KA output to that of top results returned by Google and those found useful by multiple crowd workers.

3.6.1 Method

Participants were recruited through the AMT US-only pool and paid $1.50 for the evaluation task. Each participant was randomly assigned to compare the output from the KA system with an existing top website for that question. An individual could only provide one rating per question, but could do the rating task for more than one question. We removed 34 of the 1385 unique participants who provided an evaluation rating who also participated in a KA system task.

The "top websites" used in the comparison task were the top five Google results, as well as any additional Google results that were highly cited (mentioned by 3 or more turkers) during the sourcing phase of the system. Some questions had a larger number of highly cited sources, resulting in more additional websites, as can be seen in Figure 3.6.

In the evaluation task, participants were first asked a series of questions that would cause them to read and understand both sources. In order to encourage quality through defensive task design [119], for the output from the KA system and the existing web page, they were asked to list the different sections on each and three different keywords that would describe those sections. After they read and parsed each web page, they were presented with a brief persona of a friend who was having the problem posed to the KA system. Workers were then asked, for that problem, to rate the comprehensiveness, confidence, helpfulness, trustworthiness, understandability, and writing of each web page on a seven point Likert scale (from 1 to 7) and provide an explanation for their rating on each dimension. We averaged ratings on these dimensions into a single score representing the overall perceived quality of the page.

Question	N	Score
Q1: *How do I unclog my bathtub drain?*	116	0.292 *
Q2: *How do I get my tomato plants to produce more tomatoes?*	177	0.420 *
Q3: *What are the best attractions in LA if I have two little kids?*	158	-0.044
Q4: *What are the best day trips possible from Barcelona, Spain?*	98	-0.109
Q5: *My Worcester CDi Boiler pressure is low. How can I fix it?*	139	0.878 *
Q6: *2003 Dodge Durango has an OBD-II error code of P440. How do I fix it?*	138	0.662 *
Q7: *2005 Chevy Silverado has an OBD-II error code of C0327. How do I fix it?*	135	0.412 *
Q8: *How do I deal with the arthritis in my knee as a 28 year old?*	139	0.391 *
Q9: *My Playstation 3 has a solid yellow light, how do I fix it?*	119	0.380 *
Q10: *What are the key arguments for and against Global Warming?*	138	0.386 *
Q11: *How do I use the VIM text editor?*	138	0.180
* = significant at $p < 0.01$ after Bonferroni correction		

Table 3.1: Average difference between the KA output and top websites for the eleven questions (positive indicates higher ratings for KA, negative indicates higher ratings for the competing website). Each rating was an aggregate of 6 questions on a 7-point Likert scale.

We selected 11 target questions for evaluation by browsing question and answer forums, Reddit.com, and referencing online browsing habits [38]. For some questions, we added some additional constraints to test the performance of the system for more personalized questions. In addition to this external evaluation, we also had the crowdworkers who participated in the KA system fill out a short feedback form detailing their experience using the system. We ask three questions about the difficulty of the task, the clarity of the instructions provided, and the easy of use of the user interface. We recorded some brief demographics about our workers, including to the country they were from.

3.6.2 Results

Aggregating across all questions, KA output was rated significantly higher than the comparison web pages, which included the top 5 Google results and sources cited more than 3 times (KA: $\bar{x} = 2.904$ vs Alt. Sites: $\bar{x} = 2.545$, $t(1493) = 13.062$, $p < 0.001$). An analysis of individual questions corrected for multiple comparisons is shown in Table 3.1.

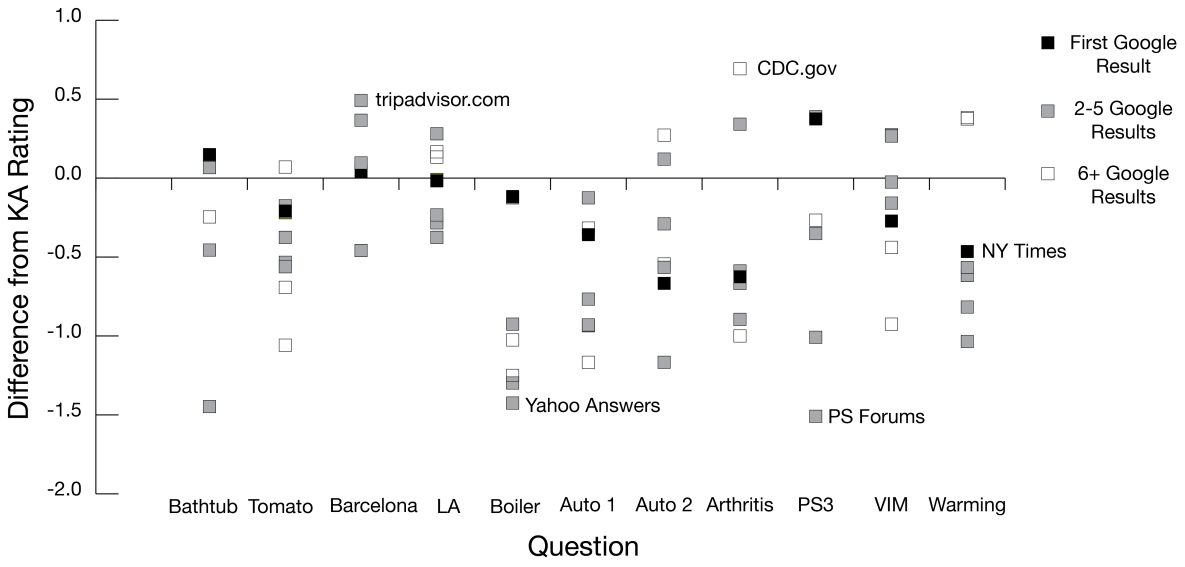


Figure 3.6: Results across questions and websites. Points represent the average aggregate score difference between the KA answer and an existing site. Sources above the line were rated higher than KA, while sources below the line were rated worse

The strongly positive results found were surprising because some of the websites in the comparison set were written by experts and had well-established reputations. Only on the two travel questions, Barcelona ($\bar{x} = -0.109$) and LA ($\bar{x} = -0.044$), and the VIM question ($\bar{x} = 0.180$) did the KA output not significantly outperform the comparison pages. A closer examination of these pages suggests that for the two travel questions, because of the strong internet commodity market surrounding travel, a considerable amount of effort has been spent on curating good travel resources. Even with the slightly more specific LA query, there were still two specialized sites dedicated to attraction for kids in LA (Mommypoppins.com and ScaryMommy.com). The VIM question represented a mismatch between our output and the question style. A number of the sources for the question were tutorials, however in the clipping phase, these ordered tutorials were broken up into unordered clips, creating an information model breakdown. This points out an interesting limitation in the KA approach, and suggests that adding support for more structured answers (e.g., including sequential steps) could be valuable future work.

As an additional external evaluation, for the two questions (Q6 and Q7) related to automotive systems we compared the discovered categories from the KA system with two commercial knowledge service products generated by expert technicians. We compared the KA response's accuracy and comprehensiveness, and found that it discovered all the categories referred to in these two commercial products for each question. Furthermore, the categories from the KA output provided more categories not mentioned in the commercial product (average 2.5 categories from two commercial products, while average 9.5 categories from KA). We validated these additional categories with expert automotive professionals who evaluated them as also being plausible and reasonable for the given questions. There was one instance in which two distinct categories (Encoder Motor and Encoder Motor Sensor) from the commercial products were clustered into the single

category named Encoder Motor Assembly in the KA output. However, the full text answer from the KA system for Encoder Motor Assembly did still contain these two sub-components with different repair procedures.

It may seem surprising that KA would work well for questions such as automotive error codes, where the response relies heavily on technical knowledge and jargon. On further inspection we believe this is because there are many online resources that have valuable information pertaining to these questions but are in unstructured and dialog oriented forms. Workers in the sourcing phase found rich sources of online information from many car enthusiast discussion forums, in which members tried to diagnose and help each other solve their automative problems. Although crowd workers may not understand the esoteric jargon of the automative domain, their understanding of grammar, semantics, and argument structure was sufficient to let them find, filter, cluster, integrate, and edit this domain-specific information. These results suggest a interesting avenue for future research leveraging human understanding of semantics and argument structure to extend crowdsourcing to process expert domain knowledge and to understand the limits of where such an approach breaks down.

On average, running a question through the KA system cost a total of $108.50 (Table 3.2). Although our primary goal was to establish a proof of concept of accomplish big picture thinking in small pieces, we return to the issue of cost in the Discussion. From the self-report crowdworker feedback, workers mostly found the tasks to be easy to complete, with the clustering phase having the most difficult task.

Phase	**Task Pay**	**Avg. # of Tasks**	**Avg. Cost**
Sourcing	$0.25	15	$3.75
Clipping	$0.50	21.6	$10.80
Clustering 1	$1.00	10	$10.00
Clustering 2	$1.00	10	$10.00
Integrate	$0.50	37.2	$18.60
Edit 1	$0.75	28.8	$21.60
Edit 2	$1.00	28.8	$28.80
Images	$0.50	9	$4.50
Total		160.4	$108.05

Table 3.2: Average number of worker tasks and average cost per phase, and overall, to run a question.

3.7 Discussion

The strong performance of the system is perhaps surprising given that its output was generated by many non-expert crowd workers, none of whom saw the big picture of the whole. We do not believe that this should be interpreted as a replacement for expert creation and curation of content. Instead, the power of the system may actually be attributable to the value created by those experts by generating content which the crowd workers could synthesize and structure into a coherent digest. This explanation suggests that the approach would be most valuable where experts generate a lot of valuable

information that is unstructured and redundant, such as the automative questions in which advice from car enthusiasts was spread across many unstructured discussion forums. In contrast, KA's output did not outperform top web sources for topics such as travel, where there are heavy incentives for experts to generate well structured content. We believe its performance is likely due to its aggregation of multiple expert viewpoints rather than particularly excellent writing or structure per se, though this is a fruitful area for future investigation.

In developing the KA system, we explored a number of approaches that did not work. We initially tried to avoid a clustering phase altogether by exploring variations of the clipping task in which we provided additional context to workers in having them read through multiple sources, engage the workers who found sources in doing the clipping, or have them build on the categories that other workers had already generated rather than work independently. However, in all cases workers did not generate good labels due to a lack of context. We then explored introducing an additional "conductor" view, in which workers could be recruited as clips came in to organize those clips and close categories that had a sufficient number of clips; however, this also failed because the conductors did not have sufficient global context to create good categories. These failures motivated the hybrid crowd-machine clustering phase.

Development of the integration and editing phases also included many false starts due to the opposite problem of giving workers *too much* context. Our first integration interface enabled multiple workers at the same time to easily view and expand all the clips in a category for within-category context, and also see the current state of how other categories were developing for between-category context. Our idea was that as workers integrated clips and built out more options, exposure to the other clips and options in real time would help them create more coherent digests. However, this approach proved overwhelming for scaling up to a large number of crowd workers engaged for short time periods. This motivated us to split up within-category and across-category consistency into the integration and editing phases and the development of the vote-edit pattern.

We encountered a number of places where our approach could be improved. As evidenced in the VIM question, the lack of support for nuanced structure in our digests can prove problematic. For some sources such as tutorials or how-tos, supporting sequential dependencies between steps could be useful. While our output was able to support such dependencies in an ad-hoc way within a category (such as the sequential steps for plunging a drain), it would be profitable to be able to support sequential dependencies across categories (e.g., first try x, then try y). More structure could also be beneficial for particular domain areas, such as explicitly capturing symptoms and causes as different types for automotive or medical diagnostic questions.

The system could also benefit from including iteration. For example, after workers completed the integration phase they were asked the question "What else needs to be done to make this a complete answer?". While many obviously said the section needed be edited, one of the most popular responses was "Needs more information." This suggested to us that while our clips and categories had pulled in most of the information, there was more information in some sections we were missing. One possibility is to introduce an iterative component at this point – as workers are integrating information into the pad and notice missing information, they can request for other workers to

go out and find that additional information through clipping. Thus while the system was partially successful at taking a breadth-oriented approach rather than the deeply iterative approach typical of sensemaking [56, 58, 181, 192], understanding how to best incorporate iteration would be a valuable area for future work.

Aside from improving the quality of the system output, there is also the possibility of reusing the output for other users researching similar questions. Although users have complex information seeking needs, many of the queries they issue are similar. For example, a recent study estimated that 3% of search queries account for 1/3 of total search volume [230]. Thus at a minimum, many answers could be amortized across users with the same question. A particularly promising but challenging opportunity is if similar questions may be able to reuse components of already summarized answers; for example, a question on investing advice for a 50 year old might use some common categories as for a 20 year old, but others would be unique to the new question's context. Challenges for the reuse of information are how the system would be able to identify the similarity for possible answers during each information synthesis phase and what level of granularity should be considered to for an effective system. Spatial and temporal reasoning over the existing knowledge and new information could be considered to provide context-aware and up-to-date answers.

We hope the design choices embodied in the KA prototype system and the design patterns discussed here may be useful for other system designers working to distribute cognitive complex tasks. Some domains that might benefit from this include micro-task markets, which could benefit from supporting more complex tasks; volunteer crowdsourcing efforts such as Wikipedia [121] or friendsourcing in which many small contributions are readily available [24]; or self-sourcing in which the crowd within could accomplish complex tasks in small increments (e.g., waiting for the bus) without needing to load the entire task context into working memory [215]. Overall, we believe this approach represents a step towards a future of big thinking in small packages, in which complex and interdependent cognitive processes can be scaled beyond individual cognitive limitations by distributing them across many individuals.

3.8 Individual Implications

While the KA system was successful for enabling crowdworkers to perform sensemaking with microtasks, the inherently individualist nature of sensemaking for many tasks makes it difficult to accomplish with crowd workers. Users often have very situational restrictions or preferences surrounding a task that can be difficult or even impossible to fully externalize to someone else performing the task (i.e. they have children, are allergic to something, or might have previous experience with part of the subject). The documents produced by the KA system are excellent as general purpose answers to questions, but might be too general or unnecessarily verbose for certain situations. However, KA had certain unique qualities that suggested opportunities to support the more individualized, personal sensemaking, mainly how though its workflow it enabled workers to externalize and pass along useful, refined data to other workers. Could we use the KA workflow as a lens for supporting individual sensemaking?

The KA workflow served as one of the foundations for the synthesized framework

(Figure 1.1) and the subsequent systems based on that framework. The individual phases of the KA workflow map directly to the framework:

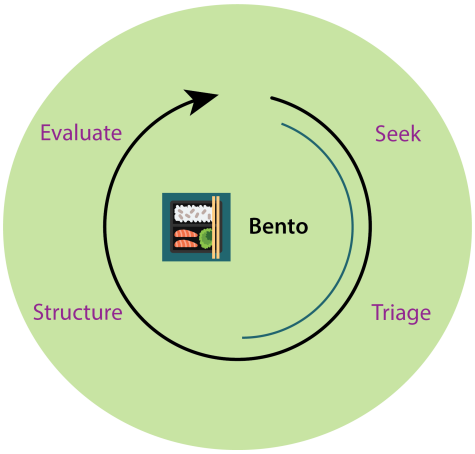
- Source Selection: Seek
- Clipping: Triage
- Clustering: Structure
- Article Generation: Evaluate

Using these phases as potential intervention points for improvement, we began to consider how we could structure each of those points in the process to better support individual sensemaking. Notably, we focused on enabling easier / streamlined externalization of user's mental models for each point in the process, with the aim of reducing cognitive burden and allowing individuals to work with and track larger amounts of data. The KA workflow pointed to potential units of data (websites, clips, categories) that could serve as the foundations for externalization in the different phases of the process. We began this exploration with the step in the process, finding sources of information (seek).

Chapter 4

Bento: Search and Triage [1]

Through the Knowledge Accelerator work, I was able to identify a reasonable workflow that was able to support distributed sensemaking across a group of crowdworkers. However, this workflow was able to work because all the individuals in the process were extrinsically motivated to do so. In an individual scenario, the overhead introduced by such workflow driven tools might not offer an easily perceivable intrinsic benefit over current methods. This makes it difficult for users to justify the adoption of digital tools, as the burden of using these tools can be significant, especially early on in the process where tasks might be simple enough to just work in memory. However, as a user's mental capabilities begin to become taxed or overwhelmed, they might have to resort to suboptimal satisficing due to their limited ability to process and manage a large amount of information.



By using the KA workflow as a guidepost for breaking apart the different activities and phases of the sensemaking process, I began to develop systems, that through optimized workflows and computation, allow users to manage and track the extraordinary amount of data they encounter during sensemaking. With these systems, I aim to give users a way to fluidly manage, structure, and evaluate the data collected during sensemaking in a natural way consistent with their current process. In this next set of chapters, I discuss a few different systems and tools I built to help provide this support for certain portions of the sensemaking process: Bento for helping to manage sources during the seeking and triage phases and Distil for streamlining source clipping and structure generation, and finally Meta which helps users collect and evaluate potential options from a variety of sources. The first system I discuss, Bento, looks at the processes of search and triage, and introduces several mechanisms to support those activities over multiple sensemaking sessions.

[1] Portions of this chapter previous appeared as Hahn, N., Chang, J. C., & Kittur, A. (2018, April). Bento browser: complex mobile search without tabs. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).

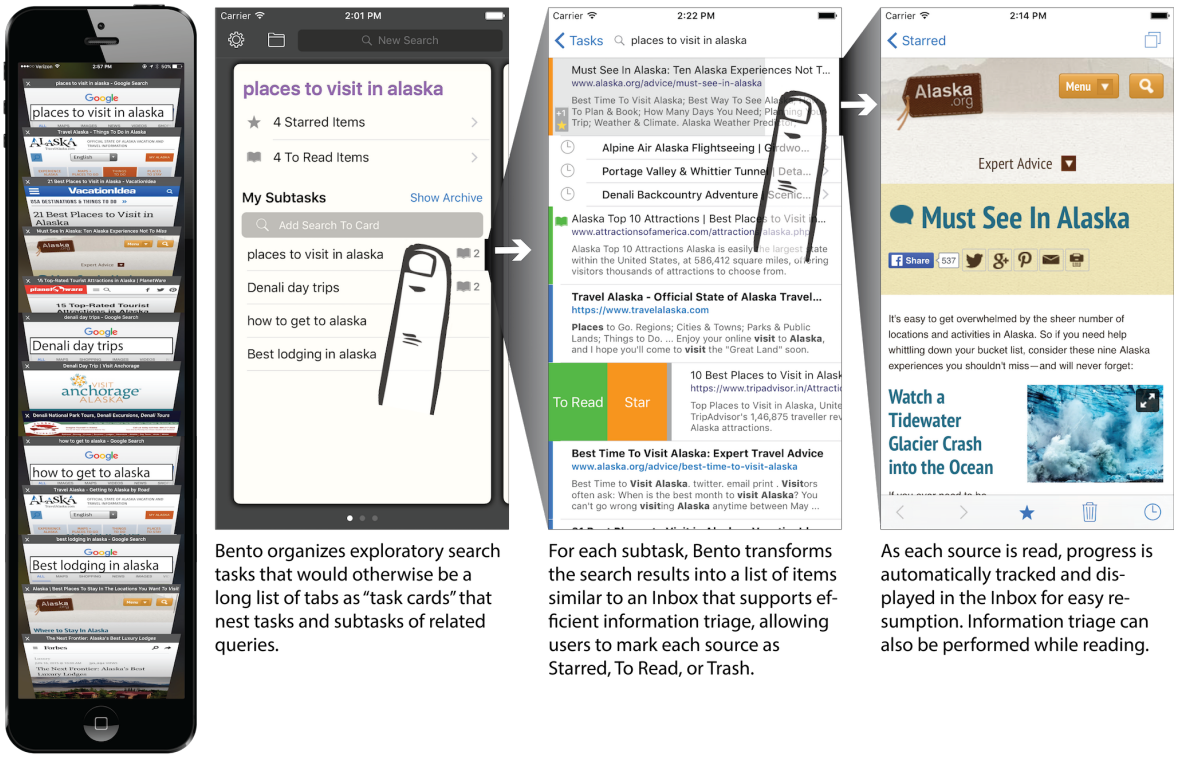
4.1 Introduction

As users begin their information foraging process, they often look to multiple sources of information to provide a clear picture of what the entire information space looks like: which options they can choose from, the full scope of a topical area, and all the different opinions around a subject [135, 150]. Due to the plethora of information, users divide their attention among these sources by going through a triage process - they perform a lightweight evaluation of the information available, and if deemed potentially useful, will either mark it for follow up, or dive deeper into it [18, 152]. In modern browser environments, this behavior largely occurs through the tabbed interface: users will queue up and manage potential sources of information. Consider a person planning a trip to Alaska: on a desktop they may create multiple tabs for each location or point of interest, which quickly multiplies as the user drills down into restaurants, hotels, and activities for each of those locations – potentially resulting in dozens of tabs open at once. Adding complexity to the situation, many of these foraging tasks may be going on in parallel (e.g., investigating alternate destinations such as Anchorage vs. Homer), may be suspended and resumed in various states of progress over time, and may be interleaved with other tasks (e.g., finding a place to eat tonight).

In this chapter I discuss Bento, and mobile interface for sources management. We performed this exploration using the practical and interesting design constraint of a mobile device. Mobile devices are used now more than ever for information seeking activity [61], however they are significantly smaller, are operated in short bursts of time, and activities are frequently interrupted on them [20]. Addressing sensemaking in a mobile device context thus is not only timely and important, but provides additional generative constraints for new approaches.

Through this work, we introduced an alternative approach to tabbed browsing for source management that also addresses the additional constraints involved in a mobile context. The key insight we built on is that tabs are often performing two distinct functions: 1) they serve as a way to organize and juggle multiple tasks that may be going on at once; and 2) they serve as a workspace to triage and build a mental model for a given task, for example queuing sources for later consumption, performing comparisons between sources and saving information of uncertain value for further review. Because tabs are overloaded in such a manner, we argue that they accomplish neither task very well, especially when used in a constrained mobile environment.

To overcome the limitations of tabs, we introduced a scaffolded process that separates the task management and workspace functions of tabs into two distinct interfaces. Instead of having many open tabs we transformed the search results page into a mutable workspace that allows users to triage and keep track of their progress on any given search, with those searches collected into tasks and subtasks. We instantiated this approach in a novel mobile web browser, Bento Browser, and evaluated its effectiveness through three user studies. Our results suggest opportunities for the development of novel systems of online information seeking for both mobile and desktop platforms that both better suit the nature of complex searching as well as constrained mobile environments.



Bento organizes exploratory search tasks that would otherwise be a long list of tabs as "task cards" that nest tasks and subtasks of related queries.

For each subtask, Bento transforms the search results into a list of items similar to an Inbox that supports efficient information triage, allowing users to mark each source as Starred, To Read, or Trash.

As each source is read, progress is automatically tracked and displayed in the Inbox for easy resumption. Information triage can also be performed while reading.

Figure 4.1: Comparing a typical list of tabs (left) with Bento's search centered navigation from the same exploratory search task.

4.1.1 Mobile Sensemaking

First, we wanted to gain a clearer picture of how users were using their mobile devices for exploratory search. To explore this we performed a short survey, first partially published in [42], with 164 smartphone users (98 Male, 66 Female, Age: $M = 32.29, SD = 8.72$) on Amazon's Mechanical Turk platform. We asked a series of questions about a user's exploratory searches, how often they perform them, what were some past searches, as well as the interface tools they use. Surprisingly, we found that people reported frequently conducting complex exploratory searches either partly (70%) or completely (45%) on their phones, ranging from planning a vacation to researching woodworking projects. However, 47% of the users also agreed with the statement that "It would be frustrating to do a complex search on a smartphone."

We asked participants a number of questions about their current habits, based on a 5 point Likert scale (Rarely – A Great Deal). When queried about which exploratory search activities they currently perform on their phones, the most common activity was simply "Reading web page" ($M = 4.03, SD = 0.89$). Text entry during search ($M = 3.59, SD = 1.14$) and keeping track of multiple pages ($M = 3.24, SD = 1.12$) were the next to most common activities, with saving web pages ($M = 2.40, SD = 1.17$) and collaborating ($M = 2.21, SD = 1.14$) being the two most uncommon activities.

We then asked about future support. 80% of participants agreed with the statement "I would find it valuable if smartphones had better interfaces for doing complex searches." Delving deeper into this question, at least 1/3 of the respondents reported extreme

difficulty (highest Likert rating) with "saving web pages", "keeping track and switching between pages" and "sharing findings with others." These suggest that the current browser interfaces on smartphones do not well support the constant context switching and task suspension present in exploratory search. Conversely, participants cited the advantage of being able to do searches "on the go" and the general "convenience" of smartphones. These results suggest users think there are significant problems with managing exploratory searches on smartphones, even though they currently do them, and would like to continue to do them. This suggests addressing complex searching in the mobile context may have both real world practical value as well as being a source of potentially generative design constraints that could also translate to less constrained device footprints such as the desktop.

4.2 Understanding Tabs

Tabs are a ubiquitous feature in every major web browser today, where they serve multiple functions ranging from organization to triage to reminding [66, 102]. In particular, they serve two primary functions in the exploratory search process. First, tabs provide task management functions – by separating out tasks [224], acting as a reminder to resume a task, and allowing for quick efficient switching between tasks [102]. Second, they provide a workspace for triaging sources [95], performing comparisons between sources [124], and saving good resources for further review.

We note three specific problems with the ways that tabs try to support these two function simultaneously. First, tabs are only loosely coupled to their generating activity. As a result, tabs during exploratory search become disconnected from their search results page, potentially causing negative effects such as users losing track of why they opened a tab, where they were in their task progression, and which pages belong to which tasks [66]. This is particularly problematic early in the exploratory search process as users are uncertain about the future value of the information contained within them [124]. Second, tabs are ordered based on the sequence in which they were opened and which tab spawned them, in order to keep them co-located to the other tabs in their task. This can become inconsistent as an organizational model as tabs are closed or opened in the middle of other tabs, and also misses an opportunity to provide more meaningful organizational structure, either for separating tasks or as a workspace. Lastly, tabs have limited context (e.g., a favicon and partial title) which can make it difficult to find a tab, know the state of progress on using it, or understand which tabs belong to which tasks. All three of these challenges are exacerbated in the mobile context, where there is little space to show multiple tabs at once or to provide context for them.

4.3 System Design

Seeing the issues surrounding tabs, we developed Bento, a novel interface for scaffolding complex search which obviates the use of tabs while still supporting their underlying task management and workspace functions of tabs [2]. This can be seen in see Figure

[2]See supplementary video Bento_Demo.mp4 for a demo of the first version of the system

4.1, which shows how a user might perform planning a trip to Alaska with the current paradigm of tabbed browsing on a mobile device versus Bento. The fundamental component enabling the approach is transforming the search results page in place into a mutable workspace that allows the user to queue pages to read (analogous to the common practice of opening a search result link in a new tab), star pages they found useful, trash unhelpful pages, and, critically, to see the progress they have already made in reading each page they opened (Figure 4.1). Unlike previous approaches (e.g., collaborative search [166], history management [12, 164], or activity workspaces [96, 114, 222]) which require a separate interface for managing and surfacing individual tabs, Bento's approach provides a natural centralized workspace in the search results page that is already a fundamental and familiar element of navigating complex searches and obviates the need for tabs altogether.

Of course, for complex searches a single search is often not enough; for example for planning a trip to Alaska one might have additional searches for day trip destinations, how to get there, and where to stay. For managing tasks and subtasks Bento bundles search result pages together into task cards, drawing inspiration from previous search-based task management tools such as SearchBar [164]. However, one difference from tools which focus on surfacing the past history of searches is the prospective nature of Bento's scaffolding, in which users can (and did) create searches as placeholders and reminders of subtasks they would need to work on (like finding a place to stay) before actually doing any of the work. Together, these elements suggest a radically different way for people to manage complex searching than traditional tabbed browsing. Below we describe the design rationale for developing Bento and details about its various interface elements.

4.3.1 A Sensemaking Workspace

When creating this workspace, we initially considered leveraging approaches utilized by previous information triage systems (e.g., [95, 201]). However, these tended to rely on spatial organization which were not a good fit for the limited real estate of smartphones. After a number of design iterations, we settled on a representation evoking the affordances of an email inbox. Email inboxes are designed for quick and efficient triage by users, providing information to users about what information is important, has been dealt with, and what still needs to be read / triaged. They accomplish this in a simple list format – not requiring the larger spatial requirements of other information triage systems. Email inboxes provide users with organization strategies ranging from flagging or starring items (which can pin them to the top of the list), archiving undesired items, and marking items to be read later (e.g., through marking as unread). We found these strategies useful for organizing searches, allowing search results to be flagged as important, archived if irrelevant or not needed in the future, marking items as potentially relevant and of interest to come back to, and supporting an awareness of where search results are in the list (in this case, ranked by relevance to the query if acted on, or in their original search result order if not). The items in the inbox are ordered with starred items at the top, followed by to read items, and finally any other search results in their original relevancy order. Trashed items are placed at the very bottom of the search results list to enable undo if needed. The search results

also have a natural progression of states: unviewed search results show up with bold text and a blue dot similar to an "unread" email message. Users can then manually mark a page as being in an intermediate state, with a 'to read later' annotation, or as a particularly useful reference source, with a 'star'.

We not only considered triage to be important in this interface, but also resumption and information provenance (Figure 4.2). Resumption is managed through a couple of factors: the search results persisting in appearance, read / unread indicators, and progress bars. Initially, we combined the progress indicator and read/unread indicator into the same space, however this caused a number of misinterpretations or disregard of the progress indicator all together. To increase visibility, we separated the progress indicator from the priority indicator (Figure 4.4). We instead represented the progress indicator as the background fill of each search result. As the user scrolled further down the page for a result and spent more time on that result, a gray bar would fill up the background of the row. In early piloting users found this to be much more intuitive and easily parsed. The read / unread indicator was adjusted to be a colored bar on the far left of the row. Additionally, when a user reopens a page, they are automatically scrolled to their last position on that page, letting them quickly resume what they were reading. We believe this novel approach of showing progression information directly on an information source gives users a way to understand progress without having to visit the source.

Lastly, to maintain information provenance, all subsequent pages visited from a search result are associated with that result. In a normal, tabbed environment, there is no obvious connection from a new tab to a previous page. Even the tacit relationship of being next to each other can be broken if tabs are opened in between. In Bento, there is a fundamental connection between each page and the search result it was opened from. If a user is reading a page that is a deep link from the search result, and they return to the list of results, that page is surfaced under the result with a small clock icon, representing it was a page the user was reading and paused. Similarly, if a user stars, or marks a subsequent page as to read later, it appears in the search results list under its parent search result (Figure 4.4). This provides an easy way for a user to resume their progress even from a page nested deeply within a search result.

4.3.2 Managing Sensemaking Tasks

Bento not only assists with managing the information from one searching task, but also gives users a way to juggle multiple information seeking tasks at once. Bento features a separate, second interface for managing the higher level tasks users are working on, and their multiple sub-components. Noting how users utilize tabs, this management interface needed to allow for quick switching between tasks, act as a reminder to resume a task, and create separate workspaces for each task. Inspired by the previous work [96, 114, 222], we decided to utilize an activity-centric management interface based on a task-subtask hierarchy, with the search queries acting as the task unit [164]. However, instead of using this hierarchy as a post-hoc way of revisiting and organizing tasks, we have users actively build their tasks in this structure.

The tasks and subtasks are organized into cards, designed to give the user a quick overview of the current status of their sensemaking activities (Figure 4.3). In order

Figure 4.2: The different manipulations that can be applied to a search result

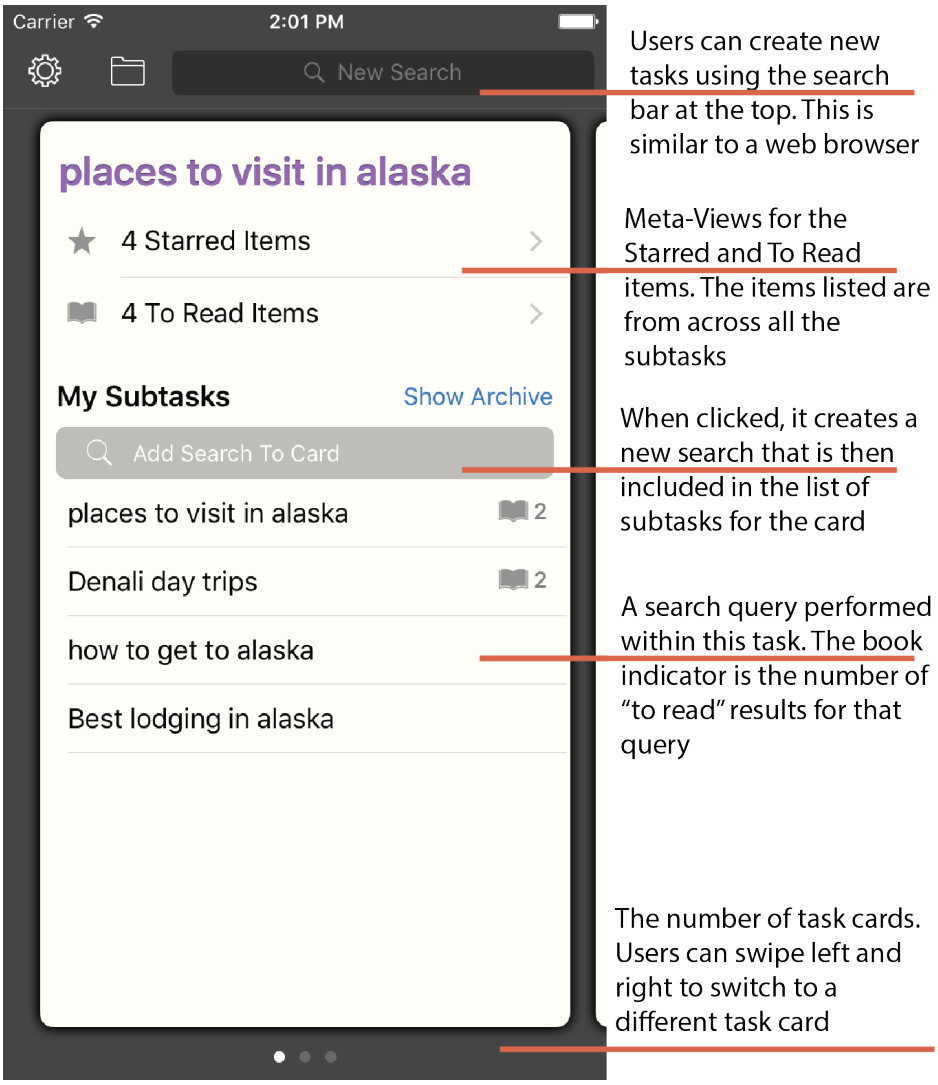


Figure 4.3: The task screen for the task "Places to go to Alaska"

to make the structure as lightweight as possible, when a user creates a search based on a new topic of research, we create a new task for them, naming it the title of their search. If a user adds a search to an existing task, a subtask appears under the task, with the query as the subtask's name. These task cards allow users to actively switch between tasks using a simple swipe, and their isolated "card" presentation provides the user a clear and present workspace for organizing, and quickly resuming their tasks activities. This is dissimilar from a tab environment, where everything is presented in a flat, undifferentiated structure.

To assist with the tedium of reorganizing tasks, the cards are automatically reordered based on recency – similar to adaptive human memory [14]. When an activity is performed in a task, it is pushed to the top of the list, and older ones never resumed slowly fall to the bottom. The subtasks within a task card follow a similar ordering. More recent queries, shown towards the top, serve as both a reminder for users about subtasks they need to complete, and allows users to orient their work chronologically as their understanding of the information changes. The temporal organization also allows users to scroll down their list of searches so that they can gain a retrospective understanding of how their mental model has evolved over time and restore the context they had when they stopped the search. This structure removes the ambiguity from tab ordering, creating a consistent interaction for later resumption [174].

To create a new task, a user types in a new information need (query) into the top search bar. Initially, we required users to create a task card, and then used it's naming as the information query. However, a normal browser provides a one touch experience in order to create a new search, with a large target for initiating the search. Recognizing that our initial approach broke the user's mental model of searching, we modified the workflow to use a more traditional search bar. The search provides results / suggestions for existing cards, subtasks, as well as normal autocomplete results.

We utilized a similar approach for creating subtasks (or sub queries). Each task card features a separate search button titled 'Add Search To Card' modeled in the appearance of a normal search bar. This reduced the cost for creating a new search to a single tap on a large salient target, in comparison to needing to tap a small button to create a card then subsequently search. On the task card, we provide rich information about the status of the individual subtasks. Beside each subtask, we note the number of starred and to read items from the search results. Besides the ordering of the subtasks, these provide the user with information on the completion of each subtask, as well as the usefulness of that particular information query.

Initially, we required users open up their individual subtasks to view important saved information, or to make progress. After the first study, one participant noted that "it would be nice ... to see all of my starred items in one place for easy reference." As a result, we added summarization lists on the to-do list card view (Figure 4.4). These summarization lists allowed a user to immediately look at all of their to read results and starred results across all the searches in a task. The "to read" summarization list became a reading list for the task, while the starred summarization list as a collection of the most important information an individual had collected for the task. These views serve as a way for users to get an understanding of either the important information they have collector for a task quickly, or what information they need to process next in a task.

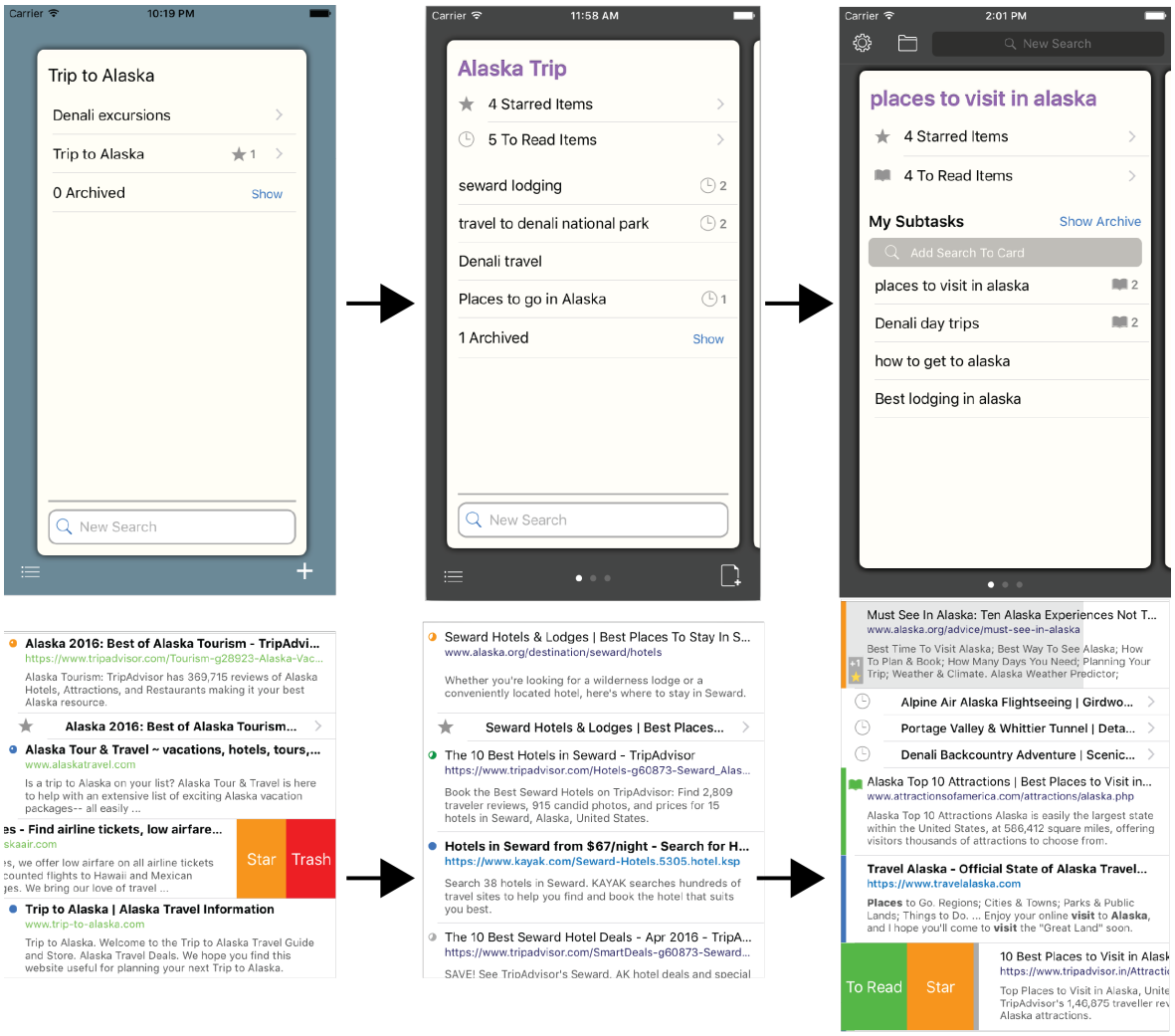


Figure 4.4: The progression of the Bento Brower design. From left to right: Study 1, Study 2, Study 3

4.4 Implementation

The Bento Browser application was built for the iOS platform and was available to participants running iOS 8.0 and above. The application utilizes Google's Firebase real-time database and analytics platform to collect telementry data from users. We utilize the Bing API to fetch search results for queries made by users.

4.5 Evaluation

We completed three studies to provide converging evidence on the value of our approach: a lab evaluation, a qualitative real world deployment, and an expanded quantitative deployment. Between studies we used the feedback to iterate and refine the design of the system. We explore whether our dual interfaces of task management and an information triage workspace are able to more effectively accomplish what tabs try to. We specifically look at the pressure points caused by a tabbed interface: organization of tasks, a workspace for information processing and sufficient context for quick resumption of activities. In each of these studies, we optimized the design to promote maximum motivation to actually work on the complex searching tasks by having users work on their own tasks. Rather than trying to evaluate an outcome from a fixed search task, we wanted to capture how individuals found Bento to be useful for a variety of complex searching tasks.

4.5.1 Study 1 - Understanding Triage

The goal of the lab study was to evaluate our approach while controlling for differences in the complexity and nature of the tasks that users engaged in, which would otherwise vary in a field trial. We focused on the triage interface in this study, having users only work on one task while in the lab. It also provided an opportunity to get feedback and iterate on the system's features that might otherwise cause critical issues in a lengthier deployment.

In order to make the lab study task realistic and providing internal motivation, we collected multiple real search tasks from each participant and randomly assigned one to the Bento condition and another to an environmentally valid control, for which we chose the Safari mobile web browser (the default tabbed browser on the iPhone). The study was a within-subjects design in which participants used both browsers (counterbalanced across participant) and provided feedback on them.

Procedure

We recruited 22 participants through a local behavioral research participant pool. Participants ranged in age from 20 to 59, with the majority of participants being local undergraduate and graduate students. 10 participants identified as male and 12 as female. We required that study participants own and use an iPhone to ensure they would be familiar with the existing iPhone operating system and Safari browser. All participants were provided with an iPhone 6s with Bento preloaded onto for use during the study.

Since a single fixed search task might not provide internal motivation to every participant [134], we instead elicited four search tasks of potential interest from the participants themselves during a prescreen. We selected two of these searches based on how participants rated the topics across 4 scales: their knowledge of the topic, the importance of the topic, the expected research time to learn about the topic, and the estimated number of web pages they would have to visit to fulfill their information need. To select one of their proposed searches, we required participants mark it as at least moderately important, have less than a moderate amount of knowledge about the topic, the search would take at least several hours, and the search would require least 8 different web pages. For each participant, we selected two searches that met the criteria (if more than 2 met the criteria, we chose those with the highest values) and randomly assigned them to either the Safari condition or the Bento Browser condition. Some example searches include "How to create an Android application" and "Advice on how to enjoy being a tourist in Japan."

Participants were asked to search for 20 minutes using either Bento or Safari, then to switch to the other search with the other tool (with order of browser counterbalanced across participants). For the Bento condition, before they began they were provided with a brief tutorial that walked them through the interface and features; all participants were already highly familiar with Safari from their own phone use. After completing both searching tasks participants completed a post-survey about their experiences. The survey asks participants to directly compare their experience with the Bento tool to the Safari mobile browser, as well as review some of the features of the Bento tool.

Results

Overall, we found that participants appreciated the Bento interface, finding that compared to Safari, it helped keep them significantly more organized ($M = 4.25$, 95% $CI[3.91, 4.59]$), and would be more useful for helping them restart where they left off ($M = 4.15$, 95% $CI[3.66, 4.64]$). Despite participants' high familiarity with Safari, we did not find significant differences in the ease of search creation using our tool, nor did individuals feel less effective using it. Participant did note that Safari was much easier to learn (85% of the participants stated this), however 70% thought that Bento was more helpful in finding pages. This was especially notable considering the prototype status of the system during Study 1 and the additional steps individuals had to go through in order to make and organize their searches.

The comparison was also well supported by feedback on the individual features of the search. On average, participants reported in our post-survey (using 7-point Likert scales) that they enjoyed the software and the features provided by it ($M = 4.95$, 95% $CI[4.18, 5.71]$). They thought that Bento Browser amplified their search effectiveness on a mobile phone ($M = 5.25$, 95% $CI[4.49, 6.01]$), they felt confident searching using the tool ($M = 5.05, 95\%$ $CI[4.30, 5.80]$), and they reported wanting a tool like Bento Browser for searching on their mobile phones ($M = 5.15, 95\%$ $CI[4.37, 5.93]$).

Of all the features, participants found starring pages to be the most useful tool (over 90% reported starring being moderately useful). When asked more about this, they noted that starring pages made it "incredibly easy to save pages" and in general "it was easy to collect a large amount of relevant webpages to read and delete the irrelevant

Question	Study 1 Mean	Study 1 CI	Study 2 Mean	Study 2 CI	Study 3 Mean	Study 3 CI
Which tool did you like better	3.15	[2.45, 3.85]	3.125	[2.18, 4.06]	3.01	[1.94, 3.89]
Which one was easier to create new searches in?	3.4	[2.82, 3.98]	3.126	[1.99, 4.26]	3.38	[2.76, 3.99]
If you wanted to keep searching later, which tool would be better for picking up where you left off?	4.15*	[3.66, 4.64]	4.25*	[3.38, 5.12]	4.44*	[4.05, 4.83]
Which tool makes you feel more at peace?	2.9	[2.16, 3.64]	2.63	[2.01, 3.25]	2.69	[2.05, 3.32]
Which tool makes your information more organized?	4.25*	[3.91, 4.59]	4.13*	[3.43, 4.82]	4.25*	[3.89, 4.61]
I felt more effective using:	3.2	[2.56, 3.84]	3.125	[2.18, 4.06]	3.01	[1.94, 3.89]
It was easier to refind information with:	3.47	[2.96, 3.99]	4.13*	[3.30, 4.95]	3.31	[2.65, 3.98]
I felt more confident that I didn't miss any important sources of information with:	3.0	[2.39, 3.61]	3.38	[1.96, 4.78]	2.53	[1.89, 3.31]
* Significantly different based on 95% Confidence Interval						

Table 4.1: The direct comparison questions were asked on a 5-point likert scale. A higher score indicates preference for Bento Browser, while a lower score indicates preference for the Safari browser. A score of 3 indicated no preference for one over the other. The 95% confidence interval (CI) is shown next to the mean. Any question where the lower bound (left number) is higher than 3 indicates Bento is significantly better. Any question where the higher bound (right number) is lower than 3 indicates Bento is significantly worse. This table covers Studies 1, 2, and 3.

ones." This suggests that the triage interface made it easy to quickly sort through the search results, and persist the important information for later use. A participant directly agreed with this, stating "... I could refind my pages for future viewing. This is very useful for searches that I am more likely to come back to."

4.5.2 Study 2 - Task Management

We iterated on the initial version of Bento based on the feedback from the previous study. Several participants noted that the interface was "clunky compared to the web browser" and it needed to be more attractive. A few others were confused by some of the interactions, such as what happens when they trashed a search result. From the qualitative feedback participants provided in the lab post-study questionnaire, we worked on three areas for improvement: visual attractiveness, better feed-forward and feed-back cues, and the summarization views.

In order to better evaluate the utility of the iterated version of Bento in a more ecologically valid setting we conducted an exploratory field study, in which participants used Bento daily for a period of 4 to 6 days.

Procedure

We recruited 8 participants from the same local participation pool as in Study 1. We required that participants own an iPhone with iOS 8.0 or above installed and had not participated in the previous study. Participants ranged in age from 18 to 24, with four identifying as male and four as female.

Individuals installed the Bento Browser application on their personal mobile device in the lab, completed a short tutorial, then spent 15 minutes working on a search of their choice in the lab so that they could ask questions and get used to the tool. They were then were instructed to use it for at least 10 minutes each day. The application provided a reminder three times a day if the individual had not yet used it for 10 minutes that day. Aside from the time requirement, we did not instruct the individuals to utilize the application in any particular way. We were interested in knowing how individuals used the different features of the application, and which features were the most useful to each individual.

After a 4 to 6 day period, participants returned to the lab for an interview and to complete a post-survey. During the interviews, we asked participants to walk through their usage of the application, showing off any of the concrete tasks that they did, as well as exploring their individual queries. This probe was designed to help users ground their experience of using the app in the specific tasks that they were performing. After the interview, participants completed the same post-study questionnaire as in Study 1.

Results

Post-survey results were very similar to the results from the lab study, with participants significantly preferring Bento over Safari for the questions "If you wanted to keep searching later, which tool would be better for picking up where you left off?" ($M = 4.25, 95\%\ CI[3.38, 5.11]$) and "Which tool makes your information more organized?"

($M = 4.125, 95\%\ CI[3.43, 4.82]$). We utilized confidence intervals in our analysis of these numbers, due to individuals directly comparing Bento and Safari to each other in each question. If the range on the confidence interval was above or below three, that would indicate a significant preference for Bento or Safari respectively. Additionally, participants also preferred Bento for the question, "It was easier to refind information with (Bento Browser)" ($M = 4.125, 95\%\ CI[3.30, 4.95]$) in favor of Bento. No other questions showed significant differences. Feedback about Bento was similar to the lab study survey, except more individuals cited a desire for a desktop companion ($M = 5.25, 95\%\ CI[3.72, 6.78]$), suggesting that additional usage in different contexts incurred the desire to switch between devices with different characteristics.

The interviews provided further insight into how individuals used the application for their own needs. Participants used Bento for a variety of exploratory tasks, ranging from learning about gardening techniques to product comparison to learning about political candidates. Several of the participants brought up Bento's value in capturing their mental model during the search process, which helped them get an overview of their search and suspend and resume it more easily. P7 specifically noted that you could "see everything that you Googled ... in a straight sequence." and the different triage lists let you "archive what you were thinking about in a single moment ... it was like a screen shot of what you were thinking about." P5 mentioned that he enjoyed "just being able to quickly look at the task list and know what to do next."

Perhaps the most important value perceived by participants was in how Bento structured searches into organized workspaces in which they could make progress. Organizing searches into tasks and recording searches as subtasks were rated highly in the survey (5.5 and 5.9, respectively). This led to some unexpected benefits, such as one participant noting "how easy it is to compare prices this way rather than with a traditional browser." Participants seemed to actively want to keep their subtasks organized, with 6 of 8 mentioning that they enjoyed utilizing the "trash" feature to throw out irrelevant results. We found this interesting because eliciting explicit feedback from users about search results is traditionally challenging, as users could just skip over the search result without having to put in the extra effort to trash it. One interpretation of this is that when users perceive the search results screen as a workspace rather than simply a launching pad they are more willing to invest effort into personalizing it.

Consistent with this, all participants utilized either the starring feature or the to-read functionality. Some of them (P2, P3) indicated that they weren't sure what the point of the to-read functionality was, since they would just immediately read a web page and star it if it was good. In contrast, P1 thought that the "to-read" functionality was one of the most useful features of the application. P1 cited that the feature allowed her to "cue up what she wanted to do next", effectively creating a future list of information to absorb. In a similar vein, P4 created several subtasks at once, and didn't visit them immediately. This allowed her to "just record all of the things she might need to think about for her trip ahead of time, and then just come back to them later." This prospective task encoding was a unique and unexpected benefit of the ability to structure sets of searches together.

Finally, transforming the search results into a workspace made some participants feel a sense of stability and organization; P5 specifically noted that he "like[d] that the results froze from when you went to them" unlike when you traditionally requery

a search. These results suggest that a key benefit of the approach was being able to organize and evolve their mental model through a relatively stable workspace.

Participants were also queried about how the mobile form factor of the application either enhanced or detracted from their experience. All of the users (6) who noted something positive about Bento cited the convenience and portability of the application. For example, P3 noted "The ability to search whenever I wanted to. ie waiting in line for something." Two of the users didn't cite anything positive, saying that they preferred larger screens and physical keyboards.

We also noted a number of challenges that users faced with Bento. Some were relatively straightforward, such as confusion around the progress indicator, leading to redesigns for Study 3. However, some were more substantive issues for the general approach. Some participants found Bento useful for complex searches, but overly high overhead for simple informational searches, suggesting that they would like "being able to toggle on/off the organizing part" or "not hav[ing] to create a new task for simple searches which would not require detail planning and organization." Another common complaint about mobile phone searching more generally was the difficulty of typing, e.g., "typing on a phone screen can be arduous." Exploring the tension between low overhead for simple searches and supporting complex searches – especially when the former can transform into the latter – may be a fruitful area for further research.

4.5.3 Study 3 - Behavioral Traces

Studies 1 and 2 provide converging evidence about the value of Bento's two interfaces of task management and a triage workspace. However, although the field trial in Study 2 provides suggestive evidence and scenarios of how participants used Bento, one weakness is that it relies on self-report data which could be biased or incomplete. In order to collect richer quantitative data about Bento's usage in the field we conducted a third study in which we instrumented the browser with data collection capabilities and analyzed participants' actual usage data. This also gave us an opportunity to iterate on the design to address the issues discovered in Study 2, e.g., confusion around the progress indicator and lack of support for quick, simple searches.

Again, with the previous study, we performed some modifications to Bento's appearance based on feedback. We focused on improving the readability of the search results and improving the learnability and "first use" experience. We introduced a more coherent first use scenario, adjusted the progress indicators on the search results screen to their final form, and modified task and subtask creation.

Procedure

Study 3 followed the same procedure as Study 2 but with the updated Bento application, with more participants, and for a longer period of time (10-13 days). Utilizing the local participation pool, we recruited 16 participants with ages ranging from 18 to 25. Participants who participated in the previous field study or lab study were not eligible to participate. Five participants identified as male, and 11 as female. Twelve of the participants were undergraduate students, while 4 of them were graduate students.

Afterwards participants completed the same interview as the first field study, and completed a slightly extended version of the questionnaire.

Results

Survey results and feedback were overall similar to the first field and the lab study, with significant preference towards Bento for the two questions: "If you wanted to keep searching later, which tool would be better for picking up where you left off?" ($M = 4.44, 95\% \; CI[4.05, 4.83]$) and "Which tool makes your information more organized?" ($M = 4.25, 95\% \; CI[3.89, 4.61]$).

The key research question for Study 3 was quantitatively investigating whether participants were in fact managing complex searches and utilizing the different features of Bento. Each individual created on average 13 tasks (M= 13.06, SD= 8.433) with on average 3 subtasks (M= 3.13, SD= 1.48), suggesting that participants were indeed engaged in complex searches with multiple subtasks. There was high variability between users, with some users having as many as 14 subtasks within one task. Drilling down further, for each subtask participants opened an average of 7.7 pages (M= 7.7, SD=5.41), suggesting that they were engaging in tasks that involved significant exploration. To check this against participants' own perceptions we asked them to classify their searches as either complex or simple when they came back into the lab at the end of the study. Participants classified 35% of their tasks as complex searches, suggesting that they were engaged in complex searches but also using the system for simple searches, addressing an issue raised in Study 2. Participants found value in Bento's organization and resumption capabilities for complex searches including researching "fandom", bus routes, and radiation oncology internships. For example, one participant explicitly mentioned "I learned the sort of tasks that bento is good for – [it] requires several (subtask) searches. for e.g. transferring money internationally there's wire transfers, exchange rates, foreign check processing fees, different charges for diff banks."

The mobile form factor in this longer study offered some surprising and unexpected benefits for the sensemaking process. One user mentioned that the mobile versions of web pages were actually easier to parse: "Many result pages are mobile optimised, such that the content delivered may be more condensed and the design of the webpage more minimalistic." Another user cited a scenario where it is actually impossible to have a laptop – cooking in the kitchen. In this case, her mobile phone is the only tool she can use to perform sensemaking: "When I'm cooking and I have a recipe loaded, I will prop up my phone on the counter. My laptop would take up too much space."

Participants consistently used many of the features of Bento. Individuals reopened subtasks on average about 2.2 times, starred 7.05% of pages visited, marked 5.84% as to read, and trashed 4.24% of results (note only 20 results were loaded at the time of the search). Each individual had approximately 23 sessions over the study period, so about 2 application sessions per day. When asked what feature they liked most, participants mentioned the organization of tasks and subtasks (9); being able to come back to searches (3); marking pages to come back to later (2); the gray background progress bar (2); starring pages (1); and the overall design (1). When asked what they would like to change most there were a large variety of suggestions, most having to do with not having the features of a full search engine like Google they were familiar with

(e.g., access to google scholar or images, answering questions directly after a search, having better search results). Two participants mentioned "quick search" as desired, suggesting there may still be a need to support simple searches more easily than in the current approach.

There was high variability around the use of features and types of searches participants engaged in, with some focusing on simpler searches and some more complex ones. To examine whether the type of search affected perceptions of the tool, we correlated the ratio of complex:simple searches with perceptions of the tool from survey responses. Those with a higher number of complex searches:

- Liked Bento better than a mobile web browser ($r(14) = 0.638, p < 0.01$)
- Felt more at peace using Bento ($r(14) = 0.71, p < 0.01$)
- Felt more organized using Bento ($r(14) = 0.55, p < 0.05$)
- Felt more effective with Bento ($r(14) = 0.834, p < 0.01$)
- Wanted to keep Bento on their phones ($r(14) = 0.644, p < 0.01$)
- Felt Bento improved their effectiveness on mobile phones ($r(14) = 0.65, p < 0.01$)

4.5.4 Summary

Across the above three studies, we found evidence that users appreciated both the task based organization interface, as well as the search results workspace interface. Together, these consistently made users feel more organized and feel like they could resume their activities more easily. Based on these findings, we also have a couple of additional takeaways.

Users appeared to use a few strategies, aligned with many goal activation theory approaches [10]. For example, in study 2 P4 noted that she queued up searches for later exploration, largely a prospective planning task. Conversely, we had individuals like P2 and P3 who didn't really understand the to-read feature, another planning tool we had incorporated into our design. This suggests that different user populations might practice different planning techniques for their exploratory searches, and while the structure appeared to be amenable to most of them, having tools for both planning and retrospective recounting could be essential to the design of these systems. This information was similar to what was found in Study 3 with user preference for different features of the system. Most users liked the overall organization for easy re-finding, however some users liked the specific planning features, such as the to-read feature and the grey progress bars. Supporting both of these resumption use cases will be key proceeding forward.

We had several complaints from users about the overhead of Bento for simpler searches. In both studies 2 and 3, individuals noted that they wish they didn't have to make an entire task card for just simple searches. However, in some of our interviews, individuals noted that their simple searches, such as looking up an actor in a movie, often blossomed into more complex searches, such as looking at what else that actor was in, what roles they typically play, etc. Having a low overhead, while also supporting this transition of simple search into complex search was an issue in Bento that was not entirely resolved.

4.6 Discussion

With Bento, we introduced a novel way to manage complex searching tasks on mobile devices. Bento creates a scaffolded process that separates the task management and workspace functions of tabs into two distinct interfaces. Its two focused interfaces provide users with the necessary affordances to make progress on their sensemaking tasks even within the constraints of mobile devices. This structure is able to meet the complex demands of sensemaking and mobile work, and can be used for later transfer, hand-off, and resumption.

From Bento, we were able to discover a several important benefits regarding a scaffolded task-based structure. First, functional units that can be leveraged in future work. For example, consider the goal of making sensemaking independent of person. The tasks present in Bento could serve as the key unit of collaboration. Since they represent a specific, independent information goal, a task could be shared and worked on by a team of individuals. Subtasks could be delegated out to certain individuals to explore, and because all of the information is tracked, mechanisms such as the star feature could be expended into a voting feature. Similarly, the task could be handed-off to another person. The details about which pages were found important and which queries were used could provide a valuable starting point to another individual researching the same topic [70].

Along another line, the structure could be used to allow for easy transfer and resumption from other computing systems. Indeed, one participant (P6) noted that he wished there was a web based version. He utilized a number of different devices, many of them not his own (such as those provided by his university). Having this tool available on any platform would let him pickup his searching or find some information that he needed. For example, a similar approach could be instantiated as a virtually identical desktop browser that syncs with the mobile version. However, moving to the desktop may provide other design opportunities given the additional screen real estate; for example, the three level hierarchy (task > subtask > page) of Bento on a smartphone might be flattened to two levels (e.g., task cards and a subtask pane of search results with a reading list, similar to an email application) or even a single level (by incorporating the task card into the view). These changes would keep the integrated tab management and exploratory search support of Bento while being a more efficient way of reading and exploring. Better support for text entry and annotation on desktops could also benefit a future version of Bento on the desktop.

Finally, it is possible that using an approach such as Bento may change the strategies that people use in search [162]. For example, although we expected users to add subtasks as they encountered new information, one participant found it useful to do the opposite: "my searches were more focused because i tended to brainstorm at the start of a task and added subtasks at that time." Essentially, the list of searches in Bento were serving not only as a history list, but also as a list of ToDos that needed to be accomplished. While we didn't initially design for this behavior, it ends up being highly congruent with the synthesized sensemaking model (Figure 1.1).

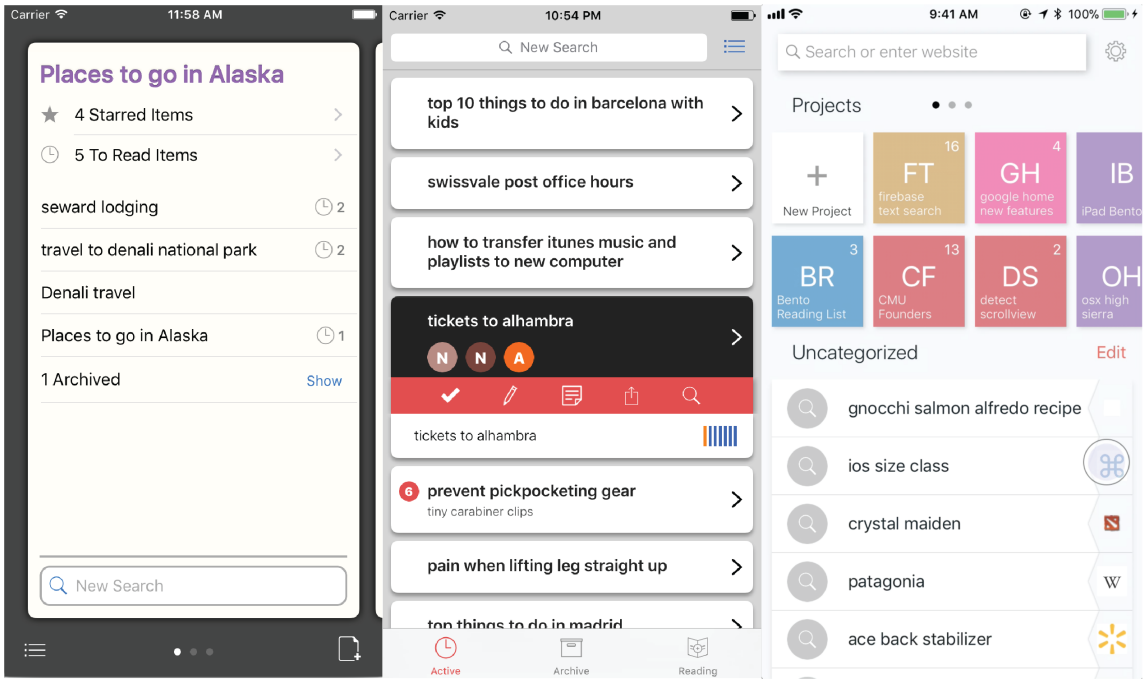


Figure 4.5: Three different iterations of Bento. The left-most was the final iteration tested in the above studies. The middle version is an intermediate prototype that allows users to stack searches together to form a project. The right-most version is the final, deployed version of Bento where users see all of their recent, uncategorized searches in the bottom half, which they can drag into either existing or a new project in the top half

4.7 Bento Iterations

While we only tested the initial version of Bento to gauge the effectiveness of a search-based task scaffold for source foraging, we realized that the interface had limitations. These centered around supporting simpler information seeking tasks, and their transition to sensemaking tasks in a more fluid manner. We went through several additional iterations of the interface (Figure 4.5) in order to refine some additional interactions in support of this: post-hoc task creation, a separate queue of recent simple searches, and quick access to the last search result page viewed for a search.

Rather than requiring users to create task cards up front in order to perform and track searches, users could create a new search instantly from the homepage of Bento. These searches would then be tracked in a "recent searches" list, which would be shown below the list of projects the user had. Users could then take one of these searches and drag them into a task in order to persist them there, while also giving them some more task-oriented features present in the original version of Bento. This post-hoc task generation was designed to further reduce the barrier required to start a simple search task — one of the most common browser interactions — but still allow users to transition those tasks from simple searches into a more complicated project with

progress tracking.

Lastly, recognizing that simple searches are often used as either purely navigational or to answer basic questions, we instead introduced a button on the right hand side of these search queries that would take you back to the most recent page you've visited for that search. This would give users the ability to quickly resume a search where they left off, or quickly access previously found information with a single tap.

4.8 Takeaways

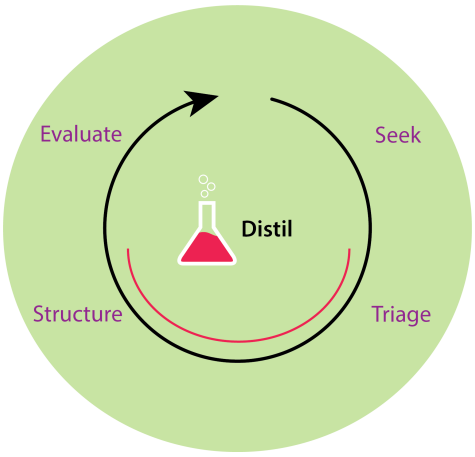
With Bento, we were able to give end users the ability to manage the large numbers of websites they visit during sensemaking through the creation of search-based tasks, and the direct triage of the search results. From this system, there were several generalize innovations that could be adapted to other sensemaking tools: creating task structures from searches, allowing for post-hoc task creation, and reusing attention signals, like scroll position, to support the user. The search-centric task structures could be adapted to any system that supports managing information sources, such as tab managers or bookmarking tools. Similarly, post-hoc task creation, could be applied to these as well – tabs and searches could be monitored, and based on visitation history, end-user task groupings could be suggested. Lastly, while Bento utilized fairly basic attention signals, future tools could begin to leverage cursor position, amount of time a paragraph was in the viewport, as well as length of visit to further support end-users in managing and navigating the wealth of information they've consumed.

Chapter 5

Distil: Extracting and Structuring Information [1]

While Bento gave users the ability to triage information at the source level, users are unable to effectively break out, triage, and categorize information from a source. Throughout the sensemaking process, users have varying levels of uncertainty [41], where early on users are trying to gain a sense of the information space, and its features, versus later on where specific detailed information becomes critical to decision making. Therefore, tools need to be able to support a fluid transitions between collecting and structuring high level, overview information and smaller, detail oriented information.

To tackle the first part of this issue, I developed the Siphon toolkit, which allows users to use a variety of selection interactions to annotate and extract information ranging from an entire page to a specific word. Uniquely, Siphon also maps any selection to the underlying HTML of the webpage, which allows the toolkit to pass along the underlying content, highlight and maintain a connection to the selected content, and rerender the selection in other contexts. I discuss this toolkit briefly in the first part of this chapter, noting the specific capabilities that support the structuring of data extracted for use in sensemaking.



This toolkit is then utilized in Distil, which aims to tackle the second issue of managing information in an ever-changing landscape. In Distil, users are able to leverage interactive "smart categories", where they can define auto-updating categorizations that automatically pull in relevant existing and new information. These smart categories allow users to more efficiently perform the processes of classification and schema induction on their collected data through streamlined categorization. With Distil, users were able to quickly create and adjust their categorizations, using them to both more deeply explore the dataset as well as organize it. In the next part of this chapter, I discuss Distil, and how it lets users

[1] The Distil portion of this chapter is currently under review for publication

generate quicker and more flexible categorizations that are easier to change and discard during the ephemeral sensemaking process.

5.1 Introduction

A key challenge when individuals are trying to make sense of an unfamiliar domain is that they don't know the complete breadth of the information space when they start. They encounter information serially by sampling web pages, foraging through them for useful information and simultaneously learning the structure of the space. For example, if someone is planning a vacation to an unfamiliar destination, they may not know which are the good locations to visit, or even the types of locations they should visit (beaches versus museums, etc.). Someone researching cancer treatments might not know the different possible side effects of radiation, or that there are different options for radiation, and that information might be scattered across many different websites or forums. An analogy for this process might be a person on an alien planet trying to classify lifeforms but only encountering them one by one, seeing new ones that seem related to old ones but also different in certain ways, and trying to identify meaningful and discriminative features to separate them into useful categories. Similarly, the process of coming to understand an information space is messy, requiring frequent refactoring, throwing out old or irrelevant information, making decisions, and distilling often verbose content into a more consumable form [192].

As users encounter information they want to save and utilize during this process, it might come in many different shapes and sizes. Depending on where a user is in the process, the amount and context of the information they need to save can be very different [41]. While efforts have focused on improving content selection [42, 90, 191, 237], simplifying or enhancing information extraction [63, 101, 208, 209] and supporting content organization [35, 104, 245], all of these improvements have occurred in silos, where tools developed for enhanced extraction or organization can't rely on novel interaction techniques for selection. These tools then require users to rely on the inbuilt browser selection tool, which is limited and not suited for many situations [41]. In the first portion of this chapter, I discuss a toolkit I developed, Siphon, that assists developers with developing and utilizing custom document selection and annotation tools. These annotations are then available in a consistent, traceable format that can be used in other tools, such as a data extraction or content organization tools. They are utilized in the next tool, Distil, as richer primitives for organization.

Second, the incremental, mutating nature of the structure of a user's mental model poses challenges for existing tools that aim to help users organize their collected sensemaking data. For example, one major class of sensemaking tools focus on providing tools for users to easily manually create structured representations of collected information, either through tags (i.e., Hearst's triage tool [95]), categories and clusters (Clipper [123, 124]), or spatial arrangements (i.e., Sandbox [234], IdeaMache [142]). While these tools provide end-users with fine-grained control over the document and information saved in their final output, doing so requires a large amount of work to manually create an organization. Furthermore, as users encounter new information and their understanding of the space grows, they have to apply a significant amount of

effort to adjust their workspace to match their new understanding or add their newly encountered content.

Alternatively, a significant amount of work has gone into building data exploration interfaces, where sensemakers can utilize a system-generated structure through facets [94, 132, 135] or clustering [44, 45, 179, 241] to explore a given dataset. These techniques allow users to quickly sift through a large set of data while also learning about its structure [94, 214]. However, these techniques often assume a fixed dataset with all the possible information already present, whereas in the sensemaking scenarios we are interested in [192], the set of documents a user considers grows and changes as a user continues to serially encounter new content areas and topics they should pay attention to. Therefore, a user would either have to wait to cluster until all information is collected, or recluster their information and risk ending up with a different automatic structure. Additionally, these automatic approaches can result in nonsensical or very surface level organizations that don't take into account the nuances of the current sensemaking problem the user has.

Traditional categories, therefore, are not a good match for the sensemaking process: they require users to manually assign items to them, don't automatically incorporate new information, and are generally fixed and difficult to change without significant manual effort. Likewise, purely automatic categorization techniques have limitations surrounding their ability to produce contextually-applicable groupings and adjust as new information is added to a dataset. To tackle these issues, we explore how user-driven, keyword-based "smart categories" can more readily support the underlying needs of the sensemaking process. A *smart category* automatically pulls in relevant portions of the information a user has collected based on a set of user-specified category keywords. This not only has the potential of allowing users to create categories more efficiently, but also allows the system to automatically assign new information to existing categories as users explore the web and save more information. These categories allow us to target a sensemaking temporal "sweet spot", where users can benefit from categories as a way of organizing an ever growing collection of data, yet not have to worry about wasting a significant amount or effort as their categorization continues to develop and change.

In the second portion of this chapter, I discuss how we instantiated a version of these smart categories in a prototype system Distil. We then evaluate Distil in a user study and explore how the smart categories are able to support users' categorization needs during the sensemaking process. Finally, we take a critical look at the design of the categories in Distil, how its implementation supports sensemaking, and how it might be further improved. Based on these critiques, we offer some design suggestions for future implementations of user-driven automatic categorization techniques that could be utilized to support sensemaking.

5.2 Siphon

In order to enable developers and researchers to quickly build the next generation of tools for accessing and saving web information, we designed Siphon, a toolkit for flexible web annotation. Siphon provides support for multiple types of website annotations through a modular, extensible interface while also providing a set of feature rich defaults

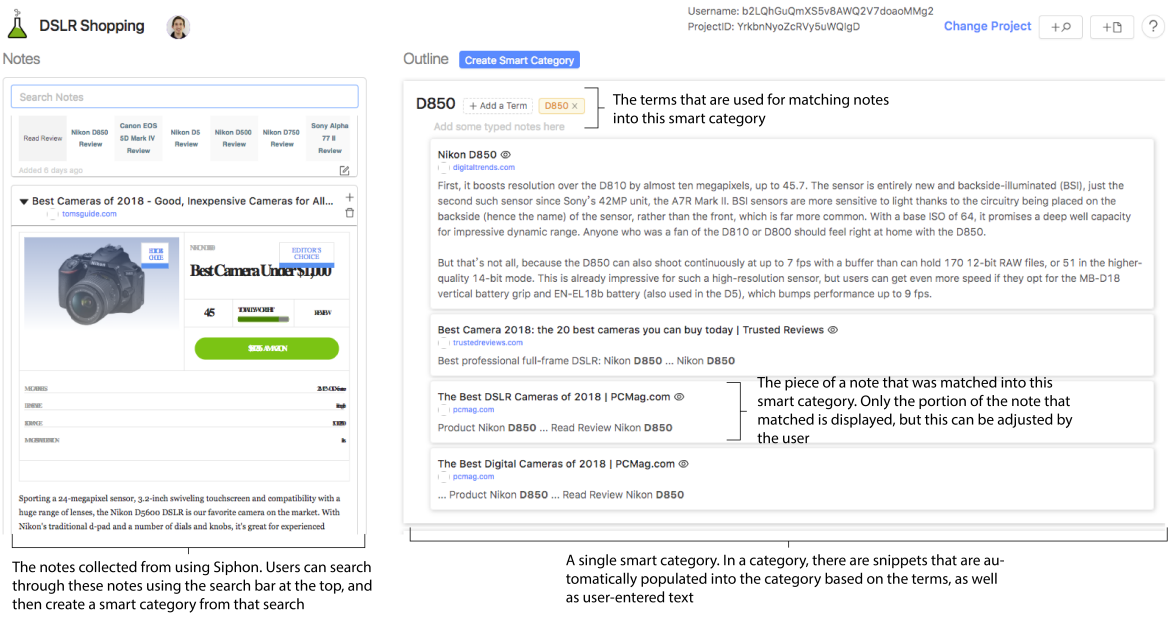


Figure 5.1: Distil's interface: The left shows the full notes a user has for this task, while the right is a free-form text editor with "smart" categories

for when developers choose not to augment a specific workflow stage. Inspired by WYSIWYG (What you see is what you get) design it provides developers with a simplified, abstracted mapping between the rendered output of a web document, and the underlying DOM model driving that output. This is instantiated in a few key ways:

- Siphon provides a unified interface for managing multiple types of selection on a single page, and simplifies the definition of a selection technique into a condition that must be true to be in that selection mode, and then three lifecycle callbacks that manage the setup, update and teardown of any elements that support that selection style.
- Content can be specified for extraction through a browser range object, a set of DOM nodes, or, uniquely, a set of window coordinates. In the coordinate case, Siphon will convert those coordinates back to the original DOM nodes generating the underlying content.
- The output of the annotations are graphical, interactive, and retain their connection to their original content. When an annotation is made, Siphon creates an self-contained, styled, interactive HTML snippet that can be shown in a standalone context. Additionally, it records the XPath of the original content, allowing developers to trace and surface already saved content.

Through these features Siphon allows developers to both create web-based annotation tools and techniques faster, while also maintaining a high level of support for activities post-selection.

5.2.1 Toolkit Description

Siphon provides support for the annotation through three major components: selection definitions, annotation objects, and a store for persisting and restoring annotations. Siphon draws its definition for the core components of annotation: selection, extraction, and reuse, from active reading [165, 210] and sensemaking literature [182, 192]. Alder et al in their taxonomy of reading [6] discuss how reading is often followed by the key activities of extracting and integrating information from different sources. Tashman and Edwards [210], evaluating digital active reading, note that tools for active reading need to support flexible selection and annotation of content, the ability to record and maintain context for annotated content, and the ability to easily visualize annotations and the relationships between them. Similarly, Pirolli et al.'s sensemaking model's [182] foraging loop consists of finding information, reading it, and then extracting it for later organization and use. These models suggest that an effective digital annotation tool should support flexible selection of content, the ability to extract content while maintaining a rich context, the support for further use of the information in a variety of contexts.

This section breaks down each of these components, and provides details about the built-in selection implementations that demonstrate the versatility and flexibility of Siphon's selection model.

Selection Definitions

A selection definition is simplified event lifecycle for triggering, updating and completing selections (Figure 5.2). A definition is composed of four main parts: a selection condition that must be satisfied and three lifecycle callbacks a developer can implement. The selection interaction begins when the selection condition is met, and ends when the condition is no longer met (such as the left mouse button must be pressed during selection). At the start of the interaction, the onSelectionStart callback is called, this allow developers to setup the selection interaction and any feedback mechanisms. For example, in the case of drawing a snippet box, inserting a bounding box element to the DOM tree at the cursor position. Whenever a Javascript user input event[2] (such a keyboard events, mouse events, etc.) occurs, the onSelectionChange callback is called. This allow developers to update the visual feedback based on users input. For example updating the boundary of the previously inserted selection box. And finally, when the selection condition is no longer met, the onSelectionEnd callback is called, allowing the developer to cleanup the selection interface, and based on if the user completed or canceled the selection action, create the appropriate Siphon annotation objects in response. In the case of the previous example, the developer would make a Siphon annotation object from the final bounds of the selection box and then remove the box.

Selection definitions allow for multiple types of selection interactions on a single page, managed through Siphon's unified event management interface. Traditionally, a developer would have to listen to the individual events associated with their selection, so in the above case the mouse down, mouse move, and mouse up, and then manage the state of their selection based on the ordering of those events. In Siphon, a developer

[2]https://developer.mozilla.org/en-US/docs/Web/API/UIEvent

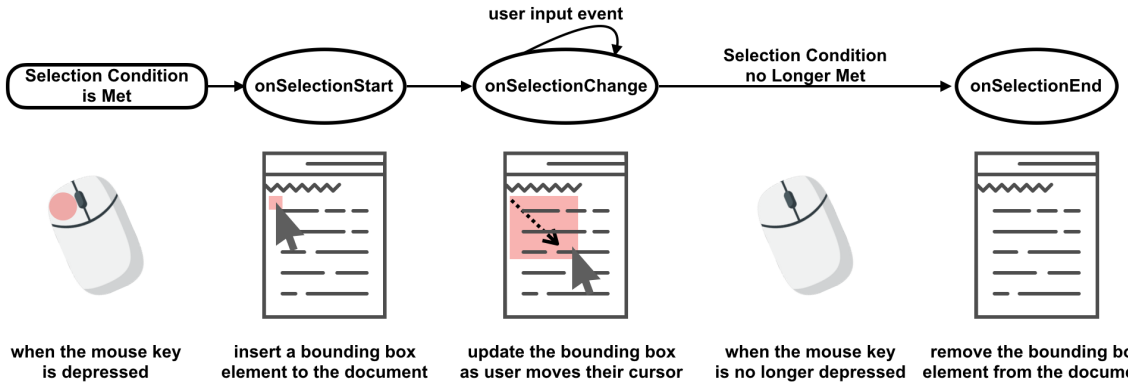


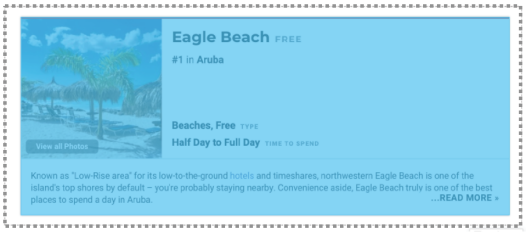
Figure 5.2: Siphon's simplified selector state diagram (Top) and an example Selection Definition of drawing a bounding box for selection (Bottom).

registers the selection definitions they want to use in their application, for example a snippet selection tool, force touch selection, and one click image saving. When a mouse, keyboard, pointer or touch interaction occurs, Siphon takes the event object and merges it with previous events to create a unified event object. This is then checked against all of the selection definition conditions registered with Siphon. If one is found to match, that selection definition is considered active until its selection condition becomes false – Siphon only allows for one mode of selection to be active at a time.

Annotation Objects

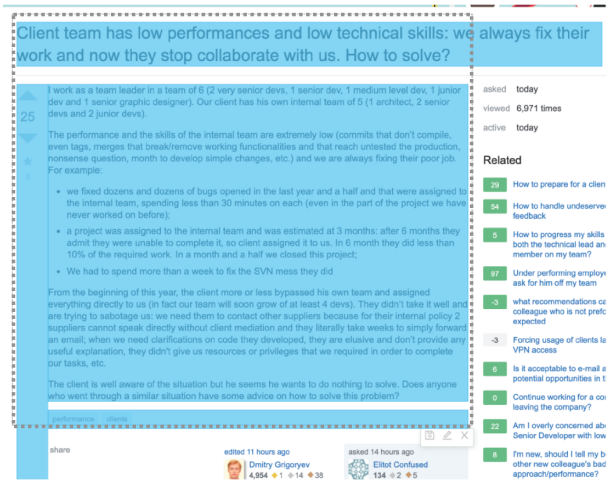
After a selection is completed, a developer can generate an annotation object based on the final output from their selection interaction. Annotation objects on creation identify the DOM nodes selected, extract the appropriate surrounding HTML including style properties, and generate a persistent reference to the original content. These objects can then be used in a variety of ways – they can act as a persistent references to the DOM elements on a page, they can be rendered as-is in external systems or interfaces, or the underlying HTML can be mined for metadata or specific content to surface. These can drive tools for consumption – such as visual organization tools, tools for performing in-page markup, and tools for automated extraction of information from pages.

Annotation objects are be generated from three core types of references: a single or a collection of DOM elements, a browser range object (generated through the default browser selection interface), or a set of graphical coordinates (i.e., a bounding box). These were selected based on the requirements of previous selection methods and tools for extracting web page data [101, 209, 245]. The graphical coordinates-based selection differs from the other two selection input in that the bounding box coordinates need to be resolved to its underlying DOM elements. Siphon accomplishes this through a bottom-up traversal of the DOM tree. First, Siphon calculates the intersection of all the block-level leaf nodes in the document with the current selection area. Then, after removing any fixed position elements, it iterates through those nodes to find the nodes with the greatest percentage of area overlap. Until at least one element is found, the algorithm adjusts the overlap threshold to a minimum of 50%. If no elements are

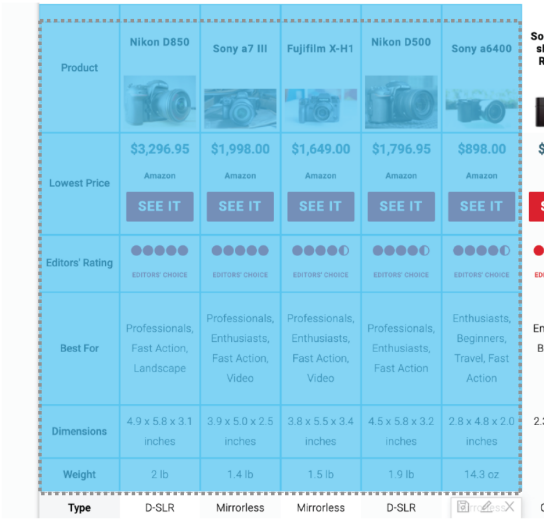


Selecting part an entire webpage component

Selecting part of a webpage component



Selecting multiple pieces of content
(note how Siphon expands selction to capture the full header)



Selecting a portion of an HTML table

Figure 5.3: Examples of Siphon's graphical selection tool. User specified bounding boxes in gray dotted lines are resolved to a set of DOM elements highlighted in blue

found at this point, Siphon assumes the user was selecting a portion of a block level element (typically an inline DOM element), and will return that element instead. After performing the leaf filtering, Siphon then attempts to reconstruct the DOM tree from the bottom up, searching for parents whose visible leaves are all present in the output of the leaf filtering. This reconstructed set of DOM elements are considered to be the selected set of DOM elements.

Additionally, Siphon allows for specifying a single point, instead of a bounding box, in the case a developer want to anchor an annotation to the page; e.g., leaving a manual note attached to a point in a page. In pen and paper interfaces, and digital document annotation interfaces [211], users might accomplish this by leaving a note in the margin near the content they are referencing. However, if a user were to click on this whitespace in a browser, this would most likely reference a large container element on the page, which not at all correspond to where they would actually want to leave a reference. In this case, Siphon assumes that the users is probably referring to some content in the center of the viewport, and uses a 1 pixel tall bounding box across the entire width of the page to try and find that center element. This allows for the reference to be attached to an element in the element in the DOM that is more closely aligned with their reference and follow it appropriately i.e., on resizing.

Once an annotation object has a set of DOM nodes to work from, Siphon then works to generate a static HTML snippet that can be rendered in alternative contexts. This is similar to the output of Hunter Gatherer [245], however instead of just creating a reference to the content, Siphon creates a fully styled, portable HTML snippet. This is uniquely challenging, due to the cascading nature of CSS, as pages are designed to be rendered as a complete entity, not piecemeal.

To accomplish this, first Siphon then creates a copy of the DOM nodes from selection and embeds all of the currently applied CSS styles into the nodes and ensures any external URL references are absolute in nature, rather than relative. Siphon performs several optimizations to ensure all possible CSS styling and HTML content are captured, as well as ensuring the snippet is properly rendered in different sized containers. First, Siphon removes any metadata / scripting tags, and also embeds the content from accessible iFrames into the snippet. Second, if any CSS pseudo elements (i.e., before and after) are present, Siphon extracts those as a separate CSS style description, assigns a unique additional class to their corresponding element, and includes this extra CSS style description in the final HTML output. Finally, in order to properly support layout reflow in different contexts (i.e., different sized containers), Siphon ensures that any CSS styles included utilize their computed, rather than their actual values (e.g., an element with a percentage based height keeps this as its embedded value instead of the currently rendered pixel based height the element would have). In order to reduce the size of the final HTML snippet, Siphon detects any default CSS values, and removes those from the embedded CSS style definition. While this works for many cases, there are usually some subtle layout differences (Figure 5.4), and in the case of highly interactive content, such as a D3 visualization, the snippet generation can fail altogether.

Finally, Siphon supports referencing the original source of an annotation using a set of XPaths and the document URL. As found in Zoetrope [5], maintaining the position and provenance of the original annotated content is challenging due to the always changing nature of web pages, as DOM elements can shift position in future revisions

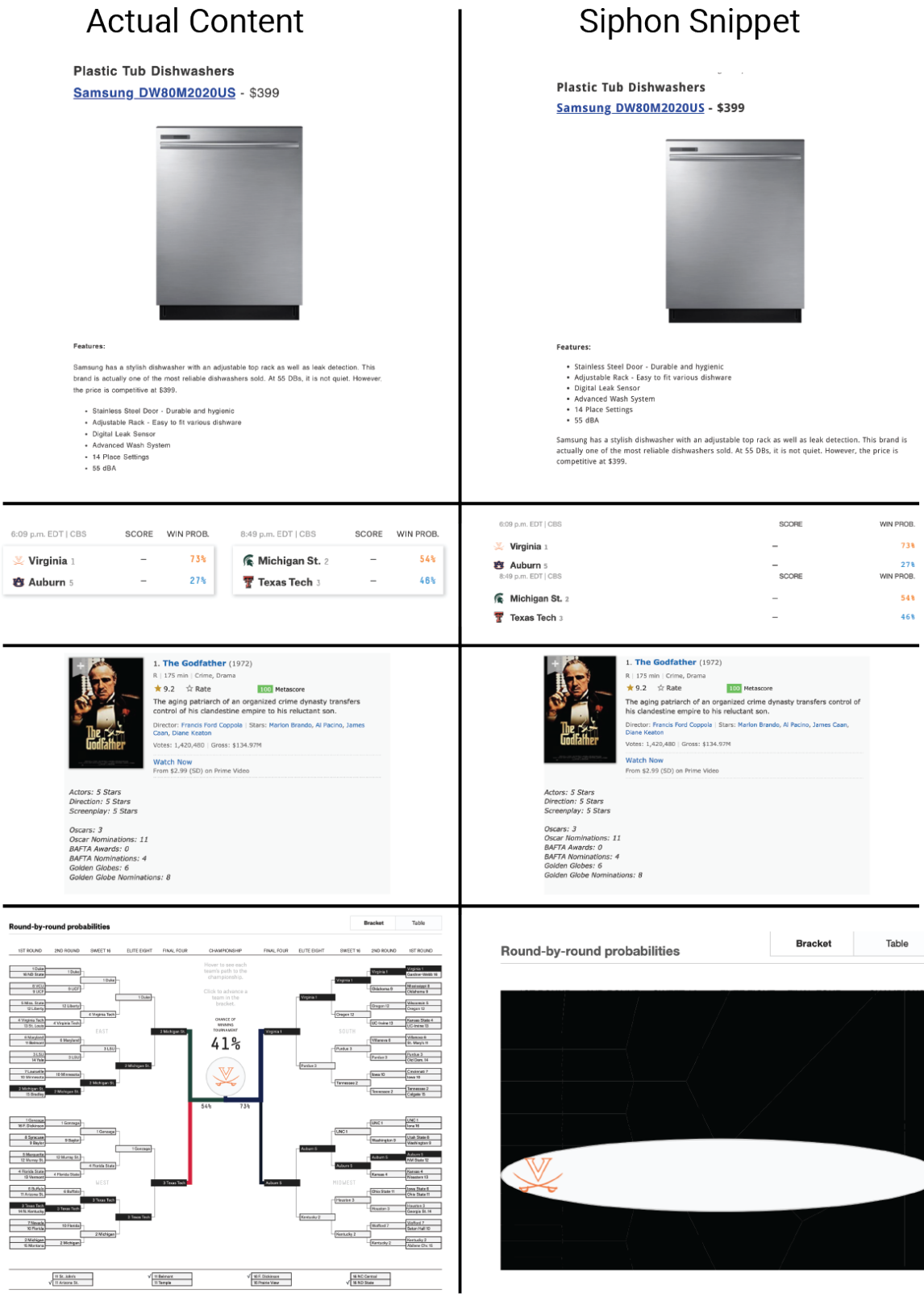


Figure 5.4: A comparison between the actual rendered content on a webpage and an extracted Siphon snippet that can be rendered outside of its original context

of a page. Siphon currently uses the open source npm package xpath-dom [100] to accomplish this, however more sophisticated XPath generation and resolution could be performed, such as that in Doncheva et al.'s work [62, 63].

Annotation Store

The final component of Siphon is a store for all the annotation objects generated by users. The default store will only save Siphon annotations to memory, as it is up to the implementer / developer to decide how and where these annotations will be saved. The store defines 4 CRUD-like core methods for saving, removing, updating and importing annotations that should be implemented by the developer using Siphon. These methods serve as the interface for serializing and deserializing the Siphon annotation objects to/from JSON, as well as, on deserialization, finding the original location an annotation was made on a page and marking that with a custom class. Because modern webpages are increasingly dynamic and utilize a large amount of client side scripting for generating the final DOM content, the store also monitors changes to the current page's DOM and searches for any annotations that have not been properly restored and attempts to restore them.

Implementation

Siphon is implemented as a vanilla Javascript library that can be utilized in a number of different contexts: as a browser extension, as part of an electron application, or embedded into a specific webpage / application. The only requirement is that Siphon has a modern DOM Document object to work with – thus it can also run in a headless browser. The library can be found on github: https://github.com/BentoBrowser/SiphonTools

5.2.2 Built-In Tools

As mentioned, the Siphon toolkit provides a number of built in selector tools that let developers utilize the toolkit without having to manually define a number of basic selection types. These built-in tools consist of:

- **HighlightSelector** Supports the integration of the built-in browser selection tool with Siphon
- **SnippetSelector** Allows a user to draw a bounding box around some content on the page
- **ElementSelector** Users can select a set of DOM elements directly. As they hover over an element, a blue overlay is displayed on the element. They can click on this element to add it to their set of selection elements. If they click on an already selected element, it removes that from the selection set
- **ClickSelector** An implementer passes in a CSS selector to this selection definition, and any click on an element that matches that selector will trigger this definition.
- **HoverSelector** An implementer passes in a CSS selector to this selection definition, and when a user hovers over an element matching this CSS selector, this definition will become active.

- **ListSelector** A more experimental selector that attempts to automatically identify "lists" of elements on a page (these can be DIVs, table elements, etc), and then provide those as a set of possible selection targets. Users, when this selector is active, can then select list elements in a manner similar to the ElementSelector.

Aside from these built-in selectors, developers can also create their own by using the selection definition framework of Siphon.

Through it's simplified selection definition tools and rich annotation objects, Siphon can enable developers and researchers to create more powerful web annotation tools with a simplified Javascript API. One instance of this is Unakite, a system for supporting developers during their software development decision making process [144]. With Siphon, Unakite allows developers to clip the relevant parts of a resource that drove their decision making while coding and structure them into a table. Beyond supporting annotation, Siphon has the potential to augment other online activities, such as discussion or collaboration. Many news articles articles now contain embedded discussion forms for users to provide their personal opinions or share additional information about the topic. While these tools offer the ability to respond to other comments, they either lack support for responding to the original content, or only provide very basic support referencing that content (i.e. only highlights). A developer, wishing for richer discussion around the article content, could implement Siphon as a way to allow for inline discussions, similar to Note Bene [242], or as a way in a comments thread, to reference the article content with a variety of selection tools. In the next section, Siphon is utilized as the input for an organization tool, Distil, which helps individuals structure the rich data they've collected into ephemeral, flexible "smart categories".

5.3 Distil

Distil is a prototype system for organizing clips of information that users have collected from the web [3]. The core component of Distil is the smart category: a keyword-driven filter that continuously categorizes new content added during the sensemaking process. Categories can be further refined and explored through a few different mechanisms: adding or removing filter terms, viewing the original source content for the snippets pulled into a smart category, adding additional clips to a category, and specifying more specific summary text for a clip pulled into a category. smart categories allow users to begin structuring their data at any point in the sensemaking process, without having to worry about the trade-off between structuring too early (thus creating obsolete structures) or too late (having to deal with an overwhelming amount of content).

5.3.1 Related Work

Many tools for assisting users during the exploratory search process do so by providing up-front categories through clustering [44, 45, 52, 129, 179, 241] and facets [132, 133, 214]. Apolo [45] takes a unique approach to this by having users build up a categorization through exemplars rather than providing a top-down set of categories produced by an unsupervised algorithm. This allows for the user to iteratively build up and adjust

[3] See supplementary video Distil_Tutorial.mp4 for a demo of the system

their information landscape as they continue to learn. However, Apolo is designed to work with a metadata-rich data source, while general web-based sensemaking often occurs with small, sparse partial sections of a document, making it difficult to provide a mixed-initiative system like Apolo with enough useful data for clustering. Grouper [241], alternatively, works on the subdocument level by clustering document snippets returned from a search. While these categories can be extremely beneficial for getting an initial overview of the space and inform users what features they should focus on [135, 214], they often do not align well with user's final mental model from their sensemaking process [124] and can become obsolete as the set of information gathered continues to grow. Therefore, while automated clustering techniques are very beneficial for gaining an overview of a fixed information space, they aren't adaptable enough for the sensemaking process where the information space and the user's mental landscape are constantly updating.

Due to the limitations of the above approaches, other researchers have focused on tools that introduce helpful metaphors or specific organizational layouts that assist users with collecting and organizing information during different phases of the sensemaking process. These include early stage tools for helping users manage and re-find sources, such as WebBook and WebForager [35] which utilize a book metaphor for managing pages, Elastic Windows [113], and Webcutter [147] which presents URL collections in a tree, star and fisheye structure. These tools provide significant support for managing sources, but are generally limited to entire documents and are focused on supporting retrieval of existing information rather than generating organizations for explanation. Alternatively, organizational tools attempt to provide flexible ways to distill and group content, such as Clipper [123, 124] which allows users to assign attributes to a clip at the same time it is being saved from a web page, IdeaMache [142] which uses a spatial layout for helping users group and manage content, and Hearst et al.'s triage tool [95] which uses a streamlined interaction for categorizing and tagging documents. While manual organization tools are extremely beneficial for creating detailed, easily interpretable artifacts, they can require a significant time and effort investment by the end user to create, and can become obsolete during the sensemaking process as a user's understanding of the space changes [124].

These existing tools largely support developing and applying structure at two points in the process: the very beginning or the very end. Users can develop a predefined categorization and apply it to documents [95], or they can defer any categorization until they have collected enough data to have a complete picture of the information space [44, 52, 124]. However, realistically in sensemaking reasonable structures can begin to emerge part way through the process [181, 192], and while they might require refinement, they can be beneficial for supporting information re-finding, and reducing the cognitive burden of having to organize all the data at the end of the process.

This suggests that there might be a temporal sweet spot in the sensemaking process, where an individual has enough information to create a rough and partially accurate categorization, and benefit from this organization as a means of sorting new information as they discover it. To that end, we propose the novel concept of "smart categories" to support the sweet spot of sensemaking in an interactive, Web-based prototype system **Distil**. Through supporting this sweet spot, the smart categories can reduce the effort that would normally have to be applied post-hoc to organize data, and reduce premature,

uninformed categorizations for the sake of an organized data set.

These categories are designed to support users in the triage phase of sensemaking [95], where they are still actively building a mental model of the information space as they gather information of interest. This differs from search and relevance feedback approaches which focus on helping them find those items in the first place [131, 193]. Existing tools, such as HyNote [158] and Scrapbox (scrapbox.io), aim to support this triage stage by allowing users to link concepts together, however they have the drawback of being a purely manual approach.

We take inspiration from IntentStreams' [15] query refinement interface – their constantly updating streams in response to user feedback allow users to iteratively adjust their search. At the core, Distil supports iterative refinement of information foragers' sensemaking workspace, by giving them low cost categories that they can quickly create, delete, and modify. Distil's smart categories uniquely promote and incentivise the deferment of organization until partway through the process by providing a mechanism that is directly beneficial (i.e., categorizing subsequent information automatically), one-shot (vs. requiring multiple feedback instances), transparent (i.e., through using keywords), and adaptable (i.e., by adding and removing keywords). In the following sections, we describe how Distil is designed to achieve this goal.

5.3.2 Clips

Distil's smart categories are designed to work with "clips" from web pages – a subset of content selected from a page by a user. For example, consider a scenario of a user trying to find good restaurants in a city they are visiting. Initially, they may go through several listicles to find popular restaurants, as well as options listed on restaurant-rating applications such as Yelp around the area they are staying in. As they browse these sites and certain restaurants catch their eye, and they might want to record the recommendation and the additional information that goes along with it. While the ideas in the Distil system can work with any sort of textual snippet, we wanted to give users a rich experience, so we paired it with our Siphon toolkit to allow for rich selection from users. The snippets collected with Siphon appear in the Distil interface on the left, and are fully interactive and searchable (Figure 5.1, left).

As mentioned in the previous section, Distil is aimed at supporting the triage stage of sensemaking, where the user is expressing interest in certain pieces of content, and is starting to form a mental picture of the information space. By supporting Siphon's rich clips in addition to text highlights in Distil, we consider how users could benefit from better, quicker structuring with richer underlying data through the keyword based smart categories.

5.3.3 Outliner

Once a user has enough data that they want to start organizing with, they can utilize the right-hand side of the Distil interface: the outliner (Figure 5.1, right). Here, users can begin to form groupings of relevant information and record any additional thoughts about the captured information (i.e., maybe a restaurant would be less ideal because they aren't gluten free). The outliner allows users to create hierarchical outlines of

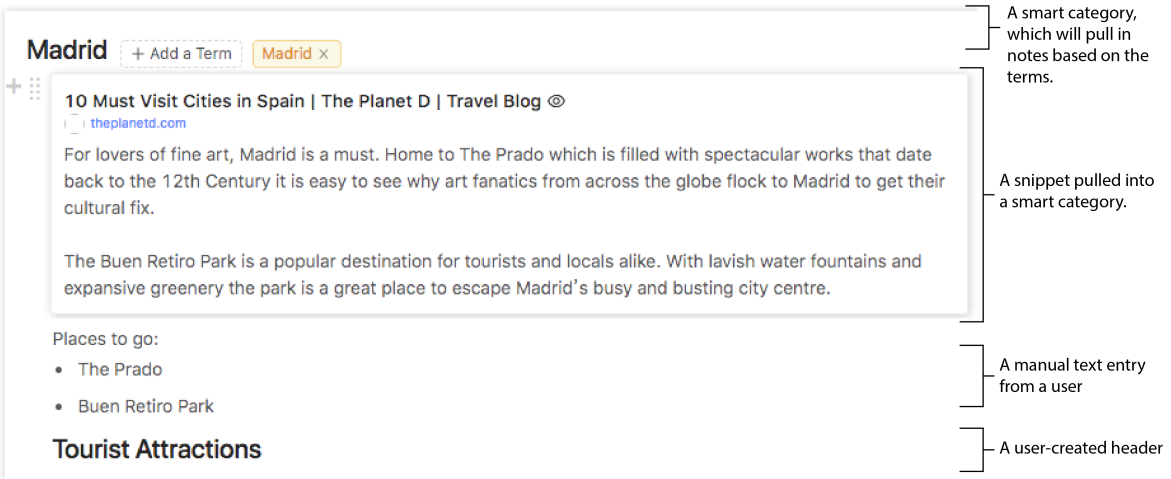


Figure 5.5: The Distil outline interface – shown is a smart category with both a snippet inside of it as well as manually entered user text

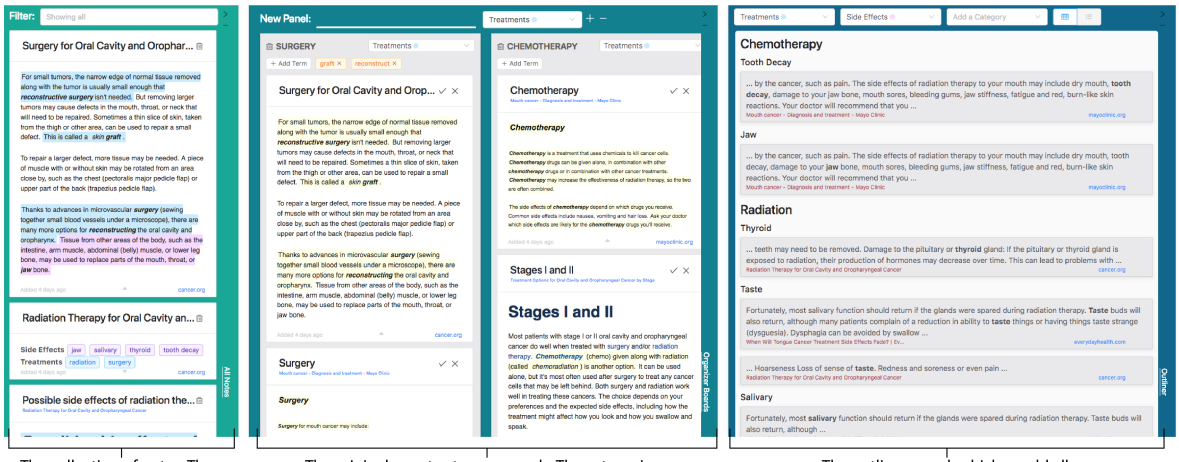
clips and manual notes using a free-form text interface, interleaving clips with personal thoughts (Figure 5.5). In this interface, clips and sections of text are treated as individual, nestable, reorderable elements. As a user continues to refine their outline, they can seamlessly add, reorder, and remove these chunks of content, similar to a popular note-taking application Notion [217].

This integrated document design was based on some initial prototyping and testing we did. In an earlier version of the interface, users had no easy method to record their thoughts or opinions about a category. In another intermediate version of Distil, we added a comments section at the top of categories, however we discovered that users were just copying and pasting text from the original notes into those sections because they had no way to associate their notes with a specific piece of text. This led us to the final, more free-form design where users could interleave matched clips with manual notes (Figure 5.5).

5.3.4 Smart Categories

From within the outliner, users can utilize the core feature of Distil: the smart category. A smart category represents a particular topic within the data a user has collected that they would like to explore further; retain for later use such as comparison or sharing; or as a way to remind themself and surface important content. Continuing with the restaurant example, say the user has gone through 5 different lists of restaurant recommendations, and now wants to apply some order to the data they have collected.

Using some prior knowledge about different types of cuisine, the user then starts with making a few different smart categories: a 'Japanese' one, an 'Italian' one, and a 'Moroccan' one. As these categories are created, the category names are used as the initial search term for gathering clips that were saved in the system, such as reviews extracted from Yelp of different listicles. Distil uses the standard Okapi BM25 ranking function for ordering the snippets [187] and a clip can be featured in multiple categories,



The collection of notes. The portions of notes that match a particular category are highlighted in a color associated with that category.

The original smart category panels. The categories would pull in the full original clips, and highlight the corresponding parts of them that matched. Different categories could be assigned to groups, like "Treatments" or "Side Effects"

The outliner panel, which would allow users to cross groups of smart categories with each other to produce an outline.

Figure 5.6: The original Distil interface with full-text notes categorized in a series of category panels and a separate "outliner" panel

as long as it matches. Rather than showing all of the text from a clip in a category, a snippet from the clip is displayed based on any matching category keywords, similar to a search snippet. Users can rearrange these snippets, delete irrelevant ones, and update their summaries to be more concise. Additionally, a user can revisit (focus on) the full clip where a snippet was extracted from, in order to gain more context or explore additional details not present in the summary.

In the original version of the Distil interface (See Figure 5.6), rather than showing selected snippets of information from the notes, the interface placed the entire note into one or more category panels. We pilot tested this interface with about 10 users, and users generally felt very overwhelmed by the amount of information, and unsure of how to consume it. Based on that feedback, we iterated on the interface to the current form, where the categories pull in summaries from the notes, rather than the entire thing.

Noticing there are not a lot of clips in the Japanese category, the user focuses on expanding their search for more potentially good Japanese restaurants. As they continue to collect clips, Distil continually checks the content of these clips against their current smart categories. Any smart categories matching that new clip will display a small notification under them saying "New Notes Found!" in red. So, in this case as the user continues to collect information about Japanese restaurants, their Japanese category will display a notification indicating new content. A user can then click on this and add any of the desired matching clip snippets to their smart category (Figure 5.7).

As the user continues to learn more about the restaurants, they might realize that the initial keyword Japanese is not capturing some of their relevant clips that did not explicitly mention it. To resolve this, they add more keywords as an additional tags to their Japanese category, such as "ramen", "sushi", and "takoyaki". The union of the results from these tags is then computed, and if anything new matches, they will get a familiar "New Notes Found!" notification under the category where they can add any desired missing snippets. Users can add additional tag phrases in a few different ways:

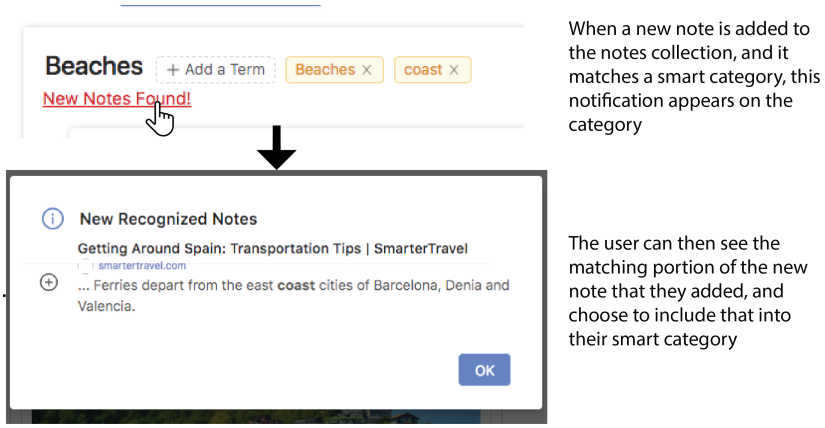


Figure 5.7: When a new note recognizes a smart category, a small notification appears. Clicking on it opens a modal dialog where users can then add that note to their smart category

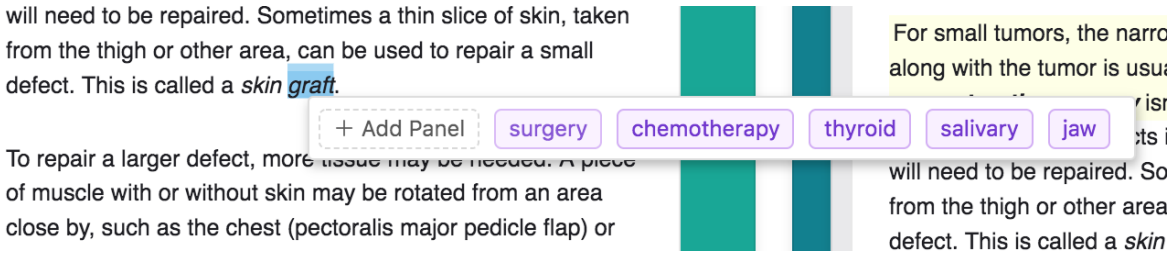


Figure 5.8: Toolbar for creating a new category or adding a term to an existing category

they can enter them directly, they can choose a phrase in the text of their clips to add a term from (Figure 5.8), or they can use the search bar at the top of the clips section to add a phrase. When using the search bar, they can preview what will appear in the resulting category before adding it. If adding a phrase from the text, we enhance the process by suggesting the top 5 existing categories that the phrase would most likely match by computing similarity of the word selected against each category name using a set of word vectors [180] and then rank the results.

As they read through the clips in the Japanese category, the user notices that the summary text is not very useful, and somewhat verbose. In response, they modify the snippet text by selecting new text from the clip to show there instead (Figure 5.9). They can also choose to show the entire clip instead of just the snippet, which can be useful in the cases that an entire clip is relevant, or it contains some highly graphical data (like a table) that would be better viewed in its original format. Finally, in the space at the top of the Japanese category, the user manually jots down the top restaurants they want to go to based on their category's data.

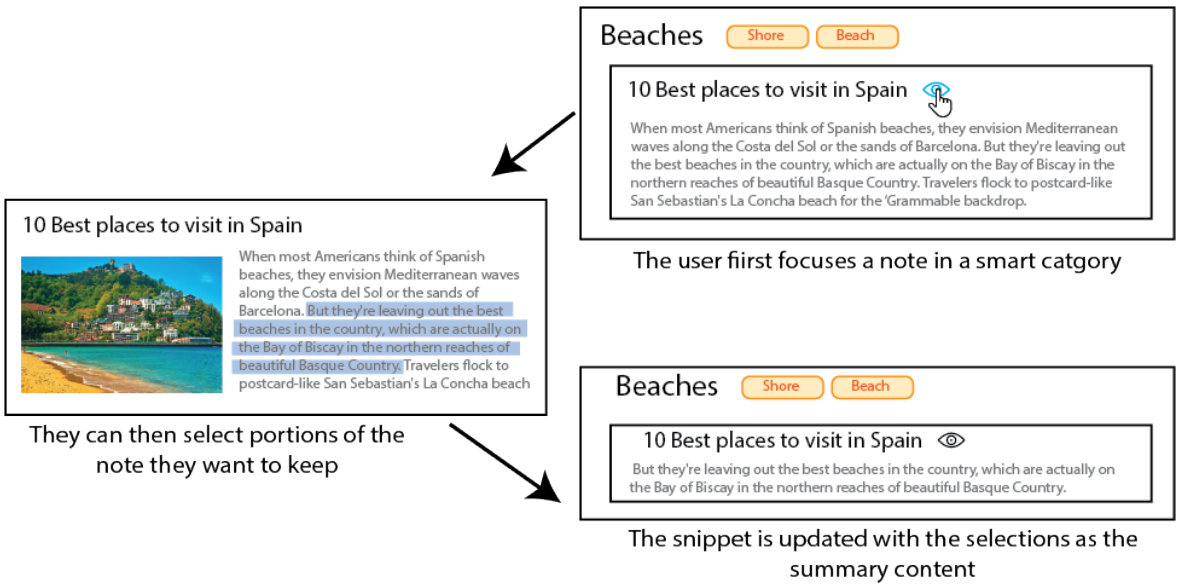


Figure 5.9: The interaction for editing a note summary – a user first focuses on the note, and then chooses a new summary

5.3.5 Implementation

Distil is implemented as a web application. The application is written in Javascript using Facebook's React library for interface rendering, and Google's Firebase database service for data persistence. Lastly, it uses the open source Lunr.js search engine to provide the filtering capabilities of the smart categories. Lunr.js indexes the text in all of the notes, which is then searched through when a user creates a new category, adds or removes a tag in an existing category, or a new note is added to the workspace. This indexing is performed on the client side, allowing Distil to be responsive to one of the above events.

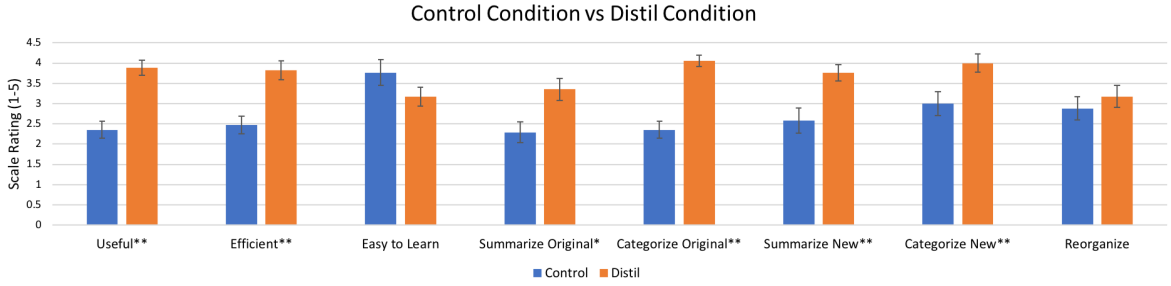


Figure 5.10: Results of the post-survey comparing the control and Distil conditions. ** $p < 0.1$, * $p < 0.5$

5.4 Evaluation

For our evaluation, we wanted to understand if smart categories allowed users to effectively sift through and generate a supportive digital artifact from data they might have collected during the sensemaking process. To that end, we designed a lab study, where individuals were asked to produce organizations for two different tasks, in two different conditions. Because we were mainly concerned about the impact of smart categories on the structuring portion of the sensemaking process, this was performed as a within-subjects study, comparing a Distil interface with smart categories, and one without (essentially the notes panel next to a rich text editor). Through these two conditions we test whether the smart categories were sufficiently accurate, intuitive, and adaptable for end user use.

To control for the difference in sensemaking task complexity as well as possible prior knowledge, we pre-selected two tasks for participants to complete. The prompts for those tasks were as follows:

- **Spain Scenario**: In this task you will organize / summarize information for planning a trip to Spain with a friend or significant other. You need to consider what areas you'll go to, what attractions you might visit, and how you'll get around.
- **Diabetes Scenario**: In this task you will organize / summarize information for a friend / significant other / child around the treatment, side-effects, and long term considerations of managing Type 1 Diabetes. They were just recently diagnosed, and are looking for some input in how to manage their condition.

In order to balance control of information with ecological validity, we chose two fairly common research tasks (travel planning and health research), but provided the clips of information participants would use for developing their structures. We were concerned that differing levels of interest around the topics would result in vastly different outcomes (for example, if a person just copied the results from the first website they found). By fixing the clips we could guarantee a minimum amount of information participants would have to consider. Clips were selected from the top 10 google results for a general query surrounding the task topics:

- **Spain Scenario**: What should I do on my trip to spain?
- **Diabetes Scenario**: Recent Type 1 Diabetes diagnosis

The clips were then collected by clipping paragraph sections from the body of these documents relevant to the queries (Diabetes: 40 clips, with 161 words on average, Spain: 71 clips, with 133 words on average). By providing this set of clips, we intended to simulate the information an individual with a strong motivation might collect.

Another key component of the smart category interaction is its ability to incorporate new found information, and allow users to easily add, remove, or enhance them in response to said data. In order to simulate this scenario in our study design, we divided the clips, providing only $\frac{2}{3}$rds of these clips initially, and then later the additional $\frac{1}{3}$rd. Through this treatment of the provided data, we wanted to explore the hypothesis that Distil would be able to support the sensemaking needs of restructuring, refinement, and gap-filling [244].

The complete design of the study consisted of 6 different sections: a demographic pre-survey, a 7-question Maximizer-Satisficer Questionnaire [170], a short interface

Type	Event Name	Percentage of all Category Events
Adding / Removing	createCategory	*
	deleteCategory	3.89%
Evolving	addTag	14.75%
	removeTag	7.24%
	changeCategoryName	2.95%
Exploring	focusNote	29.36%
Curating	deleteSnippet	32.17%
	saveFullNote	1.07%
	saveSelection	8.58%

Table 5.1: The breakdown of smart category event frequency
* createCategory is listed for easier future reference

training, the two sensemaking tasks, and finally a post-survey asking their experience. We utilized the Maximizer-Satisficer scale to estimate how much effort a participant might spend in a sensemaking task. During the training, participants first watched a short video explaining the system, and completed a simple task on researching a DSLR camera to recommend to a friend, using the interface with smart categories enabled.

For each sensemaking tasks, individuals first spent 20 minutes with the first set of clips, 5 minutes with the next set of clips, and an additional 5 minutes to perform any restructuring that they might need to do. The sensemaking tasks, as well as the conditions (with vs without smart categories) were randomized and counterbalanced to assure that no ordering or task-effects were impacting the feedback we received. In the post survey, we asked individuals to compare the two interfaces to each other and provide some additional qualitative feedback about their experience with the smart categories.

The study consisted of 17 participants (9 identified as female) from a local participant pool. The average age for the participants was 26 ($\sigma = 8.73$). Nine participants reported as college graduates and 8 of them are currently enrolled in an undergraduate program. The study took approximately 90 minutes and each participant was paid 15 USD.

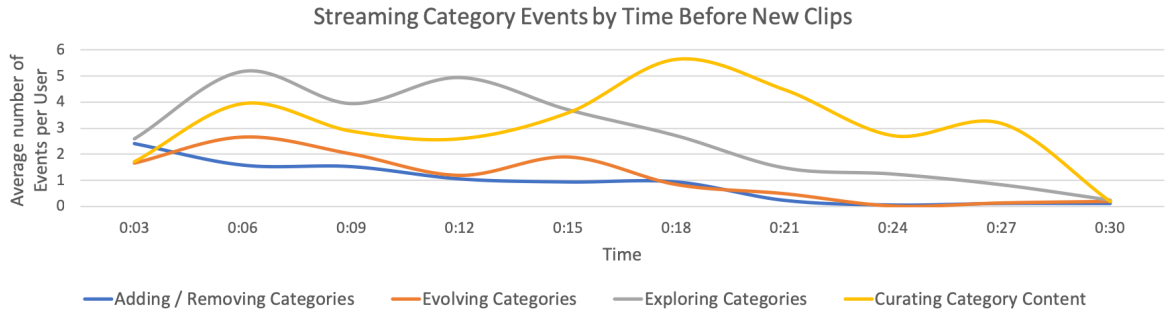


Figure 5.11: The average number of each type of event by time for a smart category before new clips were added

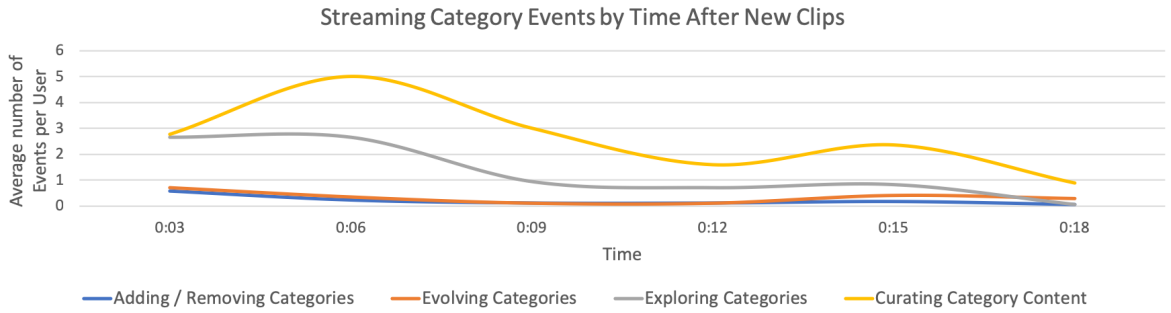


Figure 5.12: The average number of each type of event by time for a smart category after new clips were added

5.5 Results

Converging evidence suggests that participants found smart categories useful. First, participants heavily used the smart categories when they were available – on average each participant created 11 smart categories ($\sigma = 5.29$) in the condition when they were available with approximately 10 ($\sigma = 12.55$) clip snippets in each category at the end of the study. Out of the 17 participants, 11 created 4 or more smart categories during the study. Since participants could have used the search feature and copy/paste to generate categories manually, this suggests that they were perceived as valuable. In the case usage was just due to the "smart categories" being a new or interesting feature, we asked users to directly compare their experience in the condition where the smart categories were available to the one where they weren't. On 8 different likert scale (1-5) questions comparing the two interfaces, users found Distil to be both more useful ($t(16) = 5.24, p < .01$) and more efficient ($t(16) = 4.51, p < .01$) with smart categories (Figure 5.10). To ensure this wasn't correlated with the type of task (medical vs. travel), nor demographic or their Maximizer-Satisficer response, we ran a mixed-effects model with those, as well as the interface (with or without smart categories) and their interactions as predictors. Only the smart categories were found to be a significant predictor. Four of our participants directly mentioned in their free response question that the smart categories save them time, while 3 of them directly mentioned its organizational power:

> "I like the idea of using this system to form an itinerary since it can pull from multiple sources and compile information into one document."
>
> "It was visually appealing to have a lot of data / information organized."
>
> "easier and saves labor ... could quickly decide what info goes where."

While the overall response to the smart categories was a net-positive, participants noted several issues with them, especially in the way they were implemented in the Distil system. The largest complaint centered around the over-eagerness of the smart categories – several users noted issues with having to prune items and redundant info.

This resulted in a view that "felt a little disorganized" or generally was "cluttered". One participant stated that it was "just not user friendly for someone who like to have control over how to manipulate the data". Additionally, in our scenario, users were encountering the information for the first time in the Distil interface, rather than on the web. While this may have resulted in participants using the smart categories for more exploratory behavior than a normal scenario, the non-smart category condition still allowed users to use the search bar above the notes panel to search through the notes. Because search functionality was still present, we believe that the use of the more permanent smart categories represented a real attempt by users to actively categorize information in, what they believe, is a more efficient manner. This was also consistent with the continued user interaction – after creating a category, users continued to perform on average 4 additional actions ($\sigma = 4.86$) (Table 5.1), including removing around 2 snippets ($\bar{x} = 1.88, \sigma = 5.22$) from the category during the study.

Aside from the self-reported efficiency gains from the smart categories, we wanted to understand if there were any other features in their design that led to the positive user response. Below we look more closely into how smart categories were used by participants, where categories were able to work well for users, and where they fell short.

5.5.1 Category Types

At the end of the study, we went through and explored artifacts created by participants. Looking at the smart categories generated, we noticed three general trends:

- Single topic categories with only the title as the filter (i.e. Insulin, Beach, Opera, Kidney Problems)
- Single topic categories with the title and synonyms (i.e. **Where to stay**: hotels, **Beach**: ocean, island)
- Compound topic categories that used words learned while researching that topic as filters (i.e. **Near Madrid**: Royall Palace, Retiro Park, Burgos, **Managing Diabetes**: exercise, sleep, day-to-day)

All but one of the participants utilized the first type of category, eight of the participants used synonyms in their list of filter terms, and only four of the participants created compound topical categories. We think this might relate to both understanding how the system works, as well as the amount of effort required to refine the category. The first is the simplest, and most straightforward, and will work for simple concepts, while the last requires a complete understanding of what is happening, but allows for more complex expression.

In the non-smart category condition, we observed some additional structures and categorizations, including pro/con lists and nested structure (i.e. Tourist Attractions > [Name of Area] > Famous For..., Key Attractions...). Additionally, people were able to successfully combine external structure together, for example three different list of diabetic side-effects from different sources. While the smart categories were able to partition information into the appropriate section very quickly and efficiently, the categories are unable to automatically perform a proper summary or combine information sources together. However, users could utilize the free-form notes fields to

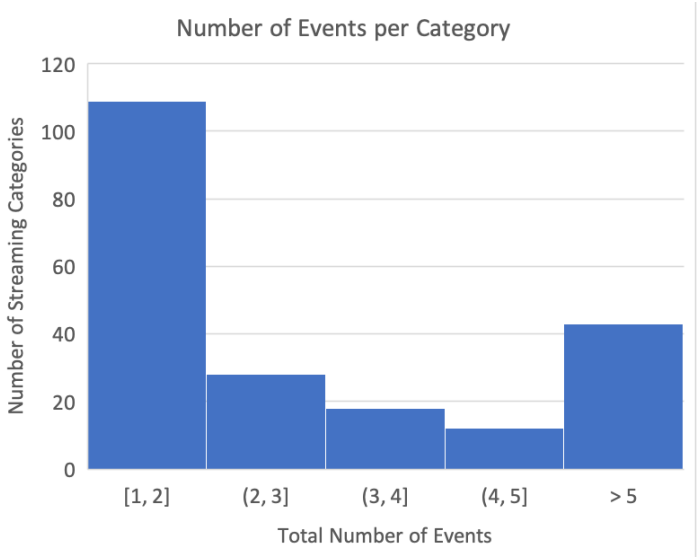


Figure 5.13: A histogram of the number of total actions performed on the smart categories

do this themselves. Therefore, as Distil is now, it might be more effective as an earlier triage tool for putting the right "chunks" of information together, however it could be further refined into an interface where users can easily go from "chunks" to streamlined summaries.

5.5.2 Category Evolution

In addition to providing categorization support, Distil was designed to allow users to easily, flexibly and transparently adjust their digital workspace to match their current understanding of the information space they are learning about. As such we expected participants to continue to add, modify and remove smart categories throughout the session; this indeed appeared to be the case (Figure 5.11). While category creation appeared somewhat front-loaded, with bumps both at the beginning and after new content was added, users were modifying and updating their structure in Distil throughout the session.

As noted in the breakdown of category events in Table 5.1, the amount of category curation was fairly high, representing almost 40% of the events occurring to a category after it was created. Users were also consistently pruning clips from their categories, about 2 clips per category ($\bar{x} = 1.88, \sigma = 5.22$), suggesting that while the categories were able to pull in relevant information, some of that content was also either duplicated or not relevant. The snippets pulled into the categories were also often adjusted (saveSelection event, see Table 5.1), which suggests while a relevant piece of information was pulled in, it was either pulled in with unneeded context or not enough. These interactions imply that while the categories were quick and efficient for sorting through data, the information in them was possibly still to verbose to send to a friend, and needed to be further curated by the end user.

Lastly, while we intended for smart categories to be flexible and allow for structural

evolution, we also noticed that users were creating, deleting and modifying categories as a way to test their effectiveness. This is suggested in the bimodal distribution of the number (Figure 5.13) of actions performed on each category, where a large number of categories were either considerably unchanged after creation or frequently refined and evolved by the users. This can be attributed to the way categories appeared to be used for exploration – on average users delete about 3 ($\sigma = 3.08$) categories and when a category was deleted, this typically occurred within the first two actions ($\bar{x} = 2.23, \sigma = 1.52$). We noticed a similar trend for filters: Users added approximately 9 additional tags ($\sigma = 7.24$) to their smart categories and approximately half of the time ($\bar{x} = 5.22, \sigma = 4.90$) they removed one of those tags. There was a high level of variation in both of these situations, suggesting that some users could be experimenting with different structures or phrases for their structures more so than others.

5.5.3 Continuous Refinement

To more deeply probe when users were creating versus refining their categories, we looked at the event timings recorded for the smart categories. There were several actions users could performing in this time period (Table 5.1) – including adding an additional tag (addTag), changing the category name (changeCategoryName), focusing on a clip added to the category (focusNote), removing a tag (removeTag), manually adding a clip not automatically captured by a category to it (saveFullNote), or changing the default clip snippets (saveSelection). About one third of the time users were exploring the data pulled in by the category, while the other two thirds of the time, they were further refining the information captured by the streaming category. This split in activity suggests that users wanted to further verify that their streaming categories were working appropriately (focusNote), while also tweaking the final output of the category so that the persisted information was more useful (addTag, removeTag, saveSelection, deleteSnippet). To see if there was any late versus early stage variation in the types of activities performed, we divided up the event space of the smart categories into three time segments of approximately 6 minutes each (due to each phase of the task taking about 20 minutes). The earliest segment had the most category and tag creation events (14% and 21% respectively), with smaller amounts of note focus and save selection events (18% and 21%). The second and third time segments featured lower category and tag creation events (8% and 13%) and slightly higher note focus and save selection events (26% and 29%). We hypothesize these numbers align with the behavior of creating categories for known or important areas up front, and then later refining those and adding less important or newly discovered categories.

As part of our study design we added a hold-out set of one third of the total clips to each user's workspace as a way to simulate gathering additional data during sensemaking. We expected this to simulate a scenario where users would have to restructure, or significantly change their structure as to accommodate the new information. However, in our qualitative survey, we didn't find a significant difference between the Distil and base condition for the question "How easy was it to change my organization in each interface?" (Distil: $\bar{x} = 3.17$, Simple: $\bar{x} = 2.88$). Surprised by this response, we hypothesized that this was due to individuals not really having to ever truly perform a drastic change to their organization; rather they were just refining and adding to it as

they went. We saw evidence of this in the free text responses as well:

> "I mainly changed the sequence in which the notes appeared. Also I added headings to specify the spots which are closer to Barcelona and Madrid."

> "I ended up moving the life factors to their respective category, which then left the insulin category all by itself, which aids in clarity."

> "With the new notes, I split one larger treatment category into 2 smaller categories. I added a new category as a result, and removed extra notes from the broader category."

> "I maintained the general structure but partitioned new information in respective categories."

Looking at the events occurring in the smart categories, we continued to see gradual refinement (Figure 5.12) similar to what was occurring before the new clips were added (Figure 5.11), albeit at a much lower intensity than initially. The majority of the activity was in exploring categories, as individuals most likely wanted to examine and verify the new data was correctly pulled in, and in curating category content, as participants continued to filter out and update some clip summaries. When we asked users about how they refined their structures in response to new information, users mentioned mainly small organizational or filtering adjustments, rather than any drastic change to their structure. Only 5 of the 17 participants during the Distil condition and 3 during the simple condition answered yes to the question: "Did your structure from the first phase change?" This suggests that the new information didn't prompt a major restructuring, possibly because it was randomly sampled and participants had already encountered the major topics involved. The current evidence fits with the structural refinement and tuning behavior noted by Zhang et. al in their consolidated sensemaking model [244]; an approach using a more curated hold-out set designed to introduce a larger structural change may be fruitful future work.

5.6 Discussion

Overall, the lab study provided converging evidence suggesting that users found smart categories useful as a way to perform quick, efficient, and continuous categorization of information they might collect during sensemaking. As users went through more and more information, they were able to continue to add and adjust their categories, refining and focusing them until they only contained the most important information. The categorization was largely "quick and dirty" – users still had to do some significant pruning of information, adjust which information was summarized, and manually consolidate similar pieces of information. However, the smart categories allowed them to quickly pull the right pieces of information into one section which they could then further manipulate to support their sensemaking needs. These categories could be updated with new chunks of information as the user continued to explore, which they could then later incorporate into their document. Participants found this process to be

more useful and efficient than searching through and manually transferring information from the information clips into a final document.

5.6.1 Limitations

While Distil is designed as a general purpose sensemaking tool, there are some situations where it provides limited support. Due to its keyword-oriented design, it can be overly greedy in pulling in information, adding irrelevant clips to a category. Additionally, there could be ambiguous phrases across clips, for example one clip could be talking about Jaguar – a model of car – while another could be referencing the animal, though past literature suggests ambiguity is rare when a term is used under the same context [75]. Three possible solutions include breaking up larger clips into smaller, more focused ones, using more advanced text processing and searching (e.g., extending the word vector approach we use for suggesting the category a phrase should be added to), or using named entity recognition techniques to address ambiguity and provide more accurate filtering functions.

Distil does use some basic information retrieval techniques to support quick organization, there are a number of opportunities to further expand on these to both decrease the user's mental effort, and ensure summarized information is comprehensive and representative of the data collected. One approach would be automatic category generation - using an unsupervised clustering algorithm and word embeddings, users could have categories suggested to them based on the data that has been collected. Another could be more structured entity extraction and presentation where Distil could pull out specific entities, such as phone number, addresses, or headings, and utilize those as components for creating even more structured organizations [27, 63]. These would be especially useful in the later stages of the sensemaking process, where a user will more likely want specifics about the topics and items they've focused in on.

One possible danger of this approach is it could make users only perform surface level evaluations of their collected information, rather than considering it more deeply. One of our participants directly pointed this out: "It was definitely very helpful but I thought it was easier to neglect some details. In the plain text, in order to filter/categorize well, I needed to really read each note. In the advanced system, I just had to type the category and it would automatically categorize information (useful and not useful information)." This might have been due to the artificial situation imposed by the study: users were categorizing information they didn't directly clip, so they were also doing a lot more discovery than a traditional sensemaking scenario. We chose this study design due to the time and mental load concerns from performing a full-fledged sensemaking task, as well as concerns that different users would collect vastly different amounts of content depending on how much they cared about the provided topic. A larger deployment study would help to provide a clearer picture about how much the smart categories are used for additional exploration versus pure categorization.

5.6.2 Design Suggestions

Based on the above feedback and observed interactions, we have a few suggestions for the design of future automatic categorization tools for supporting sensemaking.

First, transparency and low-cost adjustment interactions were key. While we had a number of users complain about having to prune notes from their categories, the cost of removing an item was one click. Users could also easily observe how the keywords they added to a category affected the content, and revert their action if it wasn't what they intended. Second, users need a way to trim down information into a more consumable / summarized form and interleave those with their personal thoughts. While the categories serve as a nice way to quickly pull together the relevant information to act on, the actual sensemaking occurs when a user parses those notes and forms some opinion about the information. Capturing this process in relation to the supporting evidence can allow users to more easily remember their decisions and share them with others. Lastly, as new information is collected into the categories, users need to be aware of the change and given the option to act on it. In Distil, this was a small notification where users could add the additional notes to a category. This could be implemented in a slightly more automatic manner, where new information appears as "unread" and would require confirmation by the user to stay in the category.

Looking forward, Distil is a first step at supporting evolving structures during sensemaking. The smart categories offer a unique capability, in that users can use them for both proactive planning – knowing they are going to come across beaches on a trip to Spain, so they proactively create a beach category – and retroactive retrieval – they just discovered that traveling by train is the best way to go, so they create a train category and see how many destinations talk about trains. This also happens at a higher level in many sensemaking tasks: a user who is planning a vacation knows it's going to be a complicated process, so they might front load some of their organizational efforts based on their past experiences. Conversely, someone researching a humidifier might expect there to be only a few options that they might choose from; however after learning of different types, brands, models, etc. of humidifier the user might recognize that they need help keeping track of all of this information. Developing techniques that are able to flexibly respond to changes in a user's research trajectory could help to produce tools that are more readily adopted by end users, as they could exist as lightweight collection and organization tools and then morph as a user's needs grow. Some possible instantiations of this would be tools that allow users to transition between structures (like convert a simple outline into a table), find additional information about an entity or topic from other web pages that they've already visited or are planning on visiting, or providing categories a user should explore / consider based on the structures produced by other individuals performing the same task.

5.7 Takeaways

Distil's unique take on categorization, where users can perform structuring at any point in their process, e.g. "just in time", follows with Bento's approach for generating sensemaking tasks. Having this flexibility in a tools allows users to leverage the support when they need it, without getting in the way early on, and without incurring huge startup costs later in the process. Similarly, because users during this process are simultaneously trying to filter out unimportant information as well as categorize and learn "bigger picture" features such as distribution, tools need to to feature both top-

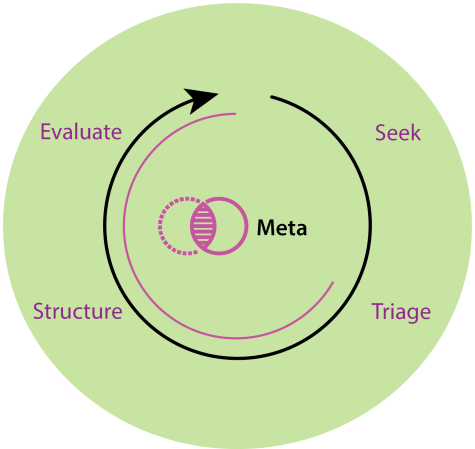
down and bottom-up filtering and aggregation tools. Additionally, user's opinions and needs will change through this process, so these tools also need to provide users with ways to continuously modify the parameters of the aggregation and filtering. Distil, with it's smart categories, allows users to pull in and aggregate topically similar information with keyword filters, while allowing for responsive changes to the structure as more information is learned. Future tools could be designed to go a step further than Distil and Bento by recognizing tasks a user is working on, extracting information from those sources, and providing the user with a set of tools for filtering that information or aggregating it to understand properties such as distribution or identify outliers.

Chapter 6

Meta: Extracting, Structuring and Evaluating Potential Options and Sources [1]

Through Bento and Distil, I was able to explore strategies for supporting the searching, triage and organizational aspects of the sensemaking process. While these approaches help users manage the data they're collecting, they don't support users while they are making decisions and taking actions – which one might argue is the entire reason for sensemaking in the first place. Therefore, I explored ways we might support this final stage of the process. Additionally, I began to consider ways we could further quicken the process of extracting information – while the clipping features of Distil were good, they still required a lot of work on the part of the user. Was there a way that we could allow users to get immediate support from a tool, which would take the information they have been looking at and process it in some meaningful way?

This led to the development of Meta – a system that leverages product entities from a user's open tab and collates them. Within the interface of Meta, users then have a number of tools available, such as a scorecard, a brand table and a sortable, filterable compiled list, to explore promising suggested options. In this chapter, I detail the design and development of this system, and how users were able to use this information to make more consistent product and source trust judgements.



6.1 Introduction

The web has become a crucial source of information for decision making tasks in categories from shopping, to medical illnesses, to travel. As the amount of information and offerings on the web has grown, for many decision making

[1] A portion of this work is currently under review for publication

tasks there's been a simultaneous increase in both the number of options to choose from [197] as well as the number of opinions on what we should choose. On Amazon alone there are more than 230 million product reviews [171], with some individual products having more than 10 thousand reviews [4]. Furthermore, many of these product reviews may be biased or fraudulent, with some estimates suggesting that the majority of reviews for certain categories — such as electronics, beauty, or supplements — may be fake [67].

As a result, there has been a rise in "second order" review sources: authorities that aim to wade through the morass of information online or to collect multiple direct reviews in order to ascertain what the "best" option is for a given category. Previously limited to a small number of authorities such as Consumer Reports, the past decade has seen the emergence of dozens of new authorities such as Wirecutter, the Strategist, Reviewed, and many more [31]. It is no small irony that the same expert review sites aiming to solve information overload by providing curated and aggregated recommendations are now themselves the cause of overload. Deciding between expert sources becomes its own challenge, as described by a popular Vox feature on the best of everything: "Do you cross-reference them all and pick the brand that appears most frequently, your brain quietly short-circuiting when you discover that some reviewers support the Contigo 'Byron Vacuum-Insulated' mug while others swear by the Contigo 'Autoseal West Loop'?" [31].

The challenge of cross-referencing expert review sites is exacerbated by the rapid rise of online affiliate marketing approaches which can result in products being promoted in exchange for pay. Hundreds of thousands of affiliate marketers generate sites aiming to drive traffic to marketed products in exchange for advertising fees, generating over £8.9B in sales in 2017 [1]. Many of these sites use a "listicle" or "best x" format similar to the expert review sites described above, and can be difficult to differentiate based on their appearance in the search results or even website content. Furthermore, even established expert review sites earn commissions through affiliate marketing: as noted by Floyd et al., these reviews and links from critics can have a significant impact on product sales [71]. Likewise, due to sponsored content and affiliate marketing programs, review sites can be financially motivated to promote certain content or brands. Adding to the challenge is that it can be detrimental for sites to reveal any association with affiliates, as consumers can adopt resistance strategies to known sponsored content [221].

In summary, consumers today face two difficult challenges: cross-referencing options from different sources to determine which are widely cited and likely to be good, while simultaneously trying to determine which of those sources are legitimate and which products on those sources might be being promoted for reasons other than intrinsic worth. These challenges are each difficult for a consumer to solve on their own due to the lack of support for cross-referencing or getting a summative view of a source, and are even more challenging because each decision is highly interdependent on the other. For example, a consumer might realize that a source is less trustworthy if it recommends products that aren't mentioned by other sources; however, doing so requires cross-referencing products from those sources in the first place.

In this chapter I identify these dual-sided challenges and explore approaches to address them. Specifically, I investigate mechanisms to cross-reference options and

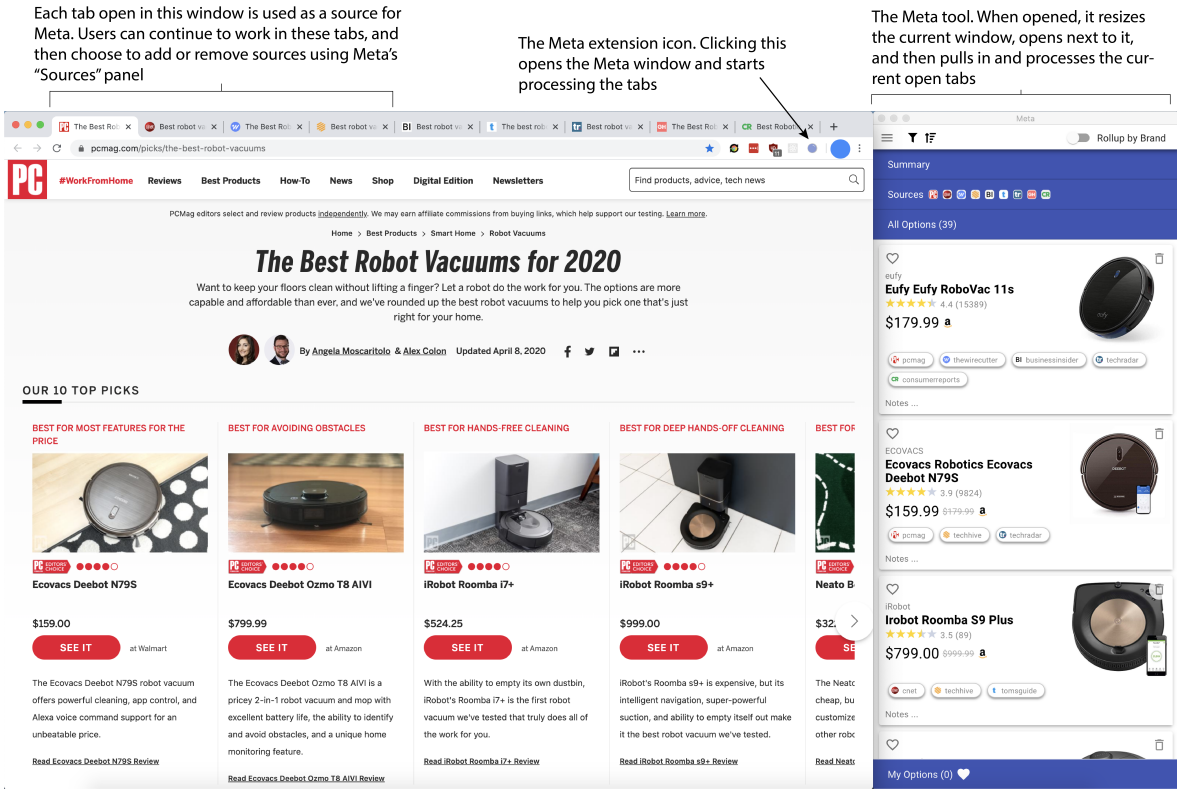


Figure 6.1: The Meta interface next to the active window it's linked to. When Meta is opened using the extension icon, it links itself to that window and continues to monitor any tabs added or removed from it

sources and surface that information in ways that help people make more informed consumer decisions. This approach is instantiated in an interactive system, Meta, and we use the system to probe whether users can make use of these mechanisms with real world, noisy data. We also probe how interacting with the system impacts users' perceptions of option and source credibility. Three studies demonstrate how Meta can: 1) combine existing metadata and simple header detection to enable accurate entity detection and resolution; 2) help users understand the landscape of options as distributed across sources; and 3) impact user perceptions of source credibility and utility.

6.2 Related Work

Significant work has gone into understanding how users utilize consensus and disagreement from multiple opinions in order to make decisions [204, 229]. Consumers utilize other's opinions as a means to reduce cognitive effort or uncertainty as the perceived risk of purchase increases [64, 190]. These opinions can often conflict with each other [99], have different weighting for certain experiential attributes [203], or not agree with the reader's underlying value system [26], resulting in mixed utility for individuals [99].

For low cost means of gathering information, such as web search, users are often

motivated to seek multiple opinions as a means of reconciling such differences [178, 203]. As they find additional information, users are performing a value judgement [228], determining the contextual utility of the information based on a variety of different attributes, including provider trust [205], credibility [153], motivation for participation [97], similarity, personality [9], and passion [47]. However, as the amount of easy to access information increases, so does the cost of analyzing it, which can result in satisficing behaviors leading to sub-optimal decisions [107] and choice deferral [59]. Steckel et al. suggests that interactive tools, such as shopbots, can potentially serve as a solution by supporting this decision making process, however the required level of interactivity from the user has to have a sufficiently observable benefit [207].

6.2.1 Supporting Trust

As users continue to find information during the decision making process, they have to choose who to trust under limited time and attention [7, 207]. This credibility and trust comes from several factors. Fogg breaks down this trust into four forms of credibility: presumed credibility, surface credibility, earned credibility, and reputed credibility [73]. In two large scale studies, Fogg et al. found that in practice the "look & feel" of a website had the biggest impact on credibility [72, 74]. This finding was also supported by McKnight and Kacmar, who found that the professionalism of a site had a significant impact on trust [156]. Aside from website appearance, search result ranking [92], the reputation of a sites domain [77], previous interactions with the site [36], and peer recommendations can have a significant effect on the perception of an individuals trust in a site [199].

Researchers have developed a number of tools and approaches for both assisting website creators for developing more "trustworthy" sites [106, 202], as well as augmenting the information for end users to evaluate trust with. Both Schwartz et al. [198] and Yamamoto et al. [235] augment users' browsing experience with additional site information about credibility, such as who is visiting the site, or the freshness of the website. Tools such as del.icio.us [173] and CredibleWeb [103] offer a more crowd-based approach, where popularity and crowd judgements can help users uncover useful and credible sites. Finally, algorithms such as TrustRank [86] and CredibleRank [37] use the link structure of the web to determine a credibility score for sites.

Rather than relying on the reputation, authority, and previous experience with a domain as a means to establish trust, Gil et al. proposes the notion of content trust, where the underlying information and claims drive trust in sources [77]. Instead of utilizing top down models of trust, quality features [172] such as information provenance [163], agreement, and contextual relevance of the information create a bottom up structure for accessing the credibility of information. Researchers have begun to further explore this concept by suggesting enhanced semantic web page structures [29, 79], analyzing the sentiment and linguistic features of sites [76, 173, 225], or by highlighting disputed content on website with outside data [68, 235]. In Meta, we explore how to support content trust in this option-centric decision making context through extracting and collating entities.

6.2.2 Recognizing and Extracting Options

In recent years, we've seen a dramatic increase in the focus on recognizing and supporting entities in web search. Up to 85% of web search traffic has an entity bearing query [82, 141], resulting in significant effort to develop large-scale entity databases [157] as well as augment search interfaces with items such as entity cards [30, 160] and answering factual questions [85]. Researchers have begun to develop tools to take advantage of this rich entity information available from databases, such as Wikify [159] and Experience-Infused Browser [91], to augment the browsing experience for users. Concurrently, more web sources have begun to augment their pages with structured, entity data through semantic web markup [22, 116, 239]. This has further bolstered search engines' ability to expose and summarize entity query information [60, 238, 240], unfortunately few end user tools exist to take advantage of this rich data [115]. Meta leverages these two developments, a large scale entity product database (Amazon) and embedded semantic entity data, to drive its unique, rich experience. By extracting and collating entities, Meta is able to serve as an end-user content trust system for entity-centric tasks. This steps beyond entity based site augmentation [91] and top-down trust systems [86, 198] by utilizing entity agreement between web sources as a means for judging source and options credibility.

6.3 System Design

To tackle these issues of overwhelming information and uncertainty, we introduce Meta, a tool for collating and cross-referencing mentions of entities on webpages [2]. Meta allows a user to select a set of webpages, which are then processed for mentions of product entities from the page metadata as well as the headers on the page. These product mentions are then collated from the different websites, and presented in an aggregate format through which the user can explore. It is designed around the concept of supporting two-way trust judgements — where individuals are trying to simultaneously evaluate which sources they should listen to and which options are actually the best. Outlier products, in either brand or rating, are tied back to the sources they came from, providing users with a way to critically judge and modify the sources that make up the list of suggestions. Meta, in its design as a prototype system, is built to work with products, but could conceptually be extended to include other types of entities, such as restaurants, locations, or even more abstract entities such as exercise regimens or healthcare plans. Additionally, due to technical limitations, it currently only works with Amazon products, but could be extended to work with other shopping sites with additional API access or parsers.

6.3.1 Formative Study

We conducted a formative study to elicit perceived benefits and challenges around the concept of collating options across sources. Our initial concept of Meta was that it would parse through a user's search results and show commonly cited products in a

[2] See supplementary video Meta_Tutorial.mp4 for a demo of the system

sidebar along with information about them such as images, excerpts from webpages that cited them, and links to read more. To see how this design aligned with potential users' processes and needs, we performed a small qualitative, speed-dating study [55].

Speed-dating participants were recruited via personal networks, as well as posts on Facebook and local NextDoor forums, all of which linked to an online survey that screened for optimizer behaviors such as seeking in-depth advice, building out detailed spreadsheets, and sifting through multiple pages of reviews that would represent target users for Meta. Those participants that demonstrated optimizer tendencies ($n = 29$) were contacted for interviews, for which they were offered $50 Amazon credit for an hour of their time. The interviews consisted of semi-structured interviews on shopping and decision-making behaviors, and speed-dating 10 conceptual mock-ups posing a variety of solutions to hypothesized problems. The speed-dating exercise allowed us to identify actual needs and challenges in the users' shopping processes, and narrow down on the pain points that had the most potential for impactful solutions.

From the responses, we noted three crucial concerns our participants had with the initial concept. The first was trust — eight of the participants noted that they would only want to have information aggregated from sources that they "trusted" rather than from all their search results. They were very concerned about sponsored content, and preferred sources that had a known reputation, were experts on the subject, or had a more blog-like unbiased appearance. Second was the use of the aggregated option list as a way to drive the creation of a shortlist for further research. Four of the participants noted that this would be a good way to generate and have a smaller set of options to prioritize investigating. One participant noted this would have been useful for their recent RV search — they were so overwhelmed with the number of options that having a list of popular options to start from would have been extremely helpful. Another individual noted that in addition to the list of options, they would like to augment it with summarized information they found. Lastly, three individuals wanted to cross-reference and pull in additional information, such as price or reviews from walmart.com. Two of these users mentioned that up-to-date pricing was a big factor in their decision making process, with one mentioning that they frequently used a tool for book prices from different sellers.

6.3.2 Meta

Due to the strong focus on the theme of trust from our initial interviews, we pivoted our design towards providing data-driven, content-based signals for helping end users evaluate trust for not only options but also the sources citing those options. However, enabling users to choose specific sources (instead of parsing their entire search results automatically) incurred additional potential costs in having users specify which pages to add, as well as potential syncing costs if certain of their tabs had been added and others had not. To address this challenge we reimagined Meta from being separate from their browsing experience to working directly on their active tabs (Figure 6.1). The intuition here is that many people open tabs as a queuing mechanism to indicate sources they would like to process later, or as an active workspace for comparison [66, 88]. Taking advantage of this, we built Meta as a Chrome extension, where users can flexibly add or remove trusted sources to be analyzed by the tool, utilizing their tabs as a means to do

so. We pull in additional information from Amazon, such as price, review rating and count, as a means to further bolster the information signals users have available to them from their sources. As sources are added and information is extracted, users can develop a sense of popular and potentially "good" options through a variety of aggregation and filtering methods, such as brand, Amazon ratings, or price. Additionally, users can also evaluate the "goodness" of sources through a source "scorecard" (Figure 6.4), where they can see how a source stacks up against others based on the brands and products recommended. As individuals proceed through their evaluation of sources and options, they can record and summarize information about the options on the product, as well as mark as "favorite" certain sources and options that they view as particularly promising. In the next few sections we go into deeper detail on how Meta extracts entity mentions from sources, how it gives an overview of the options available, and what tools it provides for allowing users to judge what "good" options and sources are.

The primary interface of Meta is a separate sidebar window, designed to accompany the primary information sources. The prototype was implemented in Typescript, using the React library for building UI components, with a Redux store for managing the source and entity information. Because it is a Chrome extension, Meta is able to readily access and extract the information in the individual tabs using content scripts, as well as pull from multiple cross-domain sources and APIs, such as Amazon.

Extraction Pipeline

As mentioned above, Meta is designed to work with a user's tabs as a base for its collation. When a user clicks the Meta extension icon, Meta considers any open tabs in the current window to be a source, and begins to analyze them (Figure 6.2). For a source, Meta checks if there is any JSON-LD, RDFa, or Microdata metadata present on the page. These formats can describe structured semantic data about a page, including places, products, reviews, or businesses mentioned on the page. Next, for each source, Meta proceeds to enumerate all of the visible headers on the page and group them by their styled appearance: font weight, font family, font size, as well as color. These groups of headers are then sorted by font weight and font size, with the intention that the more prominent headers are more likely to be the entities the page is recommending. Finally, for each of these groups, Meta finds the parent container for each header, for each of which, it searches for an HTML ID attribute that can be used as an anchor back to this content, any Amazon links, and the text that would represent the corresponding body content for this header.

After this data is collected from each source, Meta then attempts to determine the primary set of entities being discussed on the page. If there is any metadata on the page, Meta checks for one or more Schema.org Product types[3] in this data. If found, it adds these to the list of potential entities; however if there are none, it falls back to the groupings of headers. To avoid navigation style headers, and general directional headers (i.e. What to Look For, Explore More), we do some additional processing of the header groups. We check for both a large number of duplicate headers as well as headers composed of mostly common English words, and remove those from considerations. Then, starting with the four most prominent common header groups, we check to see if

[3]https://schema.org/Product

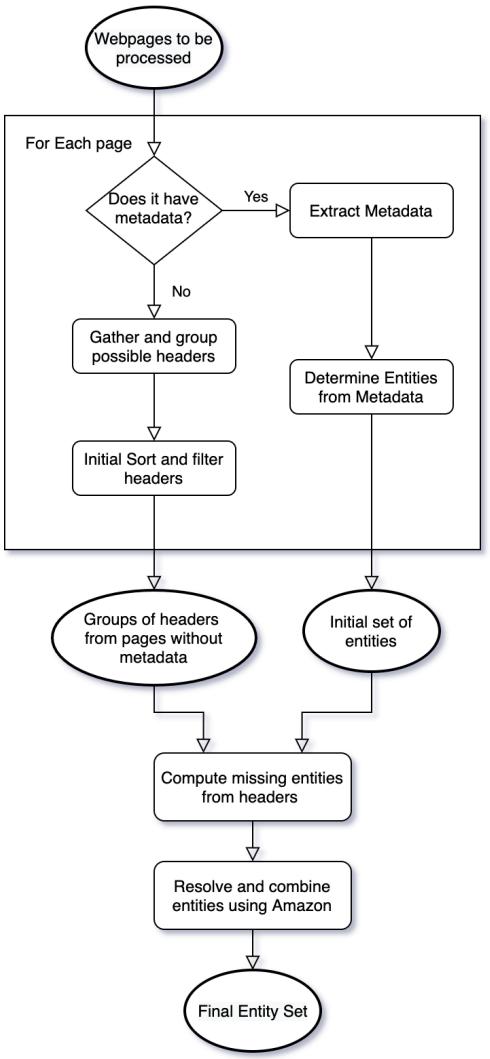


Figure 6.2: The Meta entity extraction pipeline

there is a strong string similarity between one of the headers in the groups and either an existing recognized entity or a header from another page using Jaro-Winkler string matching. While this algorithm was designed and evaluated for extracting products from "listical" style pages, more traditional entity recognition solutions [168] could possibly be applied for blog-style pages.

After the proposed set of entities is determined, these entities are resolved and deduplicated using Amazon. For a proposed entity, if they had either an on-page Amazon link or a link in their metadata, the corresponding product number is associated with that entity. In the case where a link is not present, the proposed entity name is searched on Amazon. For the top five results returned, a token-based string matching is used to rank the results. The result with the highest match's product number is then used to resolve the proposed entity, and in the case of a tie, the result with more reviews is used to break the tie. Finally, after all of the proposed entities have been resolved to Amazon products, entities with the same product number are combined, with this

serving as the final set of entities used by the Meta interface. After the initial set of entities are calculated, users can choose to add or remove additional sources, using the sources panel. When that occurs, the entire entity set is recalculated. In order to judge the accuracy of this heuristic process, we ran an evaluation against top listical websites which is discussed later.

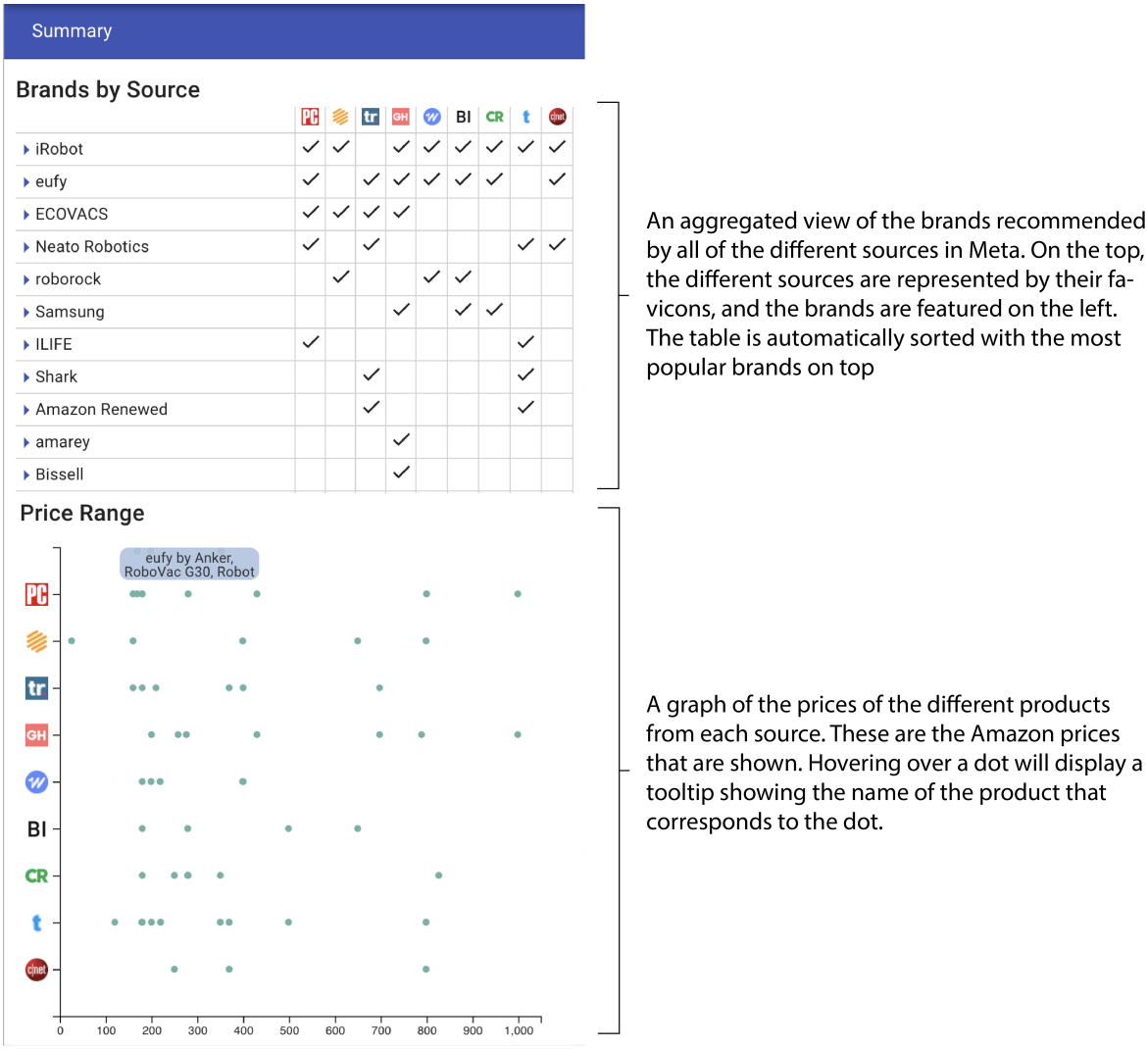


Figure 6.3: The top Summary section of the Meta interface. This shows an aggregated view of the data collected by Meta

Because Meta is performing this collation at a number of levels (between sources, between options, against Amazon) we designed Meta in a series of four panels, progressing from a high level overview of the options and source space, to a detailed look at the options and sources available.

Summary Panel: Providing a Landscape

The top-most panel is the summary panel (Figure 6.3), which aims to help users gain an overall perspective on the distribution of products and prices across sources. This

is meant to more quickly contextualize a source through two means. The first is a table that breaks down which sources mention which product brands. The table is automatically sorted with the most popular brands at the top, allowing individuals to quickly gain a sense of which brands are more common and, conversely, which sources mention those brands. Aggregating models by brand proves useful for two reasons: sidestepping issues where slight variations in product models or titles make cross-referencing difficult; and helping users get an overall view of the manufacturers in a space which could be more manageable than models. Users could also click on the brand name to expand the table and see specifically which models are mentioned by each source. The second means is a price range chart, which shows the distribution of Amazon prices for each source, with each dot providing hoverable details. Here users can get a sense of whether sources focus on high or low end options, what the typical price range looks like, and whether there are any outlier models or sources. While these two views offer a decent first step at showing aggregated information, future interface could feature more flexible diagrams that could be linked to generalized features of products, such as literary genre for books instead of brand.

Source Panel: Determining Source Credibility

As a user has gained some perspective about the landscape of the option qualities, they can proceed to the sources panel (Figure 6.4) to see how these qualities stack up on a per-source basis. From here, they can also add or remove sources from Meta as a way to further bolster or refine the information presented for each option. The sources panel enables these judgements through a "scorecard", which enables a user to see (from left to right):

- How many options are mentioned by the source
- How many options from this source are mentioned by other sources
- How many brands from this source are mentioned by other sources
- How many brands are mentioned by only this source
- How many options from this source have comparatively low Amazon ratings (lower 25% percentile)
- How many options from this source have a comparatively low number of Amazon reviews (lower 25% percentile).

Together, these numbers serve as proxies for investigating the potential credibility of the source. By clicking on a source, a user can dive into these numbers to see specifically which options are mentioned by this source, along with a small badge with details on options that are unusual by either brand, low ratings, or low review count.

A crucial part of Meta's design is tied to the fact that it is integrated into users' browsing flow, and works on users' active tabs. However, through initial interviews we discovered that while users appreciated this integrated flow, they also wanted more control over the set of tabs Meta processed. They also noted two intentions when closing tabs: either they closed a tab because they didn't want to include the source, or if its content was added to Meta they wanted to close it to reduce tab overload. To support these intentions we designed Meta so that when a user launches Meta, it automatically includes all of the tabs from the window it is launched from (Figure 6.1). Meta only pulls from the current set of open pages, not history, as many of those pages

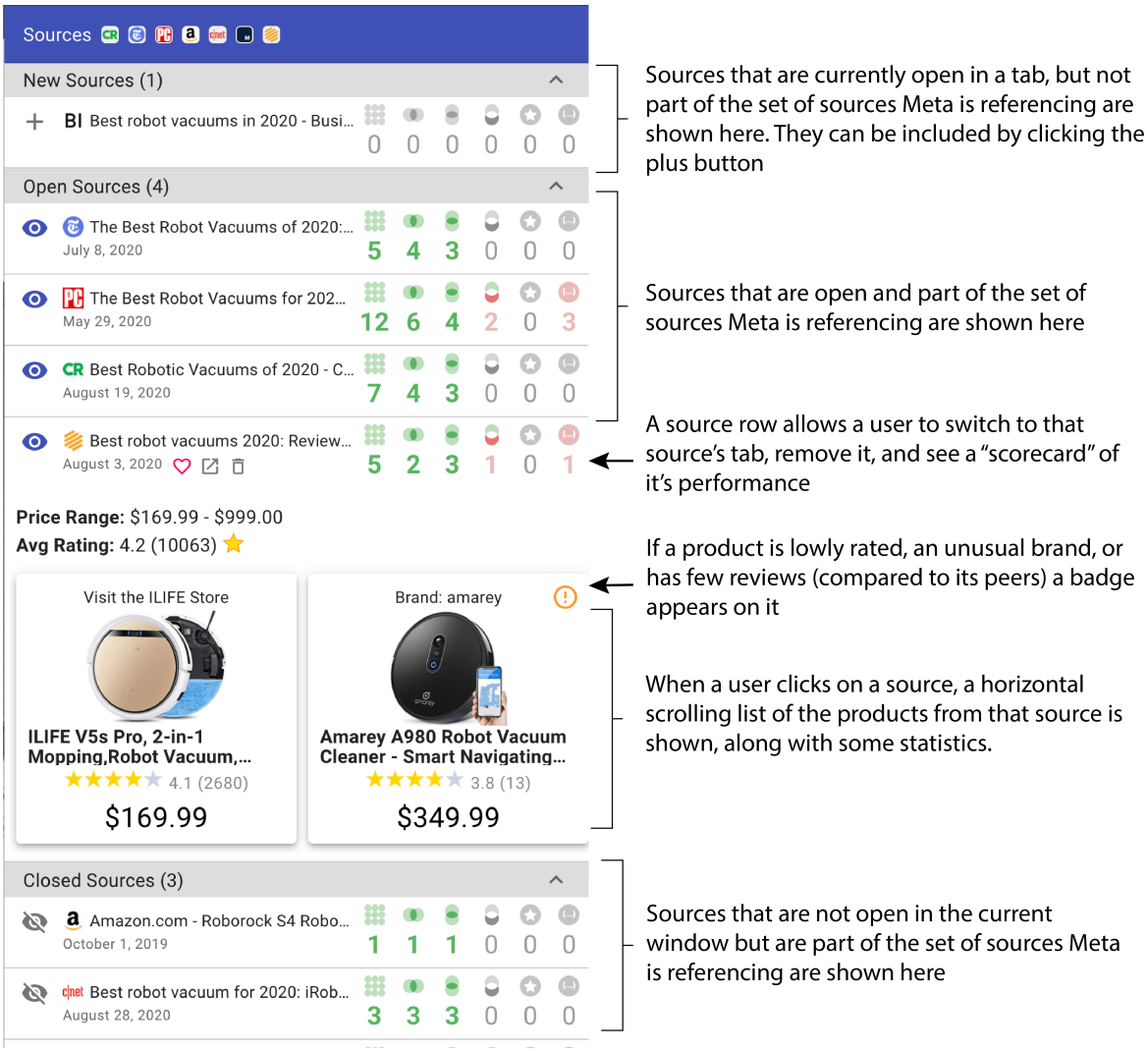


Figure 6.4: The top Source section of the Meta interface. This highlights all of the tabs / sources being referenced by Meta, and allows users to modify that list as they go

might be irrelevant at this point in their process. As a user continues to search for more information, Meta continues to track the tabs that are opened and closed in that window.

In Meta a source can exist in one of four states: open and in Meta, closed and in Meta, open and not in Meta, or closed and not in Meta. Open sources not in Meta are listed as "New Sources", and with a single click it can be added to the Meta analysis. Open sources that are being used in Meta's analysis appear in the "Open Sources" list where they switched to (using the eye icon) or removed from the Meta analysis. If a user closes a source that is included in Meta's analysis, it's preserved in the sources panel and moved down to a separate "Closed Sources" section. From here, users can quickly reopen the source, or continue to use it in their analysis while reducing their number of tabs. Regardless of if a source is open or closed, if it is in the Meta analysis, it will appear in the Summary and All Options panes. Lastly, Users can also choose to

favorite specific sources they think are especially helpful, which promotes all options they cite to the top of the "All Options" list (see below). This results in the list showing the most cited options from favorited sources, followed by the most cited options from all other sources.

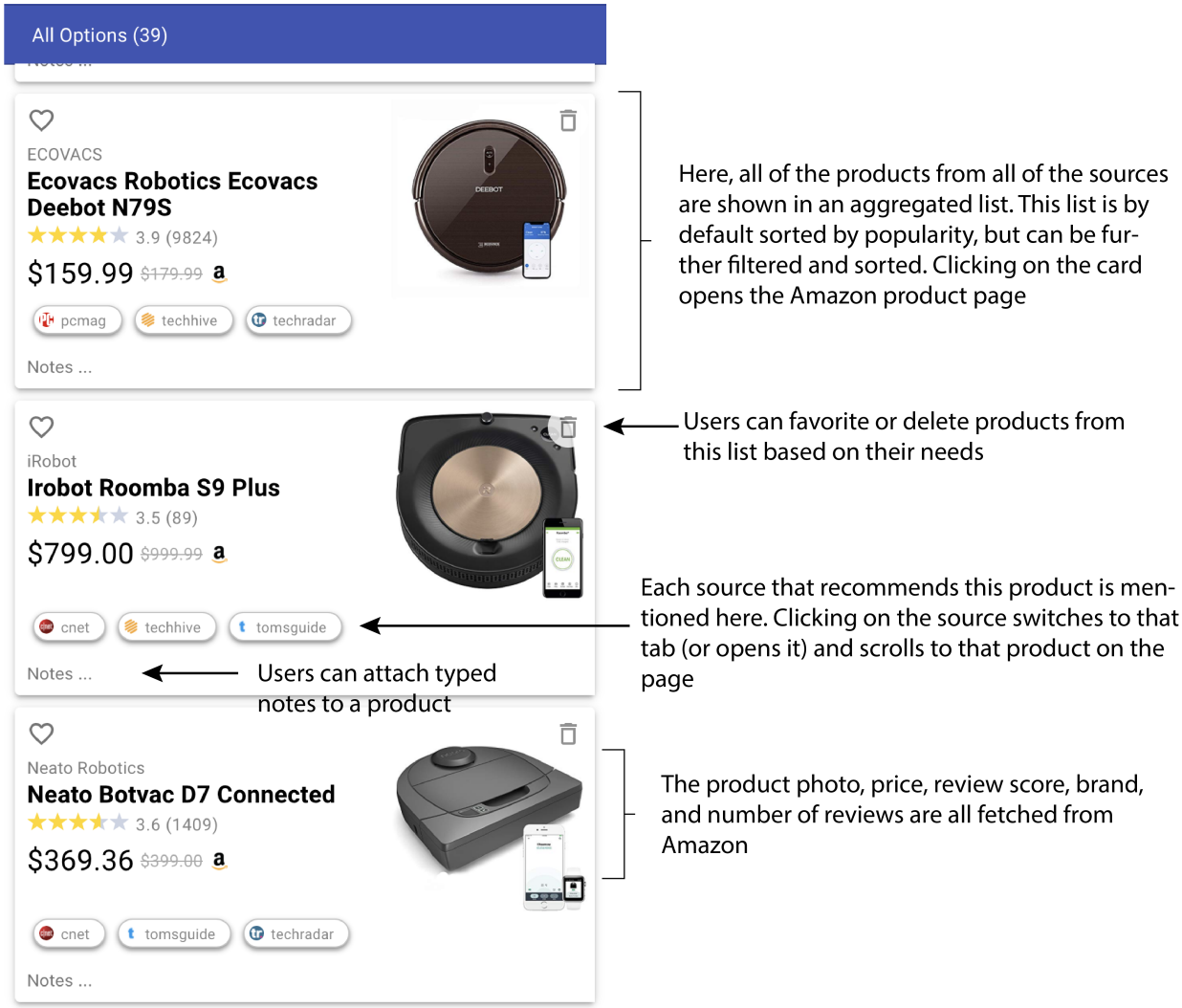


Figure 6.5: The "All Options" section of the Meta interface. Here users see the collated list of products sorted by popularity

Options Panels: Focusing on Individual Options

Finally, this information culminates in a view that combines and collates all of the options from the different sources together. The "All Options" panel (Figure 6.5) shows all of the options from all of the sources, with each option represented as a product card. Each product card includes basic Amazon data for the product, a photo, a list of sources that mention this option, a notes field, and a favorite and delete button. These cards serve as a way for the user to dive into a specific product, and quickly

record their thoughts and opinion about it. By clicking on a card, a user is taken to the Amazon page for that product, where they can dive into reviews there. A user can also dive into what each source said about that product by clicking on the corresponding source attribution. This will bring that tab to the front in the browser window and automatically scroll them to the position in the source that cites this product, allowing meta to function as an option/source based navigation for their decision. Lastly, a user can summarize any findings they have in the notes field, and if an option is particularly promising, they can favorite it. Any favorite options show up in the separate "My Options" panel for easy reference and refinding.

This list, by default, is ranked with the most popular options (options mentioned by the most sources) at the top, allowing users who are looking for a quick shortlist of products to work from a starting point with minimal interaction effort. However, users might have additional criteria or priorities, such as only wanting to consider certain brands, or only wanting the cheapest option available. To that end, Meta has a set of filtering and sorting tools similar to those found on search engines (Figure 6.6). Users can filter the range of prices, minimum Amazon rating, and select certain sources and brands.

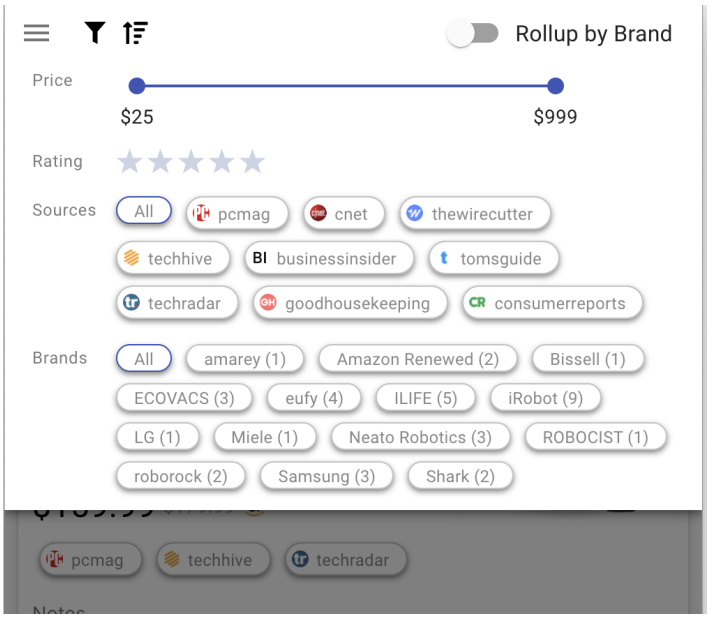


Figure 6.6: The set of filtering tools that can be applied to the list of products in the "All Options" section

While testing Meta with various scenarios, we ran into the case where there would be similar product models that should be considered as a single product in the analysis (i.e. different color models, the 2018 version of a shoe vs the 2019 version). To support this use case, we provided a "roll-up" feature, where users can logically group two products together with drag and drop (Figure 6.7). Rolled up products are treated as one product in the rest of the system (including the summary and sources tab). We also gave users the ability to do this automatically for all of the brands by toggling a switch in the toolbar. When enabled, the "All Options" section becomes a hierarchical

list of brands, with the product models nested underneath (Figure 6.8).

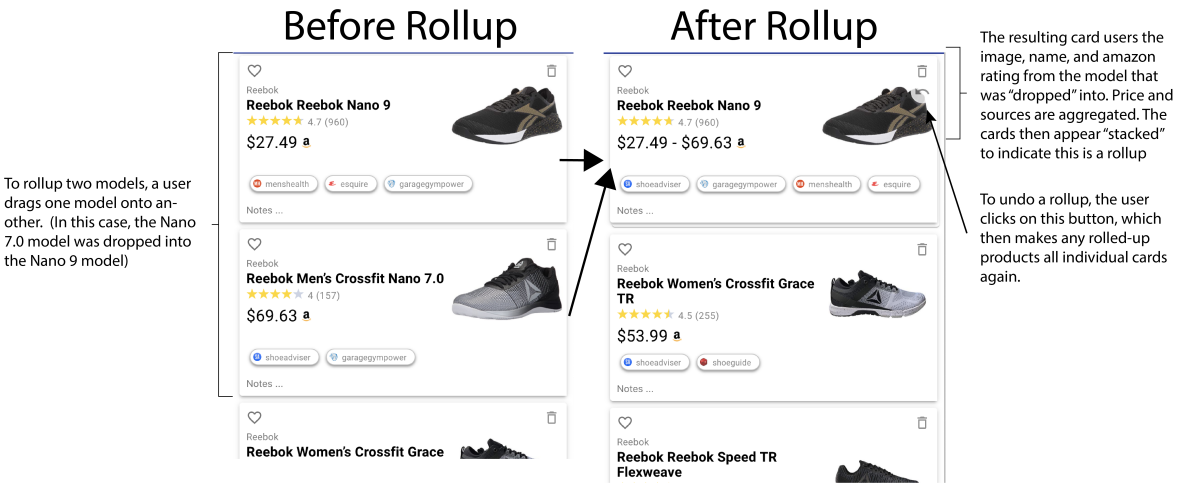


Figure 6.7: An example of two models being "rolled-up"

6.4 Evaluation

By building Meta, we sought to give end users the ability to more accurately and efficiently consider multiple sources of information when making option-centric decisions. Primarily, we sought to bolster the user's ability to judge which sources of information, and which options suggested by sources, were of potential value.

We evaluated the benefits of Meta through three different studies. In the first, we evaluated how well Meta is able to accurately and reliably capture the options being mentioned by different sources. In the second, we explored how Meta impacts the ability of end users to gain a quick and accurate overview of a set of options from different sources. For two different shopping tasks, we were able to show that users were able to more accurately and more quickly understand summary features of a group of options, such as the cheapest product and most mentioned brands. Additionally, we show that afterwards, users chose to explore options with more mentions when using Meta over a baseline. Lastly, we evaluate how Meta is able to impact an individual's judgement and trust in sources in the early stage of the process. We see that with Meta, users are quickly exposed to the certain qualities of the different sources, resulting in a significant change in the sources trusted. Through these evaluations, we show that Meta allows users to gain a quicker and more complete understanding of the options recommended from several sources, and how the data-driven, summary information exposed by Meta impacts users' trust in sources and in options.

6.4.1 Performance Evaluation

Due to Meta's design as a Chrome extension where the end user can control the collection of sources it's pulling from, Meta needs to be able to accurately and reliably recognize entities mentioned on pages. In our series of evaluations, we first wanted

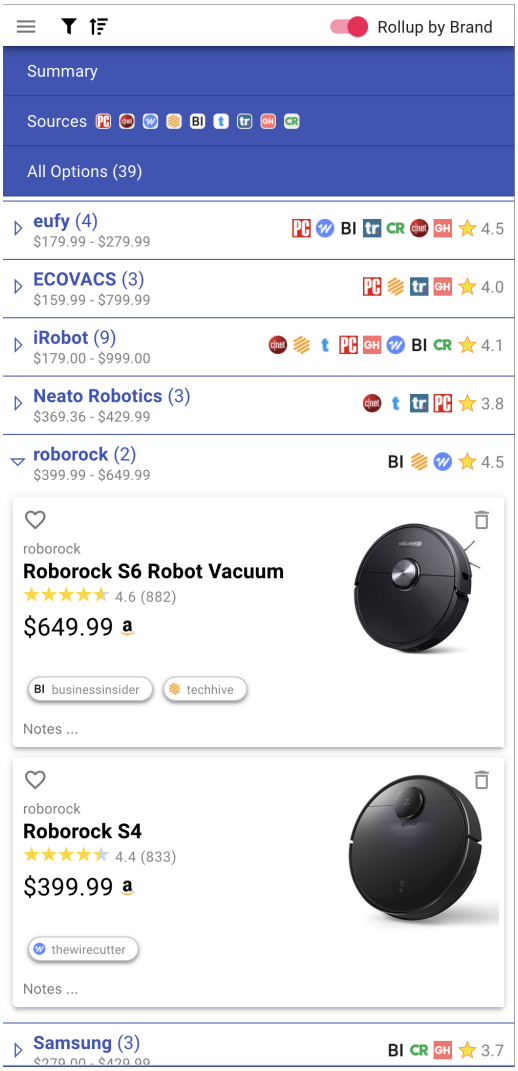


Figure 6.8: The "All Options" section when the brand rollup is enabled

to test how well Meta is able to perform this recognition, and use this as the basis for our additional evaluations. In order to collect a sample of pages that would be representative of general use, we referenced a popular review site, the Wirecutter. We collected their five most popular product guides listed in their 2019 year-in-review, as well as their five most recently updated guides for April 2020, and used the 10 categories of products discussed as the basis for our evaluation. For each category, a Google search was performed with the query “best (product).” For each query, the top seven search results were collected, excluding any results that led to marketplace or retailer SERP pages, such as Amazon or Best Buy. We used the top Google results as our reference due to its similarity to how an end-user might approach looking for sources, and to leverage the existing work done on the Google results to ensure their authenticity and credibility. This collection of webpages was used as the basis for our evaluation for that category in Meta, computing accuracy and recall for each site in the set.

Results

The overall accuracy for Meta's pipeline was 79% — 508 correctly extracted and recognized entities versus 133 either not extracted or incorrectly matched. Of the 133 entities missing, 94 of them were matched to the wrong Amazon product, while 39 of them were missing from the results. From the 70 different pages surveyed for the 10 different categories, 42 unique domains were found, with several review-centric domains appearing multiple times (i.e. thewirecutter.com, businessinsider.com). These pages were all single-page, listical style pages, featuring multiple product suggestions.

We found this level of performance adequate as a probe to explore how users would interact with a system that enabled them to do an on-the-fly meta-analysis of option and source popularity, and somewhat surprising given the historical difficulty of entity extraction and identification [81]. One reason we believe performance was relatively high is the recent growth of curated lists due to their effectiveness in affiliate and content marketing [1]. Creators of such lists are incentivized to provide accurate metadata to promote their pages' popularity, such as in search engines [3]; the 46% of pages we found with metadata provided a base to bootstrap identification even for pages that did not include metadata. Taking advantage of similar trends in public curation may prove beneficial to other system designers as well.

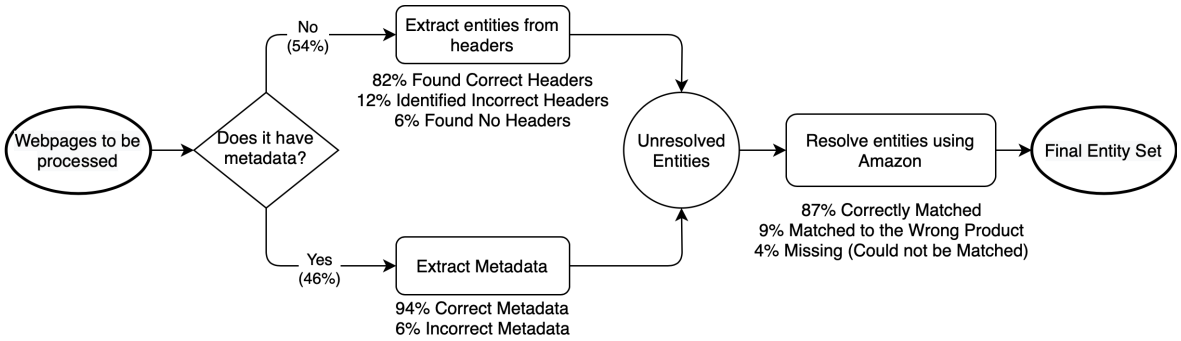


Figure 6.9: The performance of the individual stages of the Meta pipeline.

There are three primary stages in the processing pipeline where Meta failed. Generally, failure occurred in recognizing the correct "set" of entities on the page (the headers were not identified or the wrong set of headers was identified), incorrectly including non-entities into the list of products (i.e. headers like Contact Us or Top 10) or when matching the entity name to the Amazon product – either it was missing for Amazon or matched to a different model (Figure 6.9). In most cases of failure, Meta did not miss just one or two entities on a page, rather it either completely missed a page, or the entire set of entities from a page were false positives. Due to the cascading nature of the pipeline, these early errors propagate, making issues with recognizing the primary set of entities on the page particularly troublesome. This would, from the user's perspective, result in a webpage recommending a set of seemingly random products, as headings like "Contact Us" or "What to Look For" would resolve to the closest matching Amazon product, often times a book.

For each of these errors we provide some means of discovery and remediation. If

incorrect non-entities are included, users can delete the entity from the "All Options" list and it will be removed from all views. For incorrectly matched Amazon products, on each product card, a user can hover over the source attribution to see the original entity name mentioned by that source. If the entity shown in the tool is not the one recommended by that source, the user can then choose to delete it. Finally, users can also choose to exclude an entire source if Meta is unable to recognize the entities within by deleting it from the sources panel.

While the prototype version provided simple ways for users to delete incorrectly recognized entities, a deployed version would need better ways for users to recognize and fix these errors. One possible solution would be to rely on the expected underlying homogeneity of the entities recognized (i.e. they would most likely all belong in the same product category – e.g. humidifiers). Sources that featured a number of products outside this category could be flagged for further review by the end user. Similarly, the string edit distance between the product name on the webpage and its Amazon product name could be utilized to identify situations where the Amazon entity resolution process failed, or spurious entities (e.g. Contact Us) came from incorrect headers. With those identification tools, users could then be lead to manually select the right "set" of headers on the page for Meta to use, and then in the future that training data could be leveraged to build a more comprehensive entity header extraction process.

The other major class of issue, where the product isn't featured on Amazon, could be remediated in a couple of ways. First, more a more robust set of product databases could be leveraged, such as Google Shopping. We found that, depending on the type of product, the entity resolution accuracy differed greatly – for the "Office Chairs" category, the products were only matched on Amazon 63% of the time, while Robotic Vacuums were matched 95% of the time. The other alternative would be to perform an entity resolution process without using Amazon as a ground truth, and instead using a variation of string matching to de-duplicate the entity mentions.

6.4.2 Option Evaluation

Given the above adequate performance on entity extraction and identification as well as methods for ameliorating errors, we now explore how collating entity mentions would impact a user's ability to make sense of a product space. First, we conducted a study aiming to characterize the degree to which Meta would allow users to make faster and more accurate judgements about the range of options mentioned by several sources.

To evaluate the degree to which Meta might be faster and more accurate as compared to reading a set of pages alone, we developed a set of 10 objective questions a user would have for a real-world shopping task. These were based on the needs users reported in our formative study. For example, two of these questions were "What is the most common product or brand mentioned" and "Which is the most expensive product." For each of these questions, we measured the time an individual would take to provide the answer, as well as the accuracy of their answer. This evaluation was performed across two different product categories: "Humidifiers" and "Robotic Vacuums", both of which were used in the previous performance evaluation. Users were instructed to answer these questions using the top seven Google results for that category along with any external information required (e.g. Amazon). While real world usage of Meta would

	Robotic Vacuums				Humidifiers			
	B Time	B Acc.	M Time	M Acc.	B Time	B Acc.	M Time	M Acc.
Model Questions	Baseline N=13		Meta N=15		Baseline N=13		Meta N=14	
Most Popular 3 Products	288.5	23%	**85.9**	**93%**	223.1	33%	117.5	**93%**
How many different sources mentioned product X?	65.2	31%	**22.2**	**100%**	76.6	23%	**26.0**	**86%**
How many different models of brand X are there?	13.8	8%	32.3	**80%**	88.3	15%	**26.0**	**100%**
Brand Questions	Baseline N=15		Meta N=15		Baseline N=12		Meta N=14	
What is the most popular brand?	105.2	73%	**64.9**	93%	155.5	31%	69.3	**93%**
How many different sources mention brand X?	76.5	40%	**19.3**	**100%**	62.9	8%	**15.5**	**100%**
For source X, which brand was not mentioned by any other source?	94.7	53%	**55.1**	**87%**	72.5	50%	**40.9**	**93%**
For source X, how many brands were not mentioned by any other source?	72.3	27%	**30.8**	**80%**	47.7	25%	34.9	**86%**
Cross Reference Questions	Baseline N=16		Meta N=15		Baseline N=12		Meta N=14	
What is the cheapest product on Amazon?	221.4	75%	**43.2**	93%	231.2	67%	**67.6**	86%
What is the most expensive product on Amazon?	109.9	69%	**79.1**	93%	146.4	42%	**37.0**	79%
For source X, which product had less than X reviews on Amazon?	58.8	63%	49.9	**100%**	160.8	67%	**103.5**	93%

Table 6.1: Results from the Meta **(M)** versus Baseline **(B)** speed and accuracy evaluation. Any bolded time result is significantly faster than the baseline using a two-sample t-test, with $p < 0.05$. Any bolded accuracy **(Acc.)** result indicates a significant improvement in accuracy as determined by a two-proportion z-test, with $p < 0.05$. N indicates the number of participants who answered questions for that question segment and task type.

have end users choosing which set of sources to work from, the sources were fixed as a way to control variance and accurately measure the effect of using Meta.

This evaluation was run on Amazon's Mechanical Turk platform. While going through several trial runs of the task, individuals in the baseline conditions didn't have enough time to answer all of the questions. This was due to the questions being significantly harder to answer without Meta, some taking more than five minutes to answer correctly. Because workers on the Mechanical Turk platform generally receive shorter, simpler tasks, the questions were divided up into three groups based on the type of data used to answer them: a product model-centric grouping, a brand-centric grouping, and an (Amazon) cross-referencing grouping (Table 6.1). Individuals in the baseline condition were presented with only one of these groups instead of all three to ensure better task completion. Individuals in the Meta conditions, however, were presented with all three question groupings in a randomized fashion, due to the advantages the Meta interface offered for answering the questions. Individuals in that condition also spent a non-trivial amount of time learning the Meta interface; therefore we wanted to make sure they felt like it paid off (Table 6.1).

Participants in all conditions were first asked a series of demographic and background questions about their previous experience performing online shopping. Next, if participants were assigned to a condition using Meta, they went through a brief video tutorial, and a sample training task where they were asked to answer three questions using Meta for the category "Best Dish Rack". Participants had to answer these questions correctly in order to continue, as to ensure that participants effectively familiarized themselves with the interface before the timed answers.

To motivate participants to provide correct answers to the questions, a \$0.75 bonus was given for each question answered correctly, in addition to the \$2 base amount paid for the task. We aimed to equalize fairness across conditions in terms of incentives; although the Meta condition had additional opportunities to earn more bonus money faster, this was balanced out by the time taken for a training task and learning a new interface. Through this payment structure, we aimed to pay workers at least \$10/hr, and found after completing our tasks, on average workers were paid at least \$11/hr. After participants answered all of the questions, they were asked to complete some qualitative questions about the easiest and most difficult questions. Finally, participants provided feedback about the credibility of two of the sources utilized in the task, based on their previous experience with the sources as well as their in-task experience. We chose a high and low agreement source based on their numbers in the Meta scorecard. Credibility judgements were elicited for four aspects: source reputation, review process, products recommended, and perceptions of bias. These categories were selected from Gil and Artz's evaluation of end user trust in web sources [77]. For each of these aspects, individuals rated the source on a five-point Likert scale from very untrustworthy to very trustworthy. Lastly, we had them provide the top three products that they would either recommend to buy or perform deeper research into. These were compared across conditions to see if Meta usage significantly impacted the models individuals trusted and recommended.

Results

The Humidifier task included 52 participants (age 23-68 $\bar{x} = 37.1; \sigma = 9.8$; 33 who identified as male and 19 who identified as female) and the Robotic Vacuum task included 59 participants (age 19-65 $\bar{x} = 36.5; \sigma = 12.2$; 40 who identified as male, 18 who identified as female, and 1 who identified as non-binary). For both of these groups, individuals who spent less than 30 seconds on average for the main set of questions were automatically excluded, as that would have been the minimum interaction time to even attempt to accurately answer the question. A total of eight responses were excluded from the Robotic Vacuum task and seven from the Humidifier task (these were also excluded in the above demographics).

For baseline participants, the Humidifier task took approximately 20 minutes to complete in the baseline conditions ($\widetilde{x} = 16.7, \sigma = 11.3$) and 40 minutes to complete in the Meta condition ($\widetilde{x} = 38.8, \sigma = 19.3$). Including the bonuses, individuals were paid on average \$12/hr in the baseline conditions ($\widetilde{x} = 11.6, \sigma = 8.1$) and \$16/hr in the Meta condition ($\widetilde{x} = 14.7, \sigma = 7.7$). For the Robotic Vacuum task, baseline participants took approximately 20 minutes to complete in the baseline conditions ($\widetilde{x} = 17.8, \sigma = 11.3$) and 38 minutes to complete in the Meta condition ($\widetilde{x} = 36.0, \sigma = 19.2$). Including the bonuses, individuals were paid on average \$12/hr in the baseline conditions ($\widetilde{x} = 9.0, \sigma = 6.9$) and \$12/hr in the Meta condition ($\widetilde{x} = 10.8, \sigma = 6.1$).

To characterize how Meta affected performance, we examined both the accuracy and speed with which participants answered the 10 questions. It is possible that given the difficulty of some of the questions, some participants would trade accuracy for speed, and we found evidence for this in the baseline conditions. Even within participants, some appeared to "give up" on certain questions and move on, though which questions were abandoned differed by participant.

However, despite these tradeoffs we found that participants using Meta were both more accurate and faster overall than the baseline condition (Table 6.1). Overall accuracy in the Meta condition was 92% ($\sigma = 0.07$) versus 41% ($\sigma = 0.22$) in the baseline conditions. As seen in the above table, the accuracy rates for the baseline condition were very low for a number of questions as compared to the Meta condition. For all Meta questions, individuals were able to achieve a minimum accuracy of 80% for all of the questions. Even including all of the incorrect answers, individuals using Meta were still able to answer questions on average 2.6 times ($\sigma = 1.25$) faster than the baseline condition. This speed improvement is a fairly conservative estimate since it includes incorrect questions in the baseline with much shorter (e.g., $<$ 10 second) response times that were likely "given up" on.

Affecting Trust

After participants answered the above questions, they were asked to rank the credibility of two of the sources, as well as choose the top three product models that they would either buy or would be their top candidates for additional research. By observing a difference in the baseline source credibility rankings, as well as the users' current top models, we aimed to explore how Meta was impacting how users trusted sources and the options those sources recommended.

Robotic Vacuum Task

	Baseline N=132		Meta N=45	
	Average	St Dev	Average	St Dev
Mentions	2.26	1.45	2.80	1.39
Price	444.77	325.33	496.44	325.32
Rating	4.15	0.38	4.26	0.34
Num Reviews	4,108.37	4,804.83	4,346.49	5,144.34

Humidifier Task

	Baseline N=126		Meta N=45	
	Average	St Dev	Average	St Dev
Mentions	3.48	2.18	4.13	2.18
Price	91.35	109.91	77.00	91.89
Rating	4.12	0.23	4.23	0.22
Num Reviews	5,470.13	7,056.24	7,963.33	9,012.72

Table 6.2: Summary statistics from the model recommendations made by participants. Each model was broken down into 4 features, which were averaged together across the participant recommendations. Any bolded feature was found significant by a univariate ANOVA analysis, with p $<$0.05.

To evaluate whether surfacing the number of sources mentioning an option impacted option trust we analyzed whether the three products recommended in the Meta condition would differ from those recommended in the baseline condition. We compared products recommended from the two conditions on four metrics we hypothesized might influence the decision: number of mentions by sources, price, Amazon review rating, and Amazon review count. A MANOVA was conducted on each of the task types, where the Meta condition was the predictor, and the four objective components served as the response variables. In both scenarios, a significant statistical difference was found in the products recommended based on if participants were using Meta or not (Humidifier: $F(4, 166) = 3.23, p < 0.05$, Wilk's $\alpha = 0.93$, Robotic Vacuum: $F(4, 172) = 4.16, p < 0.01$, Wilk's $\alpha = 0.91$). A follow up univariate analysis found that in both scenarios products with a higher number of mentions across sources were more likely to be recommended in Meta (Humidifier: $F(1, 169) = 4.53, p < 0.05$, Robotic Vacuum: $F(1, 175) = 5.14, p < 0.05$); additionally, in the Humidifier condition, products with a higher Amazon rating were more likely to be recommended ($F(1, 169) = 7.46, p < 0.01$) (Table 6.2). These results suggest that Meta's collating and surfacing of sources mentioning a product impacted participants' recommendation decisions.

To evaluate Meta's impact on source trust, participants were asked to judge the credibility of a high, and low agreement source on four factors: source reputation, review process, products recommended, and perceptions of bias [77]. Our hypothesis was that exposure to Meta's source scorecard would result in increased credibility for the high agreement source and lowered credibility for the low agreement source. For the Robotic Vacuum task, there was a significant increase in credibility of the products recommended for the the high agreement source (PC Magazine, $t(39) = 2.08, p < 0.5$), and a significant decrease in the credibility in the low agreement source (Good Housekeeping source, $t(22) = 2.18, p < 0.5$). However, there were no observed significant differences for the

Humidifier task.

Although the results on the Robotic Vacuum task confirmed our hypothesis about impacting source trust, the lack of impact on the Humidifier task prompted us to carefully explore possible explanations. In addition to the possibility that Meta's interfaces do not impact source trust, this exploration resulted in two in-depth explanations.

First, because individuals were performing these evaluations after the objective questions, their exposure to data required to answer the previous object questions allowed them to judge source trust similarly in the baseline condition as in the Meta condition. A study design that asked participants to provide their initial trust judgments about sources (e.g., at the time of seeing them on a search results page) would have provided additional data and would be more representative of the typical user experience, but may have compromised validity for the current study if participants anchored on their initial judgments.

Second, due to the high correlation between the four factors to source reputation ($r > 0.5$ between all factors), we would only see a significant difference where a credibility feature response was counter to the website's initial perceived reputation. Our low agreement source for the Humidifier task was Health.com, but it had an average reputation of 3.5 and 3.4 in the baseline and Meta condition respectively, which suggests individuals in both conditions didn't have high expectations for it and it may have been subject to floor effects. In the Robotic Vacuum task, Good Housekeeping had a higher average reputation of 4.0 and 3.9 in the baseline and Meta conditions, which might have resulted in a larger decrease in trust from exposure to the Meta source scorecard. Unfortunately, the same limitation on the study design in which we did not elicit initial judgments of source trust meant that it was not well suited to distinguish between the possibilities above.

6.4.3 Source Trust Evaluation

To address the limitations in the previous study design, we conducted a followup study in which we elicited participants' perceptions of source trust before and after exposure to Meta's source scorecard. Furthermore, we also wanted to more deeply examine Meta's ability to influence trust early in the decision making process when users are still deciding which sources to explore in more depth, for example on a search results page immediately after typing in a search query. In this scenario a user currently might be faced with many potential sources to explore with a sparse set of signals on which to decide (e.g., domain name, title, search result snippet). To test this hypothesis, we performed a within-subjects study where participants were asked to provide us with three sources they would consider credible before using Meta, and then asking them for an amended list after they interacted with the Meta source scorecard.

This evaluation was run for three different categories of products: the Humidifier and Robotic Vacuum categories from the previous task, as well as for Travel Backpacks (taken from the performance evaluation). For each task, an individual was first presented with a Google search results page for the product category, with the search query being "best [X]", cached such that all participants would see exactly the same page. From the search results participants were asked to pick out three sources they trusted, with which to initially investigate the query. Next, participants were given a link to Meta with all of

	Coeficient	T-Value
Humidifier		
health.com	-0.32	-3.73
consumerreports.org	-0.22	-2.6
popularmechanics.com	0.33	3.87
thespruce.com	0.35	4.15
Robotic Vacuum		
cnet.com	-0.48	-5.06
consumerreports.org	-0.45	-4.78
businessinsider.com	0.34	3.6
pcmag.com	0.44	4.72
Travel Backpacks		
thewirecutter.com	-0.3	-3.45
indietraveller.co	-0.23	-2.62
thebrokebackpacker.com	0.23	2.62
gearhungry.com	0.49	5.66

Table 6.3: The list of sources for each task with a significant change in reported credibility ($p < 0.01$)

the sources from the search results pre-loaded into it. To get participants familiar with the scorecard interface and scaffold their knowledge for answering subsequent questions accurately [119], they were asked to answer three objective questions in Meta: which source had the most overlapping brands, how many products by source [X] had a rating lower than four, and which source had the most brands not mentioned by any other source. The selection of three questions was aimed to balance participants' attention to and familiarity with the information on the source scorecard without requiring them to exhaustively process every source and score. Participants were then asked to choose the three sources they considered the most trustworthy and would use for further research into the task. They were subsequently asked why they chose those sources, if they changed from their initial set of sources, and if so why they changed.

Participants were recruited from Amazon Mechanical Turk, with 50 participants for each of the three scenarios, including 82 who identified as male and 68 who identified as female. The median time for task completion was 11 minutes ($\bar{x} = 14.6, \sigma = 12.1$). Individuals were paid \$2.50 for their participation, resulting in a median hourly wage of \$13.63/hr. Prior to analysis we removed 27 responses that included incorrect answers to the initial objective questions, resulting in 42 responses for the Humidifier and Travel Backpack tasks and 39 responses for the Robotic Vacuum task.

Results

We found that Meta had a significant impact on the sources individuals chose as trustworthy, as measured by the change from before and after exposure to the scorecard. A Chi-square analysis on the distribution of three selected sites before and after participants used Meta revealed a significant difference for each task. To explore how each source was impacted, we constructed a general linear model. For each participant, we calculated a change score for each of the possible sites they could have recommended

before and after Meta. If a site was recommended before and not after using Meta, it was given a score of -1. Conversely, if a site was not recommended before but was afterwards, it was given a score of 1. Otherwise, the site received a score of 0 for the participant. Because this was a repeated-measures design task, we constructed a model with sources as a categorical predictor of the change score, and participant as a random variable. The model reported that for each of the tasks, source was a significant predictor across participants for the change source (Travel Backpacks: $F(7) = 7.55, p < 0.1$; Robotic Vacuums: $F(8) = 10.13, p < 0.1$; Humidifiers: $F(7) = 7.88, p < 0.1$).

With this model, we were able to discover which sources saw significant increases and decreases in perceived credibility after users were exposed to the Meta scorecard (Table 6.3). Qualitative responses from participants suggest several causes for changes in perceived credibility, including the importance of outlier products that did not have overlap with other sources:

> "Health.com seemed like they threw more random, unverified products in, making it seem like they can be bought easier than the others. These sources make me feel comfortable."

> "[My choice] had good overlap, implying consensus picks of quality. They had relatively low outliers and an acceptable average of reviews. I dropped Health.com from my initial selection due to the high number of outliers, implying its list might be out of date or contain substandard items."

However, in the Travel Backpack task, there was a much larger diversity in the backpack recommendations from the sources — each source on average had four outlier brands (x=4.125, SD=4.90). In this case, unique brands were perceived as a positive to some participants:

> "Travel and Leisure has the highest number of unique brands to browse, which is what I want to do. The other two sites are ones that interested me while I was looking at them."

However, other scorecard qualities, such a low ratings, were still consistently useful in this case:

> "I liked the ones that had few bags with low ratings. I also considered the number of outlier products but it wasn't my main consideration. I looked at the percentage of outlier products to the number of product recommendations. In the end I didn't choose the personal blog because it seemed less reliable. I didn't like the fact that he recommended multiple products with low ratings."

These responses illustrate that individuals were actively using the scorecard information to adjust their attitude toward which sources they found credible and useful. However, the interpretation of what was found useful differed, with overlap between sources used to either find popular options and avoid outliers (high overlap), or find unique options (low overlap) depending on the type of task.

6.5 Discussion

Across these three studies, we found that Meta significantly influenced users' perception of trust of options and sources for decision making tasks. Users utilizing Meta to answer questions about the landscape of products were, on average, 51% more accurate and 2.6 times faster. Additionally, users were more likely to choose products with more source mentions and higher Amazon ratings with Meta over the baseline. By using the source scorecard, users made more critical credibility and utility evaluations of websites, based on factors such as unique brands, low Amazon ratings, and product agreement. Finally, our performance analyses suggest that even in its current state as a research prototype, Meta might work reasonably well for many popular tasks, being able to accurately analyze and cross-reference 79% of the entities present on webpages using existing page metadata and simple header detection.

While we were able to demonstrate that Meta was able to significantly impact user's perception of options and sources for product shopping in two lab studies, we believe a field study would provide additional ecological validity in evaluating the impact Meta can have on end users. As noted in the performance evaluation, although Meta was fairly accurate at recognizing the entities from sources (79%), in a real world scenario, missing the last 21% of entities can be detrimental and possibly skew the data. Our two lab studies were designed to avoid misclassification issues that might occur by utilizing sources where recognition was known to be 100%; however some additional solutions could be implemented to either bolster this recognition number or help users recover from recognition errors.

One possible approach for improving header recognition would be to train an ML model to perform the header classification portion of the Meta recognition pipeline. With an ML model, additional features such as surrounding content and semantic header meaning could help to boost the recognition accuracy [81]. The other portion of the pipeline where accuracy could be greatly improved is the matching of Amazon products. Right now, Meta is limited to performing a search and then parsing the Amazon results to find a possible matching product; however the product might not appear on Amazon or could have a slightly different name variant than what is reported as the Amazon title. We could either enhance the title string matching algorithm we use to be more flexible, using more token-based features and weighting, or possibly find an alternative, more complete product entity database to rely on.

Aside from improving the performance for the product recognition Meta prototype, the other direction for exploration would be to expand the system to deal with additional types of entities, such as locations for travel, or even exercises. Expanding the pipeline's ability to recognize other entities presents the challenge of linking with an additional entity datastore, such as Google Places. The other challenge would be adapting how Meta presents and manages the underlying features of the options. For example, brand is an feature specific to products, whereas for a use case like travel, the preferred feature would likely be location. This would involve adapting the summary views to showcase these features appropriately, such as mapping out the travel attractions collected in Meta onto a map. As entity recognition approaches and databases continue to improve, we hope the interaction paradigms explored here may prove useful in improving real world decision making across a variety of domains.

Our motivation in developing Meta was spurred by the proliferation in expert-curated as well as affiliate marketing review sites and the corresponding difficulties that users have in trying to ascertain what sources and options are the best for them. A natural question is how the approach introduced here would perform if this growth of secondary sources continued. On the one hand, we believe that situating Meta as part of the user's natural browsing process means that it should maintain its relevancy as long as web browsing models continue to follow a similar process flow in which a user queries a search engine and chooses from a relatively small set of sources to further investigate. However, widespread use of tools like Meta could raise the possibility of improvements to that process, ranging from better reranking of search results to novel paradigms in which "Meta data" could be used to provide richer interfaces for exploring search results, such as spatial layouts or clusters based on option and source overlap. More generally, the trend of secondary sources providing increasing amounts of accurate metadata (through incentives such as boosted search result rankings) suggests a promising future for tools that can harvest this data and improve user decision making online.

6.6 Takeaways

Meta provided a way for users to gain a high level overview of an information space with a single click. This was accomplished through the automatic extraction and collation of product entities, which in turn enabled users to get an overview without the legwork of processing everything manually. This, again, provided this "just in time" sensemaking benefit we saw in Bento and Distl – when a user starts to encounter too many products to keep track of, they can utilize Meta to combine the numerous recommendations they might recieve. Meta takes it one step further than Bento and Distil through automated extraction, as well as this notion of "information compression", where similar pieces of information from different sources could be combined together. As the amount of information on the web continues to grow, this feature of compression will be essential to ensuring that users aren't overwhelmed.

Chapter 7

Conclusion and Future Work

In this thesis, I explored several ways that tools could be designed to support sensemaking in-situ, with lightweight interaction, and just in time capabilities. From Bento's workspace that provides an enhanced workflow for tracking tasks, to Meta's ability to identify and extract options from listicals, each of these systems was designed to augment user's existing workflows and support their cognition throughout the process. Through tighter integration with user's current sensemaking practices and existing interfaces (i.e. browsers), I developed systems that simplified the means for users to track, manipulate and act on the large amount of data they were exploring.

Because sensemaking is a "human" big data task, dependent on a person's preferences and context, there isn't an easy way to shortcut the process for individuals. While tools like recommendation engines provide a means for users to receive suggestions based on other's activity [133], users are often left asking "why", not understanding how the option applies to them [171]. Websites, such as Metacritic, the Wirecutter, and Trip Advisor, attempt to offload a large portion of the sensemaking to "professionals" and then aggregate and rank the options reviewed, however these sites are biased based on the reviewer's situation [31]. Therefore, users are stuck with the task of reading, compiling, and then evaluating these different sources of information in order to make an informed decision for their situation.

7.1 Strategies

How can we support and simplify this cognitively demanding process so that it doesn't feel so overwhelming and burdensome? Throughout this work, I considered the question: how can we allow users to externalize their sensemaking process so they can juggle the vast amount of information they have process? Across the three systems, Bento, Distil and Meta, I have developed several different strategies that help to accomplish this challenge.

7.1.1 "Just in Time" Sensemaking

Because of the wide set of strategies utilized for different sensemaking tasks, there isn't an easy "one size fits all" tool. Some users might feel like they have enough information

to make a decision after looking at two information sources, while others might feel like they could never have enough information. What makes this more difficult is that users don't often know which of these strategies they're going to use when starting a task. For example, searching for a humidifier might seem like a simple task at first, however as a user starts to discover options and what differentiates them (such as ultrasonic, warm vs. cool mist), the task might suddenly grow to be more than they can handle without cognitive support. Ideally, tools should be able to "step in" at this point and begin to provide support, rather than having individuals have to commit to a tool upfront, creating categories or organizations when they might not need them, or requiring users to go back through the data they've already explored and be forced to copy it into a tool.

Bento, Distil and Meta all explore aspects of this concept, by trying to naturally be useful when a user starts to become overwhelmed. Bento makes a first attempt at this by automatically recording all the searches a user performs and showing it as a "recent searches" list. When a user recognizes that a task is starting to get complicated, they can easily just drag searches together to indicate a more complicated task. Distil's companion, Siphon, provides multiple different ways to collect information, from full pages, to highlighting, to individual sentences. This, combined with the ephemeral Distil categories, lets users collect and form easily adjustable categories, although the idea of "collection" might still be heavy for some individuals. Lastly, Meta lets a user jump from a series of open listicals to a summarized, collated list of available options, allowing the user to click a button and have a tool step in to help them with an overwhelming number of options.

Future tools then must recognize that giving users some initial payoff or reason beyond just "storing information" is required for real adoption. Users already have numerous general purpose productivity tools, such as notepads, spreadsheets, affinity diagramers, but unless a user has recognized early on in their process one of these tools will be beneficial, they'll can easily go unused. Rather, tools that are able to support and even track users through their process, and step in when they are required, are important next steps into introducing consistently used sensemaking support tools.

7.1.2 Reuse Attention Signals

User's browsing activity is rich with signals about what's important and meaningful to users. These signals are mined and utilized by other individuals to gain insight into the most useful features of sites, as well as generalized user interests and profiles [83, 109]. But what if we start to leverage and mine these signals as a way to support an individual user? In the health and productivity domains, researchers have already started to take advantage of these signals in the "quantified self" movement [145], helping individuals gain insight into their daily habits and routines.

During browser-based sensemaking tasks, users provide explicit signals about their information needs through searches, and subtler signals of tab activity, page scroll position, and highlighting text on pages. Bento takes advantage of searches and page activity to help users construct a sensemaking workspace. Rather than having users manually construct the names of tasks and subtasks, as well as copy and paste which resources are beneficial, Bento leverages searches, page attention, and source "stars"

as a way to scaffold a sensemaking workspace. Meta leverages a user's open tabs as an signal about what sources they want to evaluate and leverage for their evaluation of potential options. While Bento and search session approaches [183] can help search engines and users logically group their activities together, these signals also have the opportunity to signal when a system should step in. For example, if a user has too many tabs open relevant to a particular task, there could be a prompt that would help them consolidate and close the information from those tasks. Other tools could recognize if there are any consistent entities or phrases an individual is researching and begin to "queue up" other relevant resources. These signals have significant potential to assist with users, however, more work has to be done around modeling and classifying these user input signals.

7.1.3 Automatic Sensemaking Tasks

Sensemaking naturally occurs over multiple sessions – ranging from days to even months. These are often interleaved with other tasks, including simple transactional activities and other sensemaking queries. As a result, as a user might do with their tasks in a physical space [222], users leave digital remnants of their tasks around as reminders to continue them or as references [66]. This behavior has spurred the development of research tools like SearchBar [164] and task management tools like Workona [2] as means to support this workspace organization. Bento took inspiration from these tools in the design and structure of its workspace. SearchBar similarly used searches as persistent pointers back to results, and Bento allows users to elevate these to larger tasks.

These sensemaking task workspaces, as seen in Bento, give users a place to logically group and store the resources and judgements. Similar to activity workspaces [96, 114, 222], these sensemaking workspaces can allow for quicker task resumption and lower burden when users are managing multiple tasks. For something as ephemeral and uncertain as a sensemaking task, users might not be willing to spend a significant amount of energy to establish and maintain these workspaces. However, by evaluating the type of information and the signals a user is providing (e.g. searches), we can start to create and maintain these workspaces automatically. Bento takes advantage of searches, but future tools could use data like product categories for a shopping task, or location for a travel task as a way to generate and managing these project groupings. Then, when users want to clean up or revisit a task, all of the resources they've explored, as well as potentially relevant new sources, are automatically organized for their use.

7.1.4 Triage as First Class

As users encounter information during the sensemaking process, they're continuously making value judgements about the potential utility of data. These judgements are often quick, and depending on the phase of the process they are in (early versus late) ephemeral and not concrete. Badi et al [18] and Hearst et al [95] describe this process as triage, or "the practice of quickly determining the usefulness and relevance of documents in a collection of documents." As users work with these quick, ephemeral judgements, their information storage and manipulation needs are vastly different from longer term

storage needs. Interaction costs for making these value judgements need to be low to match the quick and constant nature of the triage process. Additionally, users need the capability to not only quickly add, but also delete, de-prioritize, and organize information as its value can change quickly during the process.

Tools, such as Hearst et al.'s visual organizer [95] and VKB [152] recognize and implement this when users are making value judgements about documents or full web pages. Bento provides a similar benefit by allowing users to mark certain resources as important with a "star" or indicating some resources need to be revisited with a high information potential through a "to read". Extending beyond documents, Distil (using the Siphon toolkit) allows users to select different regions of pages to save, as a user might not know early on in the process what boundary of information is important. Distil then allows users to reference and organize these loose judgements through its "smart categories" – creating persistent, quick categorizations with keywords. Meta takes this a step further by allowing users to triage individual entities mentioned by pages, where they can pull out a subset of entities that they find particularly promising.

In each of these systems, users are given the ability to apply their judgements with simple, one-click interactions. Allowing users to record their judgements in a quick and natural way is essential for not only helping users keep track of their own thoughts during the process, but also if they are collaborating with or plan to pass off the task to another individual. In this work, we've found that giving users the power to attach judgements (such as stars, notes and tags) to information primatives (such as sources, clips, images, or entities) can provide a familiar and natural way externalize and capture these thought processes.

7.1.5 Two-way Information Flows

As users search and encounter information, they are simultaneously applying their existing knowledge to support their searching and organization processes, as well as learning from the new data to form a new structure they can leverage for further research. Numerous tools have taken advantage of both of these process models as a means to support the sensemaking process. For example, Scatter/Gather [51] presents information in a hierarchy that users can delve into, starting from the most general topics. Alternatively, Apolo [45] leverages examples as a bottom-up way to expand the current set of papers a user is working with.

Recognizing and supporting both this "bottom-up" learning simultaneously with the "top-down" application of existing knowledge in a single tool is a significant challenge. Bento took advantage of the "top-down" search structure a user has as a means of organization. The search queries acted as top-down guides and pointers back to the original information, allowing users to quickly refind information without having a ton of tabs open. While that was useful for refinding, it was limited in supporting any organization based on new data the user has learned during their searches.

Distil embodies this organizational schema in full by using featuring persistent, lightweight keyword based categories. Users can use these categories to both apply known structures, and explore potentially other useful segmentations, with low consequence if they have to change or update a category. By supporting this sweet spot of bottom-up and top-down application of knowledge, Distil can allows users to leverage the tool in a

way that matches their existing mental process. Future interfaces could strongly benefit from supporting both directions of learning in a single interaction or interface – letting users easily uncover interesting signals from the data while giving them the power to apply their personal perspective.

7.1.6 Information Compression

Lastly, even if users are able to collect information seamlessly and apply appropriate structures and segments to their tasks, individuals are still dealing with an overwhelming amount of information. Helping users explore and make conclusions with this data in a more efficient manner is essential. Researchers have made significant progress using visualizations to help users explore more quantitative data [195, 233]. Tools such as DataSquid [194], through focus and context interactions, offer users the ability to explore and find patterns in large sets of data, and then dive in to explore the variety of certain subgroups. While more and more information on the web is starting to be available as structured semantic meta-data [116], a large portion of the web is still mostly unstructured text content. Tools such as multi-document summarization offers some potential help on this front [87, 223, 236], but are still largely designed for scenarios like news stories which have a significant amount of content in a single document, rather than reviews which can often be just a few sentences.

Distil attempts to tackle part of this issue by using keywords as a way to drive organization and quickly pull in potentially relevant pieces of information. Rather than showing the full portions of the matching note, Distil instead surfaces a snippet similar to a search engine as a way to compact and reduce the irrelevant data for that particular category. Meta leverages entities as a means to support this information compression. Recognizing that individuals make decisions in the domains of shopping and travel are often deciding between well defined entities in online databases, Meta leverages this as a way to collect and collate the mentions across several pages. This significantly reduces the individual user burden of manually tallying these entities, and allows the information to be presented in several alternative formats, such as a table and a filterable list. This suggests that future systems could leverage either keyword or entity-based extraction of data as a reasonable next step to bootstrapping websites missing more robust semantic markup.

7.2 Future Work

Through this work, I was able to develop and test several strategies for reducing the burden on users during sensemaking. While this work was able to test the concepts behind these strategies in mostly a controlled lab environment, it doesn't yet look at how these concepts could be aggregated and tested in a real world sensemaking scenarios. Going forward, these strategies serve as a foundation for driving the design and development of more complete and complex tools.

7.2.1 Deployment

A deployed, real-world system that takes advantage of the ideas from Bento, Distil and Meta would further bolster the claims from this thesis, and would allow end users to take advantage of these improvements. We made a brief attempt at deploying just the ideas from Bento and Distil, however we found that clipping / highlighting and organizing the resulting data was still too much overhead for most users during sensemaking. Individuals appreciated the visual nature of the tool we created, and used it for a lot of longer term bookmarking, however still felt that it was too heavy, and involved for most scenarios, which was problematic for our goal of having a lightweight tool users would use for most sensemaking scenarios. As a result, we began to consider other ways that we might have an even less costly intervention, and developed Meta that relied on existing on page entities as a way to bootstrap getting information into the system. This removed the reliance on clipping / highlighting, while still creating a functional, useful tool. Moving forward with developing a deployed, end-user tool, we have to intelligently consider how we can either further persuade users to save data into the system with some sort of analysis benefit, or we can leverage the signals from information a user has explored as insight into their current thought process.

7.2.2 Collaboration and Reuse

Sensemaking is often not a solitary process — there are other stakeholders such as spouses, friends or even children who need to be consulted, especially during the decision making portion of the process. The judgements of others can also serve as the foundation for decisions, and with our increasingly connected society, it's become relatively easy to solicit and collect opinions. For most modern productivity tools, the ability to share and collaborate is a necessary first class feature — thus users expect tools to support and feature seamless collaboration.

Tools such as SearchTogether [166], CoSense [176], Cogamento [200] and IdeaMache [142] have all taken excellent first steps as outlining the necessary features users need to coordinate and share information during collaborative sensemaking tasks. Handoff-centric approaches, such as Clipper [123], have pointed to the importance of structure and schema when sharing sensemaking outcomes. However, these tools like many others require that very heavy up front commitment from users in order to take advantage of their benefits. I see this as an opportunity to leverage some of the sensemaking support strategies outlined in this section and apply them to collaborative or hand-off-centric scenarios.

7.2.3 Maximizers Abound

During our development of these tools, it became clear that individuals had some very strict constraints around the amount of time and effort they would put into sensemaking. These constraints and the resulting strategies used varied widely between users, and within users for certain types of tasks, as the amount of time and effort is highly correlated with how much a user cares about the outcomes of the task. A user taking a maximizer-oriented approach is more likely to rely on external cognitive aids, such as

tables or a document of notes, while one taking a satisficing approach is more likely to feel comfortable with existing suggestions, such as the top product on the Wirecutter or on Amazon.

However, what if tools made it so easy to be a maximizer, that the reasons for satisficing disappear? Individuals satisfice as a way to save time and effort, because being a maximizer in this scenario is just not beneficial enough to them. Smarter ways to aggregate and compress the data and options from multiple sources could obviate part of this problem – Meta explores this and makes it trivial to compare a large number of "listical" sites. Tools that help prevent tab overload, like Bento, could help reduce the stress individuals feel while approaching these problems, again encouraging more "maximization" behavior. Instead of feeling overwhelmed by options and data, something like Distil with its lightweight categorization, could help to keep things organized and tidy. In this thesis work, I hoped to create tools that empowered individuals to do more in depth research, and not feel limited by their ability to keep track of, categorize, and consume data from multiple different sources. Going forward, systems that inspire immediate delight when users start to interact with them, and can provide an immediate payout without a significant amount of data entry or overhead could help to encourage individuals to work with more data, and be more willing to exhibit maximizer behavior.

7.3 Conclusion

In this thesis, I explored how we might better support users during their digital sensemaking tasks with lightweight, in-situ tools. These three tools, Bento, Distil, and Meta, provided a tight coupling to users' existing browsing experiences by leveraging existing information such as tabs and searches, while also introducing cognitive scaffolds that could be easily generated and modified as a user's collected information changed. By using these tools, users were able to feel more organized, more readily generate organizations, and evaluate potential options more discriminatively. These tools allowed users to externalize their sensemaking models in four key stages of the sensemaking process: seeking, triage, structuring, and evaluation, and more effectively juggle the large amount of information required to perform sensemaking

Bibliography

[1] 2018. IAB / PwC Affiliate Marketing Study 2017. *IAB / PwC Affiliate Marketing Study 2017* (May 2018). https://www.iabuk.com/adspend/iab-pwc-affiliate-marketing-study-2017

[2] 2020. A better way to work in the browser. (2020). https://workona.com/

[3] 2020. Enable Rich Results with Structured Data. (Mar 2020). https://developers.google.com/search/docs/guides/search-gallery

[4] 2020. eufy Anker, BoostIQ RoboVac 11S (Slim). (2020). https://www.amazon.com/dp/B079QYYGF1

[5] Eytan Adar, Mira Dontcheva, James Fogarty, and Daniel S Weld. 2008. Zoetrope: interacting with the ephemeral web. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*. ACM, 239–248.

[6] Annette Adler, Anuj Gujar, Beverly L Harrison, Kenton O'hara, and Abigail Sellen. 1998. A diary study of work-related reading: design implications for digital reading devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 241–248.

[7] Niv Ahituv, Magid Igbaria, and A Viem Sella. 1998. The effects of time pressure and completeness of information on decision making. *Journal of management information systems* 15, 2 (1998), 153–172.

[8] Salman Ahmad, Alexis Battle, Zahan Malkani, and Sepander Kamvar. 2011. The jabberwocky programming environment for structured social computing. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 53–64.

[9] Anat Toder Alon. 2005. *Rediscovering word-of-mouth: An analysis of word-of-mouth talk in the context of online communities*. Boston University.

[10] Erik M Altmann and J Gregory Trafton. 2002. Memory for goals: An activation-based model. *Cognitive science* 26, 1 (2002), 39–83.

[11] Brian Amento, Loren Terveen, Will Hill, and Deborah Hix. 2000. TopicShop: enhanced support for evaluating and organizing collections of Web sites. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*. Citeseer, 201–209.

[12] Corin R Anderson and Eric Horvitz. 2002. Web montage: A dynamic personalized start page. In *Proceedings of the 11th international conference on World Wide Web*. ACM, 704–712.

[13] J Anderson. 2009. Forrester Market Report: Consumer Behavior Online: A 2009 Deep Dive. http://www.forrester.com/go?docid=54327. (2009). Accessed: 2017-09-10.

[14] John R Anderson and Robert Milson. 1989. Human memory: An adaptive perspective. *Psychological Review* 96, 4 (1989), 703.

[15] Salvatore Andolina, Khalil Klouche, Jaakko Peltonen, Mohammad Hoque, Tuukka Ruotsalo, Diogo Cabral, Arto Klami, Dorota Głowacka, Patrik Floréen, and Giulio Jacucci. 2015. Intentstreams: smart parallel search streams for branching exploratory search. In *Proceedings of the 20th international conference on intelligent user interfaces*. ACM, 300–305.

[16] Paul André, Aniket Kittur, and Steven P Dow. 2014a. Crowd synthesis: Extracting categories and clusters from complex data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 989–998.

[17] Paul André, Robert E Kraut, and Aniket Kittur. 2014b. Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 139–148.

[18] Rajiv Badi, Soonil Bae, J Michael Moore, Konstantinos Meintanis, Anna Zacchi, Haowei Hsieh, Frank Shipman, and Catherine C Marshall. 2006. Recognizing user interest and document value from reading and organizing activities in document triage. In *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, 218–225.

[19] Michelle Q Wang Baldonado and Terry Winograd. 1997. SenseMaker: an information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. ACM, 11–18.

[20] Nikola Banovic, Christina Brant, Jennifer Mankoff, and Anind Dey. 2014. ProactiveTasks: the short of mobile device use sessions. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. ACM, 243–252.

[21] Regina Barzilay, Kathleen R McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 550–557.

[22] Tim Berners-Lee and Mark Fischetti. 2001. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. DIANE Publishing Company.

[23] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010a. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 313–322.

[24] Michael S Bernstein, Desney Tan, Greg Smith, Mary Czerwinski, and Eric Horvitz.

2010b. Personalization via friendsourcing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 2 (2010), 6.

[25] Michael S Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 237–246.

[26] James R Bettman, Eric J Johnson, Mary F Luce, and John W Payne. 1993. Correlation, conflict, and choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19, 4 (1993), 931.

[27] Eric A Bier, Edward W Ishak, and Ed Chi. 2006. Entity quick click: rapid text copying based on automatic entity extraction. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*. ACM, 562–567.

[28] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 333–342.

[29] Christian Bizer and Radoslaw Oldakowski. 2004. Using context-and content-based trust policies on the semantic web. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. 228–229.

[30] Horatiu Bota, Ke Zhou, and Joemon M Jose. 2016. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, 131–140.

[31] Eliza Brooke. 2018. The rise of the recommendation site. *Vox* (Dec 2018). `https://www.vox.com/the-goods/2018/12/11/18131224/recommendations-best-strategist-wirecutter-buzzfeed-reviews`

[32] Robert Capra. 2011. HCI Browser: A tool for administration and data collection for studies of web search behaviors. In *International Conference of Design, User Experience, and Usability*. Springer, 259–268.

[33] Robert Capra, Gary Marchionini, Javier Velasco-Martin, and Katrina Muller. 2010. Tools-at-hand and learning in multi-session, collaborative search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 951–960.

[34] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 335–336.

[35] Stuart K Card, George G Robertson, and William York. 1996. The WebBook and the Web Forager: an information workspace for the World-Wide Web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 111–ff.

[36] Luis V Casalo, Carlos Flavián, and Miguel Guinalíu. 2007. The influence of satisfaction, perceived reputation and trust on a consumer's commitment to a

website. *Journal of Marketing Communications* 13, 1 (2007), 1–17.

[37] James Caverlee and Ling Liu. 2007. Countering web spam with credibility-based link analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*. 157–166.

[38] Pew Research Center. 2015. Generational differences in online activities. Report. (25 July 2015). `http://www.pewinternet.org/2009/01/28/generational-differences-in-online-activities/`.

[39] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (2013), 123–133.

[40] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 3 (2011), 27.

[41] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2016a. Supporting Mobile Sensemaking Through Intentionally Uncertain Highlighting. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 61–68.

[42] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2016b. Supporting Mobile Sensemaking Through Intentionally Uncertain Highlighting. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 61–68. `DOI:http://dx.doi.org/10.1145/2984511.2984538`

[43] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. 2019. SearchLens: Composing and Capturing Complex User Interests for Exploratory Search. In *Proceedings of the 2019 ACM 24th Annual Meeting of the Intelligent User Interfaces (IUI'19)*. 498–509.

[44] Joseph Chee Chang, Aniket Kittur, and Nathan Hahn. 2016. Alloy: Clustering with crowds and computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3180–3191.

[45] Duen Horng Chau, Aniket Kittur, Jason I. Hong, and Christos Faloutsos. 2011. Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 167–176. `DOI:http://dx.doi.org/10.1145/1978942.1978967`

[46] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. 2015. Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 4061–4064.

[47] Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43, 3 (2006), 345–354.

[48] EH-H Chi, Phillip Barry, John Riedl, and Joseph Konstan. 1997. A spreadsheet approach to information visualization. In *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*. IEEE, 17–24.

[49] Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1999–2008.

[50] Shelia R Cotten and Sipi S Gupta. 2004. Characteristics of online and offline health information seekers and factors that discriminate between them. *Social science & medicine* 59, 9 (2004), 1795–1806.

[51] Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. 1992. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 318–329.

[52] Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. 2017. Scatter/gather: A cluster-based approach to browsing large document collections. In *ACM SIGIR Forum*, Vol. 51. ACM, 148–159.

[53] Richard L Daft and Karl E Weick. 1984. Toward a model of organizations as interpretation systems. *Academy of management review* 9, 2 (1984), 284–295.

[54] Hoa Trang Dang, Diane Kelly, and Jimmy J Lin. 2007. Overview of the TREC 2007 Question Answering Track.. In *TREC*, Vol. 7. 63.

[55] Scott Davidoff, Min Kyung Lee, Anind K Dey, and John Zimmerman. 2007. Rapidly exploring application design through speed dating. In *International Conference on Ubiquitous Computing*. Springer, 429–446.

[56] Brenda Dervin. 1983. *An overview of sense-making research: Concepts, methods, and results to date*. The Author.

[57] Brenda Dervin. 1992. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. *Qualitative research in information management* 9 (1992), 61–84.

[58] Brenda Dervin. 1998. Sense-making theory and practice: an overview of user interests in knowledge seeking and use. *Journal of knowledge management* 2, 2 (1998), 36–46.

[59] Ravi Dhar. 1997. Consumer preference for a no-choice option. *Journal of consumer research* 24, 2 (1997), 215–231.

[60] Stefan Dietze. 2016. Retrieval, crawling and fusion of entity-centric data on the web. In *Semanitic Keyword-based Search on Structured Data Sources*. Springer, 3–16.

[61] Jerry Dischler. 2015. Building for the next moment. `http://adwords.blogspot.com/2015/05/building-for-next-moment.html`. (2015).

[62] Mira Dontcheva, Steven M Drucker, Geraldine Wade, David Salesin, and Michael F Cohen. 2006a. Collecting and organizing web content. In *Personal Information Management-Special Interest Group for Information Retrieval Workshop*. 44–47.

[63] Mira Dontcheva, Steven M Drucker, Geraldine Wade, David Salesin, and Michael F Cohen. 2006b. Summarizing personal web browsing sessions. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*. ACM, 115–124.

[64] Grahame R Dowling and Richard Staelin. 1994. A model of perceived risk and intended risk-handling activity. *Journal of consumer research* 21, 1 (1994), 119–134.

[65] Laura Drăgan, Siegfried Handschuh, and Stefan Decker. 2011. The Semantic Desktop at Work: Interlinking Notes. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11)*. ACM, New York, NY, USA, 17–24. DOI:http://dx.doi.org/10.1145/2063518.2063521

[66] Patrick Dubroy and Ravin Balakrishnan. 2010. A study of tabbed browsing among mozilla firefox users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 673–682.

[67] Elizabeth Dwoskin and Craig Timberg. 2018. How merchants use Facebook to flood Amazon with fake reviews. https://www.washingtonpost.com/business/economy/how-merchants-secretly-use-facebook-to-flood-amazon-with-fake-reviews/2018/04/23/5dad1e30-4392-11e8-8569-26fda6b404c7_story.html, *The Washington Post* (Apr 2018).

[68] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting Disputed Claims on the Web. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. Association for Computing Machinery, New York, NY, USA, 341–350. DOI:http://dx.doi.org/10.1145/1772690.1772726

[69] Günes Erkan and Dragomir R Radev. 2004. LexRank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* (2004), 457–479.

[70] Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 247–256.

[71] Kristopher Floyd, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling. 2014. How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing* 90, 2 (2014), 217–232.

[72] BJ Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, and others. 2001. What makes Web sites credible?: a report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 61–68.

[73] Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December (2002), 2.

[74] Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*. 1–15.

[75] William A Gale, Kenneth W Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 233–237.

[76] Michael Gamon, Sumit Basu, Dmitriy Belenko, Danyel Fisher, Matthew Hurst, and Arnd Christian König. 2008. BLEWS: Using blogs to provide context for news articles.. In *ICWSM*. 60–67.

[77] Yolanda Gil and Donovan Artz. 2007. Towards content trust of web resources. *Journal of Web Semantics* 5, 4 (2007), 227–239.

[78] Dennis A Gioia and Kumar Chittipeddi. 1991. Sensemaking and sensegiving in strategic change initiation. *Strategic management journal* 12, 6 (1991), 433–448.

[79] Jennifer Golbeck. 2008. Weaving a web of trust. *Science* 321, 5896 (2008), 1640–1641.

[80] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4*. Association for Computational Linguistics, 40–48.

[81] Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review* 29 (2018), 21–43.

[82] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 267–274.

[83] Qi Guo and Eugene Agichtein. 2010. Ready to buy or just browsing? Detecting web searcher goals from interaction data. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 130–137.

[84] Vishal Gupta and Gurpreet Singh Lehal. 2010. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence* 2, 3 (2010), 258–268.

[85] Kelvin Guu, John Miller, and Percy Liang. 2015. Traversing Knowledge Graphs in Vector Space. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 318–327.

[86] Zoltan Gyongyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with trustrank. In *Proceedings of the 30th international conference on very large data bases (VLDB)*.

[87] Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 362–370.

[88] Nathan Hahn, Joseph Chee Chang, and Aniket Kittur. 2018. Bento browser: complex mobile search without tabs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[89] Udo Hahn and Ulrich Reimer. 1999. Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. *Advances in Automatic*

Text Summarization (1999), 215–232.

[90] Jaehyun Han and Geehyuk Lee. 2015. Push-push: A drag-like operation overlapped with a page transition operation on touch interfaces. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 313–322.

[91] Sudheendra Hangal, Abhinay Nagpal, and Monica Lam. 2012. Effective browsing and serendipitous discovery with an experience-infused browser. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. 149–158.

[92] Eszter Hargittai, Lindsay Fullerton, Ericka Menchen-Trevino, and Kristin Yates Thomas. 2010. Trust online: Young adults' evaluation of web content. *International journal of communication* 4 (2010), 27.

[93] Marti Hearst. 2006a. Design recommendations for hierarchical faceted search interfaces. In *ACM SIGIR workshop on faceted search*. Seattle, WA, 1–5.

[94] Marti A. Hearst. 2006b. Clustering Versus Faceted Categories for Information Exploration. *Commun. ACM* 49, 4 (April 2006), 59–61. `DOI:http://dx.doi.org/10.1145/1121949.1121983`

[95] Marti A Hearst and Duane Degler. 2013. Sewing the seams of sensemaking: A practical interface for tagging and organizing saved search results. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*. ACM, 4.

[96] D Austin Henderson Jr and Stuart Card. 1986. Rooms: the use of multiple virtual workspaces to reduce space contention in a window-based graphical user interface. *ACM Transactions on Graphics (TOG)* 5, 3 (1986), 211–243.

[97] Thorsten Hennig-Thurau, Kevin P Gwinner, Gianfranco Walsh, and Dwayne D Gremler. 2004. Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? *Journal of interactive marketing* 18, 1 (2004), 38–52.

[98] Ken Hinckley, Shengdong Zhao, Raman Sarin, Patrick Baudisch, Edward Cutrell, Michael Shilman, and Desney Tan. 2007. InkSeine. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI 07*. ACM Press. `DOI:http://dx.doi.org/10.1145/1240624.1240666`

[99] Stephen J Hoch and Young-Won Ha. 1986. Consumer learning: Advertising and the ambiguity of product experience. *Journal of consumer research* 13, 2 (1986), 221–233.

[100] Johann Hofmann. 2015–2019. xpath-dom. `https://github.com/johannhof/xpath-dom`. (2015–2019).

[101] Andrew Hogue and David Karger. 2005. Thresher: automating the unwrapping of semantic content from the World Wide Web. In *Proceedings of the 14th international conference on World Wide Web*. ACM, 86–95.

[102] Jeff Huang and Ryen W White. 2010. Parallel browsing behavior on the web. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*. ACM, 13–18.

[103] Zhicong Huang, Alexandra Olteanu, and Karl Aberer. 2013. CredibleWeb: a

platform for web credibility evaluation. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 1887–1892.

[104] David Huynh, Stefano Mazzocchi, and David Karger. 2005. Piggy bank: Experience the semantic web inside your web browser. In *International Semantic Web Conference*. Springer, 413–430.

[105] Shamsi T. Iqbal, Jaime Teevan, Dan Liebling, and Anne Loomis Thompson. 2018. Multitasking with Play Write, a Mobile Microproductivity Tool. In *Proceedings of the 31st annual ACM symposium on User interface software and technology*. ACM.

[106] Melody Y Ivory and Marti A Hearst. 2002. Statistical profiles of highly-rated web sites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 367–374.

[107] Jacob Jacoby, Donald E Speller, and Carol A Kohn. 1974. Brand choice behavior as a function of information load. *Journal of marketing research* 11, 1 (1974), 63–69.

[108] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. 1999. Data clustering: a review. *ACM computing surveys (CSUR)* 31, 3 (1999), 264–323.

[109] Bernard J Jansen. 2009. Understanding user-web interactions via web analytics. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–102.

[110] Jim Jones. 2013. Turkee Ruby Gem. `https://github.com/aantix/turkee`. (2013).

[111] David Jonker, William Wright, David Schroh, Pascale Proulx, Brian Cort, and others. 2005. Information triage with TRIST. In *2005 International Conference on Intelligence Analysis*. 2–4.

[112] Ece Kamar and Eric Horvitz. 2013. Light at the End of the Tunnel: A Monte Carlo Approach to Computing Value of Information. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '13)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 571–578. `http://dl.acm.org/citation.cfm?id=2484920.2485011`

[113] Eser Kandogan and Ben Shneiderman. 1997. Elastic Windows: evaluation of multi-window operations. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. ACM, 250–257.

[114] Victor Kaptelinin and Mary Czerwinski. 2007. *Beyond the desktop metaphor: designing integrated digital work environments*. Vol. 1. The MIT Press.

[115] David R Karger. 2014. The semantic web and end users: What's wrong and how to fix it. *IEEE Internet Computing* 18, 6 (2014), 64–70.

[116] David R Karger and Dennis Quan. 2004. Haystack: a user interface for creating, browsing, and organizing arbitrary semistructured information. In *CHI'04 extended abstracts on Human factors in computing systems*. 777–778.

[117] Melanie Kellar, Carolyn Watters, and Michael Shepherd. 2007. A field study characterizing Web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 999–1018.

[118] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 4017–4026.

[119] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.

[120] Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. 2012. CrowdWeaver: Visually Managing Complex Crowd Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1033–1036. DOI:http://dx.doi.org/10.1145/2145204.2145357

[121] Aniket Kittur and Robert E Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 37–46.

[122] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

[123] Aniket Kittur, Andrew M Peters, Abdigani Diriye, and Michael Bove. 2014. Standing on the schemas of giants: socially augmented information foraging. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 999–1010.

[124] Aniket Kittur, Andrew M Peters, Abdigani Diriye, Trupti Telang, and Michael R Bove. 2013. Costs and benefits of structured information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2989–2998.

[125] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.

[126] Aniket Kittur, Bongwon Suh, Bryan A Pendleton, and Ed H Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 453–462.

[127] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 2: A macrocognitive model. *Intelligent Systems, IEEE* 21, 5 (2006), 88–92.

[128] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data–frame theory of sensemaking. In *Expertise out of context*. Psychology Press, 118–160.

[129] Khalil Klouche, Tuukka Ruotsalo, Luana Micallef, Salvatore Andolina, and Giulio Jacucci. 2017. Visual re-ranking for multi-aspect information retrieval. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 57–66.

[130] Steffen Koch, Markus John, Michael Worner, Andreas Muller, and Thomas Ertl. 2014. VarifocalReader — In-Depth Visual Analysis of Large Text Documents. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (dec 2014), 1723–1732. DOI:http://dx.doi.org/10.1109/tvcg.2014.2346677

[131] Jürgen Koenemann and Nicholas J. Belkin. 1996. A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '96)*. Association for Computing Machinery, New York, NY, USA, 205–212. DOI: http://dx.doi.org/10.1145/238386.238487

[132] Weize Kong and James Allan. 2014. Extending Faceted Search to the General Web. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 839–848. DOI:http://dx.doi.org/10.1145/2661829.2661964

[133] Jonathan Koren, Yi Zhang, and Xue Liu. 2008. Personalized Interactive Faceted Search. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 477–486. DOI:http://dx.doi.org/10.1145/1367497.1367562

[134] Bill Kules and Robert Capra. 2008. Creating exploratory tasks for a faceted search interface. In *Second Workshop on Human-Computer Interaction (HCIR 2008)*.

[135] Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. 2009. What do exploratory searchers look at in a faceted search interface?. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 313–322.

[136] Anand P Kulkarni, Matthew Can, and Bjoern Hartmann. 2011. Turkomatic: automatic recursive task and workflow design for mechanical turk. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2053–2058.

[137] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 23–34.

[138] Walter S Lasecki, Christopher D Miller, and Jeffrey P Bigham. 2013a. Warping time for more effective real-time crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2033–2036.

[139] Walter S Lasecki, Christopher D Miller, Raja Kushalnagar, and Jeffrey P Bigham. 2013b. Legion scribe: real-time captioning by the non-experts. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. ACM, 22.

[140] Edith Law and Haoqi Zhang. 2011. Towards Large-Scale Collaborative Planning: Answering High-Level Search Queries Using Human Computation.. In *AAAI*.

[141] Thomas Lin, Patrick Pantel, Michael Gamon, Anitha Kannan, and Ariel Fuxman. 2012. Active objects: Actions for entity-centric search. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 589–598.

[142] Rhema Linder, Nic Lupfer, Andruid Kerne, Andrew M. Webb, Cameron Hill, Yin Qu, Kade Keith, Matthew Carrasco, and Elizabeth Kellogg. 2015. Beyond Slideware: How a Free-form Presentation Medium Stimulates Free-form Thinking in the Classroom. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition (C&C '15)*. ACM, New York, NY, USA, 285–294. DOI:http://dx.doi.org/10.1145/2757226.2757251

[143] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2010. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 57–66.

[144] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A Myers. 2019. Unakite: Scaffolding Developers' Decision-Making Using the Web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 67–80.

[145] Deborah Lupton. 2016. *The quantified self*. John Wiley & Sons.

[146] Kurt Luther, Casey Fiesler, and Amy Bruckman. 2013. Redistributing Leadership in Online Creative Collaboration. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1007–1022. DOI:http://dx.doi.org/10.1145/2441776.2441891

[147] Yoelle S Maarek, Michal Jacovi, Menachem Shtalhaim, Sigalit Ur, Dror Zernik, and Israel Z Ben-Shaul. 1997. WebCutter: a system for dynamic and tailorable site mapping. *Computer networks and ISDN systems* 29, 8-13 (1997), 1269–1279.

[148] Richard Mander, Gitta Salomon, and Yin Yin Wong. 1992. A "pile" metaphor for supporting casual organization of information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 627–634.

[149] Inderjeet Mani and Eric Bloedorn. 1997. Multi-document summarization by graph search and matching. *arXiv preprint cmp-lg/9712004* (1997).

[150] Gary Marchionini. 2006a. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[151] Gary Marchionini. 2006b. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[152] Catherine C Marshall and Frank M Shipman III. 1997. Spatial hypertext and the practice of information triage. In *Proceedings of the eighth ACM conference on Hypertext*. ACM, 124–133.

[153] Dina Mayzlin. 2006. Promotional chat on the Internet. *Marketing science* 25, 2 (2006), 155–163.

[154] Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *AAAI/IAAI*. 453–460.

[155] James McKinney. 2015. TfIdfSimilarity Ruby Gem. https://github.com/jpmckinney/tf-idf-similarity. (2015).

[156] D Harrison McKnight and Charles J Kacmar. 2007. Factors and effects of information credibility. In *Proceedings of the ninth international conference on*

Electronic commerce. 423–432.

[157] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*. ACM, 1–8.

[158] Xiaojun Meng, Shengdong Zhao, and Darren Edge. 2016. HyNote: Integrated Concept Mapping and Notetaking. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '16)*. Association for Computing Machinery, New York, NY, USA, 236–239. DOI:http://dx.doi.org/10.1145/2909132.2909277

[159] Rada Mihalcea and Andras Csomai. 2007. Wikify! Linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 233–242.

[160] Iris Miliaraki, Roi Blanco, and Mounia Lalmas. 2015. From selena gomez to marlon brando: Understanding explorative entity search. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 765–775.

[161] Frances J Milliken. 1990. Perceiving and interpreting environmental change: An examination of college administrators' interpretation of changing demographics. *Academy of management Journal* 33, 1 (1990), 42–63.

[162] Neema Moraveji, Daniel Russell, Jacob Bien, and David Mease. 2011. Measuring improvement in user search performance resulting from optimal search tips. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 355–364.

[163] Luc Moreau and others. 2010. The foundations for provenance on the web. *Foundations and Trends® in Web Science* 2, 2–3 (2010), 99–241.

[164] Dan Morris, Meredith Ringel Morris, and Gina Venolia. 2008. SearchBar: a search-centric web history for task resumption and information re-finding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1207–1216.

[165] Meredith Ringel Morris, AJ Bernheim Brush, and Brian R Meyers. 2007. Reading revisited: Evaluating the usability of digital display surfaces for active reading tasks. In *Horizontal Interactive Human-Computer Systems, 2007. TABLETOP'07. Second Annual IEEE International Workshop on*. IEEE, 79–86.

[166] Meredith Ringel Morris and Eric Horvitz. 2007. SearchTogether: an interface for collaborative web search. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*. ACM, 3–12.

[167] Meredith Ringel Morris, Jarrod Lombardo, and Daniel Wigdor. 2010. WeSearch: supporting collaborative search and sensemaking on a tabletop display. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 401–410.

[168] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.

[169] Michael Nebeling, Alexandra To, Anhong Guo, Adrian A de Freitas, Jaime Teevan, Steven P Dow, and Jeffrey P Bigham. 2016. WearWrite: Crowd-assisted writing from smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3834–3846.

[170] Gergana Y Nenkov, Maureen Morrin, Andrew Ward, Barry Schwartz, John Hulland, and others. 2008. A short form of the Maximization Scale: Factor structure, reliability and validity studies. *Judgment and Decision Making* 3 (2008), 371–388.

[171] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 188–197.

[172] Jason RC Nurse, Syed Sadiqur Rahman, Sadie Creese, Michael Goldsmith, and Koen Lamberts. 2011. Information quality and trustworthiness: A topical state-of-the-art review. (2011).

[173] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. 2013. Web credibility: Features exploration and credibility prediction. In *European conference on information retrieval*. Springer, 557–568.

[174] Antti Oulasvirta and Pertti Saariluoma. 2006. Surviving task interruptions: Investigating the implications of long-term working memory theory. *International Journal of Human-Computer Studies* 64, 10 (2006), 941–961.

[175] Aditya Parameswaran, Ming Han Teh, Hector Garcia-Molina, and Jennifer Widom. 2013. Datasift: An expressive and accurate crowd-powered search toolkit. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[176] Sharoda A Paul and Meredith Ringel Morris. 2009. CoSense: enhancing sensemaking for collaborative web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1771–1780.

[177] Sharoda A Paul and Madhu C Reddy. 2010. Understanding together: sensemaking in collaborative information seeking. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 321–330.

[178] John W Payne, James R Bettman, and Eric J Johnson. 1993. The use of multiple strategies in judgment and choice. (1993).

[179] Jaakko Peltonen, Kseniia Belorustceva, and Tuukka Ruotsalo. 2017. Topic-relevance map: Visualization for improving search result comprehension. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 611–622.

[180] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[181] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[182] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. 2–4.

[183] Karthik Raman, Paul N Bennett, and Kevyn Collins-Thompson. 2013. Toward whole-session relevance: exploring intrinsic diversity in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 463–472.

[184] Ramana Rao and Stuart K Card. 1994. The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 318–322.

[185] Brian T Ratchford, Myung-Soo Lee, and Debabrata Talukdar. 2003. The impact of the Internet on information search for automobiles. *Journal of Marketing research* 40, 2 (2003), 193–209.

[186] Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S Bernstein. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 75–85.

[187] Stephen Robertson, Hugo Zaragoza, and others. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[188] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets.. In *ICWSM*.

[189] Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. ACM, 13–19.

[190] Ted Roselius. 1971. Consumer rankings of risk reduction methods. *Journal of marketing* 35, 1 (1971), 56–61.

[191] Volker Roth and Thea Turner. 2009. Bezel swipe: conflict-free scrolling and multiple selection on mobile touch screen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1523–1526.

[192] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. ACM, New York, NY, USA, 269–276. DOI:http://dx.doi.org/10.1145/169059.169209

[193] IAN RUTHVEN and MOUNIA LALMAS. 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review* 18, 2 (2003), 95–145. DOI:http://dx.doi.org/10.1017/S0269888903000638

[194] Jeffrey M Rzeszotarski. 2017. Uncovering Nuances in Complex Data through Focus and Context Visualizations. (2017).

[195] Jeffrey M Rzeszotarski and Aniket Kittur. 2014. Kinetica: naturalistic multi-touch

data visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 897–906.

[196] MC Schraefel, Max Wilson, Alistair Russell, and Daniel A Smith. 2006. mSpace: improving information access to multimedia domains with multimodal exploratory search. *Commun. ACM* 49, 4 (2006), 47–49.

[197] Barry Schwartz. 2004. The tyranny of choice. *Scientific American* 290, 4 (2004), 70–75.

[198] Julia Schwarz and Meredith Morris. 2011. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1245–1254. DOI:http://dx.doi.org/10.1145/1978942.1979127

[199] Mirjam Seckler, Silvia Heinz, Seamus Forde, Alexandre N Tuch, and Klaus Opwis. 2015. Trust and distrust on the web: User experiences and website characteristics. *Computers in human behavior* 45 (2015), 39–50.

[200] Chirag Shah. 2010. Coagmento-a collaborative information seeking, synthesis and sense-making framework. *Integrated demo at CSCW* 2010 (2010).

[201] Frank M Shipman III, Haowei Hsieh, Preetam Maloor, and J Michael Moore. 2001. The visual knowledge builder: a second generation spatial hypertext. In *Proceedings of the 12th ACM conference on Hypertext and Hypermedia*. ACM, 113–122.

[202] Ben Shneiderman. 2000. Designing trust into online experiences. *Commun. ACM* 43, 12 (2000), 57–59.

[203] Steven M Shugan. 1980. The cost of thinking. *Journal of consumer Research* 7, 2 (1980), 99–111.

[204] Donnavieve Smith, Satya Menon, and K. Sivakumar. 2005. Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing* 19, 3 (2005), 15 – 37. DOI:http://dx.doi.org/https://doi.org/10.1002/dir.20041

[205] Donnavieve Nicole Smith. 2002. *Trust me, would i steer you wrong? The influence of peer recommendations within virtual communities.* University of Illinois at Chicago.

[206] Michael Spenke, Christian Beilken, and Thomas Berlage. 1996. FOCUS: the interactive table for product comparison and selection. In *Proceedings of the 9th annual ACM symposium on User interface software and technology*. 41–50.

[207] Joel H Steckel, Russell S Winer, Randolph E Bucklin, Benedict GC Dellaert, Xavier Drèze, Gerald Häubl, Sandy D Jap, John DC Little, Tom Meyvis, Alan L Montgomery, and others. 2005. Choice in interactive environments. *Marketing Letters* 16, 3-4 (2005), 309–320.

[208] Jeffrey Stylos and Brad A Myers. 2006. Mica: A web-search tool for finding api components and examples. In *Visual Languages and Human-Centric Computing (VL/HCC'06)*. IEEE, 195–202.

[209] Jeffrey Stylos, Brad A Myers, and Andrew Faulring. 2004. Citrine: providing intelligent copy-and-paste. In *Proceedings of the 17th annual ACM symposium on User interface software and technology*. ACM, 185–188.

[210] Craig S Tashman and W Keith Edwards. 2011a. Active reading and its discontents: the situations, problems and ideas of readers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2927–2936.

[211] Craig S Tashman and W Keith Edwards. 2011b. LiquidText: a flexible, multitouch environment to support active reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3285–3294.

[212] Linda Tauscher and Saul Greenberg. 1997. Revisitation patterns in world wide web navigation. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. ACM, 399–406.

[213] Jaime Teevan. 2016. The future of microwork. *XRDS: Crossroads, The ACM Magazine for Students* 23, 2 (2016), 26–29.

[214] Jaime Teevan, Susan T Dumais, and Zachary Gutt. 2008. Challenges for supporting faceted search in large, heterogeneous corpora like the web. *Proceedings of HCIR* 2008 (2008), 87.

[215] Jaime Teevan, Daniel J Liebling, and Walter S Lasecki. 2014. Selfsourcing personal tasks. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2527–2532.

[216] title 2015. Etherpad Lite. `https://github.com/ether/etherpad-lite`. (2015).

[217] title 2016. Notion - The All-in-one Workspace. `https://www.notion.so/`. (2016). Accessed: 2018-04-10.

[218] Daniel Tunkelang. 2009. Faceted search. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–80.

[219] Rajan Vaish, Keith Wyngarden, Jingshu Chen, Brandon Cheung, and Michael S Bernstein. 2014. Twitch crowdsourcing: crowd contributions in short bursts of time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3645–3654.

[220] Max G Van Kleek, Michael Bernstein, Katrina Panovich, Gregory G Vargas, David R Karger, and MC Schraefel. 2009. Note to self: examining personal information keeping in a lightweight note-taking tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1477–1480.

[221] Eva A Van Reijmersdal, Marieke L Fransen, Guda van Noort, Suzanna J Opree, Lisa Vandeberg, Sanne Reusch, Floor Van Lieshout, and Sophie C Boerman. 2016. Effects of disclosing sponsored content in blogs: How the use of resistance strategies mediates effects on persuasion. *American Behavioral Scientist* 60, 12 (2016), 1458–1474.

[222] Stephen Voida, Elizabeth D Mynatt, and W Keith Edwards. 2008. Re-framing the desktop interface around the activities of knowledge work. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*. ACM, 211–220.

[223] Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2016. A sentence compression based framework to query-focused multi-document summarization. *arXiv preprint arXiv:1606.07548* (2016).

[224] Qing Wang and Huiyou Chang. 2010. Multitasking bar: prototype and evaluation of introducing the task concept into a browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 103–112.

[225] Aleksander Wawer, Radoslaw Nielek, and Adam Wierzbicki. 2014. Predicting webpage credibility using linguistic features. In *Proceedings of the 23rd international conference on world wide web*. 1135–1140.

[226] Karl E Weick. 1964. Reduction of cognitive dissonance through task enhancement and effort expenditure. *The Journal of Abnormal and Social Psychology* 68, 5 (1964), 533.

[227] Karl E. Weick. 1995. *Sensemaking in organizations*. Vol. 3. Sage.

[228] Allen M Weiss, Nicholas H Lurie, and Deborah J MacInnis. 2008. Listening to strangers: whose responses are valuable, how valuable are they, and why? *Journal of marketing Research* 45, 4 (2008), 425–436.

[229] Patricia M West and Susan M Broniarczyk. 1998. Integrating multiple opinions: The role of aspiration level on consumer response to critic consensus. *Journal of Consumer Research* 25, 1 (1998), 38–51.

[230] Ryen W White, Mikhail Bilenko, and Silviu Cucerzan. 2007. Studying the use of popular destinations to enhance web search interaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 159–166.

[231] Ryen W White, Bill Kules, Steven M Drucker, and others. 2006. Supporting exploratory search, introduction, special issue, communications of the ACM. *Commun. ACM* 49, 4 (2006), 36–39.

[232] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1, 1 (2009), 1–98.

[233] Pak Chung Wong and R Daniel Bergeron. 1994. 30 years of multidimensional multivariate visualization. *Scientific Visualization* 2 (1994), 3–33.

[234] William Wright, David Schroh, Pascale Proulx, Alex Skaburskis, and Brian Cort. 2006. The Sandbox for analysis: concepts and methods. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 801–810.

[235] Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing credibility judgment of web search results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1235–1244.

[236] Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. *arXiv preprint arXiv:1706.06681* (2017).

[237] Dongwook Yoon, Nicholas Chen, and François Guimbretière. 2013. TextTearing: opening white space for digital ink annotation. In *Proceedings of the 26th annual*

ACM symposium on User interface software and technology. ACM, 107–112.

[238] Ran Yu, Ujwal Gadiraju, Besnik Fetahu, and Stefan Dietze. 2017. Fusem: Query-centric data fusion on structured web markup. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, 179–182.

[239] Ran Yu, Ujwal Gadiraju, Besnik Fetahu, Oliver Lehmberg, Dominique Ritze, and Stefan Dietze. 2018. KnowMore – Knowledge Base Augmentation with Structured Web Markup. *Semantic Web* (2018), 1–21.

[240] Ran Yu, Ujwal Gadiraju, Xiaofei Zhu, Besnik Fetahu, and Stefan Dietze. 2016. Towards entity summarisation on structured web markup. In *European Semantic Web Conference*. Springer, 69–73.

[241] Oren Zamir and Oren Etzioni. 1999. Grouper: a dynamic clustering interface to Web search results. *Computer Networks* 31, 11-16 (1999), 1361–1374.

[242] Amy X. Zhang, Michele Igo, Marc Facciotti, and David Karger. 2017. Using Student Annotated Hashtags and Emojis to Collect Nuanced Affective States. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale (L@S '17)*. ACM, New York, NY, USA, 319–322. DOI:http://dx.doi.org/10.1145/3051457.3054014

[243] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 217–226.

[244] Pengyi Zhang and Dagobert Soergel. 2014. Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking. *Journal of the Association for Information Science and Technology* 65, 9 (2014), 1733–1756.

[245] Yuxiang Zhu, David Modjeska, Daniel Wigdor, Shengdong Zhao, and others. 2002. Hunter gatherer: interaction support for the creation and management of within-web-page collections. In *Proceedings of the 11th international conference on World Wide Web*. ACM, 172–181.