7-12-2018

Information foraging through the analysis of semantic network topology

Phillip J. DiBona
Rowan University

INFORMATION FORAGING THROUGH THE ANALYSIS OF SEMANTIC NETWORK TOPOLOGY

by

Phillip J. DiBona

A Thesis

Submitted to the
Department of Computer Science
College of Science and Mathematics
In partial fulfillment of the requirement
For the degree of
Master of Science in Computer Science
at
Rowan University
May 30, 2018

Thesis Advisor: Shen-Shyang Ho, Ph.D.

Dedications

I dedicate this thesis to my loving and supportive family. My wife Patricia and daughter Catherine have been a continued source of encouragement, and have shown tremendous patience with me throughout all the long hours and nights. My parents, Phillip and Carmelita, have provided an academic foundation for which I am very grateful.

Acknowledgements

It is with sincere gratitude that I acknowledge the support and guidance of my Professor and Thesis Advisor, Dr. Shen-Shyang Ho. He has continually provided astute observations and suggestions regarding research directions and opportunities to create a more general-purpose framework that can be used for continued research. His advice to engage the larger academic community has significantly improved my research processes.

I would also like to thank my committee members, Dr. Anthony Breitzman and Dr. Bo Sun, who are more than generous with their time and advice. Their diverse work has helped me understand additional applications for this research.

Abstract

Phillip J. DiBona
INFORMATION FORAGING THROUGH THE ANALYSIS OF SEMANTIC NETWORK TOPOLOGY
2017-2018
Shen-Shyang Ho, Ph.D.
Master of Science in Computer Science

Information seekers are posed with multiple challenges in gathering an unbiased and comprehensive body of information. The costs of analyzing documents often drive searches toward a small subset of documents. Additionally, modern search tools may reinforce the confirmation bias of users by providing only those documents that closely match their search query. The end result is a decision or hypothesis that is ill-considered and substantiated by potentially biased information. Information seekers need an information foraging tool that can help them explore the document corpus to find relevant topics and text snippets, while finding the hidden information that may be buried in the corpus or may not have been known a priori. An automated information foraging tool can mitigate these challenges by automatically identifying a wide breadth of topics for the user, extracted directly from a document corpus. When documents are decomposed and reconstituted into a semantic network, there is value in the topological structures formed. Leveraging a suite of information retrieval and graph analysis algorithms that analyze the semantic network, a framework is defined for assisting information seekers in both exploring and exploiting relevant information from a corpus to support unbiased decision making.

Table of Contents

Table of Contents (Continued)

Table of Contents (Continued)

Table of Contents (Continued)

List of Figures

List of Tables

Chapter 1

Introduction

Information has become ubiquitous in the internet age, where both news outlets and individuals are continually publishing articles to the World Wide Web covering every conceivable topic and event from numerous perspectives. For information seekers, the large quantity of available information (i.e., volume) and rate at which new information is created and published (i.e., velocity)[1] provides the opportunity to research and analyze a wide array of information. This rich body of material informs decisions, arguments, and hypotheses ranging from business to government and military policy. The ideal goal of the information seeker, enabled by this volume of data, is to generate an understanding of a domain that is both complete and accurate, but this depends primarily on the accuracy and completeness of the available information[2]. However, this same volume and velocity of data may hinder that goal because completely finding and analyzing all documents within a domain's corpus is infeasible due to the time and manpower constraints required to do so. This leads to two related key challenges that diminish the overall completeness and accuracy of a full search and analysis of a domain.

1.1 Exploration-Exploitation Tradeoff

The first challenge is the Exploration-Exploitation Tradeoff, which contrasts the desirability to explore as much of the information space as possible against the inherent costs of analyzing each document that relates to a domain in order to find and extract the relevant snippets needed to inform analysis[3]. These costs are typically quantified as the time involved in reading and analyzing the documents. Because of the sheer volume of data available, the information seeker will not be able to extract a holistic set of topics and snippets that cover the breadth of the search domain. In practice, the exploration-exploitation tradeoff operates more as a

continuous spectrum where the seeker will transition between exploration of the domain and exploitation of documents in particular sub-topics of the domain based on in situ goals and an assessment of what is "good enough"[4]. For example, a person will often repeatedly transition between exploration and exploitation as the results of one search may guide later searches. Typically, the exploratory side of the spectrum strives to find relevant topics to better understand the information landscape as it relates to a domain, while exploitation strives to "reward" the seeker with documents or document snippets that can serve to provide details, evidence, and depth to the understanding of specific topics. When a seeker is unfamiliar with a domain or sub-domain, these transitions may themselves become costly and time consuming, resulting in a truncated search. The seeker ultimately must decide what level of exploration is "good enough" based on incomplete information.

1.2 Search Tools and Bias

The second challenge that affects search completeness and accuracy stems from the tools we most often use to discover information. Modern search engines offer information seekers efficient and ubiquitous access to this vast volume of data for both exploring and exploiting their corpus of information. These tools excel at matching documents against a user's query, where a seeker formulates a query using specific search criteria, and the search engine recommends a series of documents. This process is well-suited to support the exploitation side of the Exploration-Exploitation spectrum, as the user is presented with documents for further, detailed analysis. But for the seeker who is exploring the domain, these documents, or the tools presenting them, may not provide the context for understanding how they fit into the larger domain[5].

For those information seekers already familiar with a domain, these search tools may also pose an additional challenge from confirmation bias, the tendency to seek

confirming evidence by providing ranked results that closely match the search query[6]. This adversely affects completeness of the search because analyzed documents, topics, and snippets will be only a subset of the relevant information domain. This subset will be closely correlated to the specific search query terms, causing an inability to discover relevant topics unknown to the information seeker. When search engine induced bias is coupled with inherent bias in the information seeker, a resulting understanding of a domain may be skewed toward specific sub-domains, leading to a significantly incomplete and inaccurate analysis.

1.3 Information Foraging

Information search is ultimately used to develop an information product, that may take the form of a formal hypothesis or structured argument, research paper, or a mental understanding of a domain or topic, a process often referred to as *sensemaking*. Pirolli and Card defined a model for sensemaking[3] that is principally comprised of two intersecting loops, the foraging loop and the the sensemaking loop (see Figure 1). The foraging loop models how information analysts (i.e., seekers) gather information, collect it into an *analyst's shoebox*, and after identifying relationships among the data, form an evidence file. This process iteratively narrows information topics from an unstructured raw set, to a relevant set, to a lucrative set. By the time the information is organized into the evidence file, the information analyst has used the information and concept/topic relationships as well as contextual information to form loose or high-level organizational structures within the data, often in the form of concept maps (or semantic networks). The sensemaking loop builds upon this base of information and models how analysts schematize the information, adding more structure, and ultimately forming hypotheses.

As hypotheses are formed, seekers may often need to traverse down the

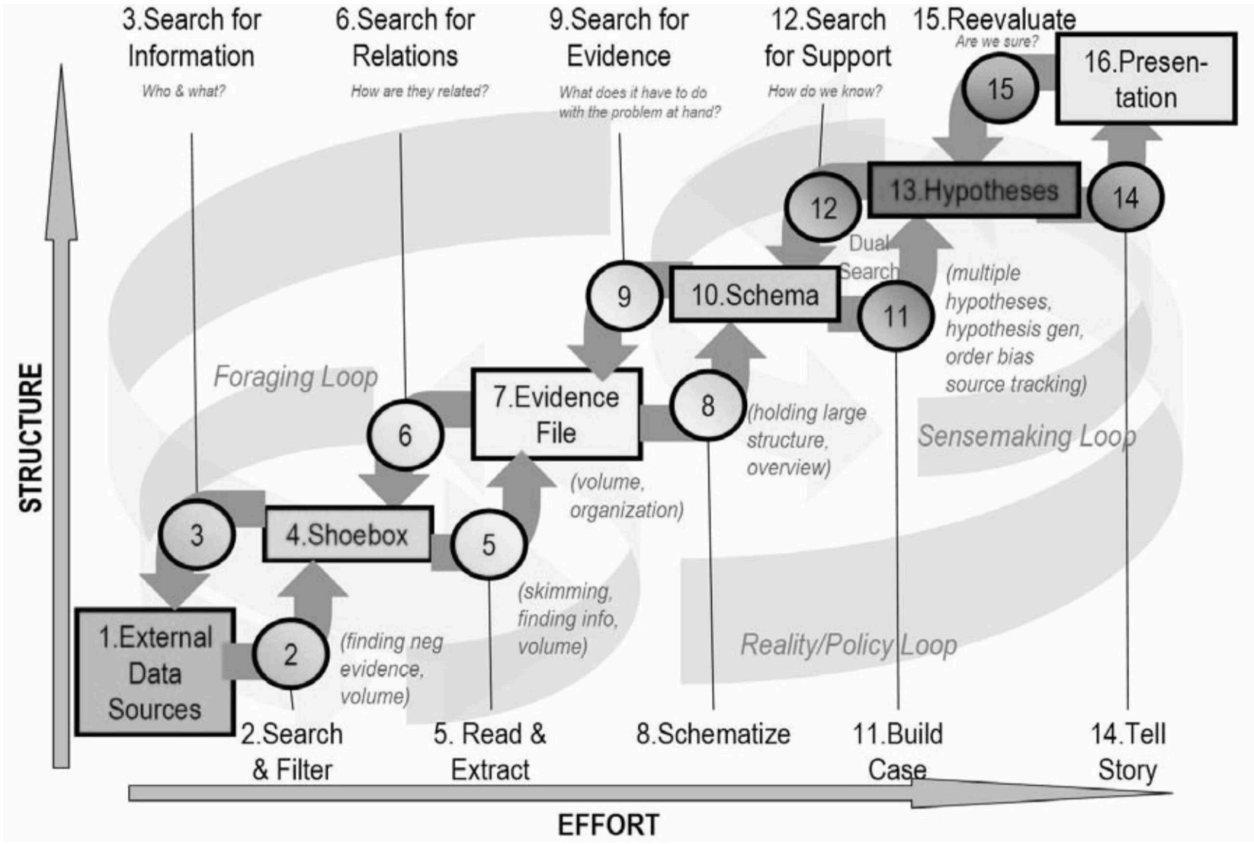

Figure 1. An information foraging and sensemaking model

sensemaking loop and back into the foraging loop to find information to further substantiate or refute the premises of the hypotheses (i.e., exploitation), potentially changing the schema and structure of that information. As premises and assumptions are substantiated or refuted, the information seeker may again transition into an exploration process, seeking to expand the breadth of topics analyzed. Depending on the information seeker's a priori understanding of the domain, the information foraging process may initiate anywhere on the exploration-exploitation spectrum. However, for a complete and accurate understanding of the information, the seeker will have made several transitions between the two ends of the spectrum, exploring a wide breadth of information and

selectively exploiting several sub-topics, gathering an evidence file. Information Foraging Theory also discusses that seekers will often identify *information patches*[7] that are rich sources of relevant information. These patches are visited by the seeker, and when the exploitation of information at that patch is exhausted (or the diminishing value of the information exploited no longer justifies remaining at the patch), the seeker will move onto another information patch for further foraging. The content of the patches are identified by the seeker via indicators or cues and may hint at their likelihood to be lucrative patches for exploitation. These cues, referred to as the *information scent*, help the seeker "follow the data" during their exploration and foraging.

1.4 Semantic Networks in Exploratory Search

Throughout the research into how information seekers discover, collect, organize, and schematize information, there is a recurrent theme regarding the benefits of modeling the relationships among elements of information. In their sensemaking model, Pirolli and Card assert that as information seekers forage for information and evidence, the seeker will identify relationships among the concepts within their corpus, often creating concept maps, or semantic networks, that model these concepts (e.g., people, locations, events, organizations) and their interrelationships. Collaborative databases, such as Wikipedia, often have explicit links between documents, concepts, and categories (i.e., topics). Exploratory search based on these network relationships can improve search performance and provide diversity in the search results[9]. Recent research approaches in exploratory search interfaces have shown efficacy from displaying and leveraging network views of the information space as means to navigate the complex relationships of data elements within the corpus, visualizing both the documents and the categories that label them[10]. These networks of information concepts and their interrelationships are in

fact semantic networks, a graph structure that represents knowledge in patterns of concept nodes and edges[11]. In practice, these networks may be modeled as either directed or undirected graphs, and the edges may be weighted to represent the strength of that relationship.

1.5 Premise

This information foraging model provides useful insights into the cognitive science of information exploration and exploitation. The constructs of the iterative foraging loop that transitions the exploration-exploitation spectrum, information patches and scent, along with the common (human) use of concept maps for identifying relationships within the information space offer insights into how the computer science community can develop algorithms and methods to facilitate lucrative information foraging. Both the exploration-exploitation tradeoff as well as cognitive bias (inherent in the seeker and exacerbated by search tools) pose as barriers to a complete and accurate exploration of a domain's information space.

Information seekers need tools to support the exploration and navigation of the wide volume of a domain's topics and concepts as well as their complex interrelationships. Decisions on how to explore this conceptual network offer the information seeker, not a few options to follow from each topic or concept, but possibly tens or hundreds of possibilities[5]. Information search tools should not support just exploration or exploitation, but rather the repeated transitions that occur throughout the search process.

Information seekers, when gathering data for important decisions and hypotheses, need an automated information foraging (AIF) capability that can analyze a document corpus for a specific domain, and provide a simultaneous presentation of the full breadth of topics that comprise that domain, where it will be apparent to the user what topics are consonant or dissonant with their extant

beliefs or biases[8]. Further, this capability should allow the seekers to narrow these topics for further analysis by providing access to related topic adjacencies (both in breadth and depth) across the exploration-exploitation spectrum. This allows them to follow the *information scent* to explore the various information patches existing in the corpus and enrich their hypotheses through the identification of snippets of information that can serve as supporting or refuting evidence for decisions and hypotheses.

The technical premise behind this information foraging framework is that when documents in a corpus are decomposed into constituent terms and formed into a semantic network, based on their co-occurrence and relationships to other terms, there is inherent value in the graph structures formed. These semantic networks organically form concept community structures that can be viewed by the information seeker as a set of *information patches* representing highly-cohesive topics within the corpus with their own interrelationships. These inter-relationships can assist the seeker in exploration of the information space through adjacent (i.e., related) patches/topics. By having an ability to explore the entire corpus via these topics, the information seeker may identify those topics for further exploitation to find the lucrative document snippets to be used as evidence for substantiating or refuting hypotheses or informing decisions. Additionally, a hierarchical graph-based schema provides full traceability between term-concept communities and the documents that originated the terms.

This thesis describes the methods of the AIF framework to:

- Decompose an information corpus,
- Construct a semantic network of concepts,
- Identify structures in the semantic network as cohesive concept communities,

- Present the concept communities to the information seeker as corpus topics,
- Explore the corpus topics to identify relevant topics to a seeker's queries.

1.6 Contributions to the Research

Pirolli and Card have proposed the need for an automated information foraging capability that can assist information analysts/seekers in exploring the raw information landscape to identify the relevant and lucrative text snippets that can be exploited as evidence for hypotheses and decisions. However, without addressing bias in the information and in the information seeker, this evidence will also be biased, potentially leading to ill-considered decisions. Heuer posits the need for rigorous methods to identify and eliminate bias in decision making through developing multiple competing lines of inquiry requiring analyzing as much of the information as possible. This thesis describes a framework for providing an information foraging capability that addresses both goals through a suite of algorithms that focus on the semantic network, a construct that is inspired by the concept maps often generated by information analysts. While other research has addressed various aspects of the foraging problem, this ensemble set of algorithms addresses the *full* information foraging loop, from identifying raw information sources, through assisting with the exploration of those documents to find relevant and lucrative text snippets. The resulting automated information foraging framework provides the software tools to study the entire information foraging process in light of the exploration-exploitation tradeoff, while also exposing the hidden lucrative information embedded in the corpus, mitigating potential bias.

1.7 Scope

The actual information landscape for any particular domain will contain a wide variety of data types and forms, such as text (structured and unstructured), imagery (e.g., photographs, figures, maps, timelines), video, and audio. Each of

these modalities offers an important element to an information foraging system. The multimedia types are specifically well suited as cues for modeling information scent.

For this thesis, the variety of information considered will be limited to unstructured text in the form of web-published articles from established news sources. Unstructured text is a rich and complex data type with unique challenges and opportunities. Additional research in expanding these concepts and methods to non-textual information would prove beneficial to the overall goals of this thesis.

Additionally, this thesis is limited to the investigation of algorithmic approaches for creating an automated information foraging capability through the analysis of semantic network topology. Graphical user interfaces nor visualizations of the information landscape have been attempted. The goal is to establish a proof of concept framework for these methods as a basis for research in automated information foraging.

The algorithms and approaches describes in this thesis require a corpus of unstructured textual documents. Although this thesis does identify one such method for acquiring a corpus, this is not a focus area under investigation.

Chapter 2

Literature Review

The purpose of this literature review is to identify the breadth of topics relevant to the construction of an automated information foraging (AIF) framework. Because AIF and its technical premise cover multiple disciplines and research areas within those disciplines, the scope of this review is intended to identify key research papers and how they relate to AIF. This review is divided into cognitive science, computer science, and the contributions to the research stemming from this thesis.

2.1 Cognitive Science

Information foraging, as a research topic, is first concerned with the processes and confounds involved in how humans search for information as well as the rationale and goals informing that search. Through cognitive task analysis, Peter Pirolli and Stuart Card empirically studied the processes and structures employed by information and intelligence analysts in sensemaking[3]. The roles studied are optimal subjects because their success requires analytic rigor, the ability to organize and sift through vast amounts of information, and minimize cognitive biases, as the hypotheses they form and substantiate are crucially important. Pirolli and Card developed a model for Sensemaking that shows the relationship between the the structure of information and the level of effort required to achieve that structure, where the pinnacle is a defensible hypothesis that has ample evidence to substantiate it. The hypotheses are formed from schemas that organize the information which are in turn built upon the relationships among the elements of information. The basis of this sensemaking model is the *foraging loop*, where the analysts search for information among external, raw data sources, identify and filter relevant pieces of information and store that information in a loosely organized manner. As the information is read and examined in more detail, more lucrative

information is extracted and organized into a more structured manner where relationships among the information is identified. One of the common methods used to structure this evidence is through visual *concept maps* or *linkage maps*. This insight into how professional information seekers and analysts organize and structure information inspired much of the technical approach for this thesis. Further, their cognitive task analyst highlights that the information foraging loop is a leverage point in the sensemaking process where technology may provide benefits to the analyst.

Pirolli and Card expanded these concepts in their book, *Information Foraging Theory*, where they further explain this information search process relative to the exploration-enrichment-exploitation tradeoff in the form of a biological metaphor[7]. This metaphor compares the tradeoff to the process modeled by predators hunting for food as they balance the tradeoff of calories expended searching for food (exploration) against the calories gained (exploitation) through several foraging strategies. One of the key strategies is the patching strategy, where predators will identify and spend extended periods of time in food patches. When the food gained in these patches is expended, they will explore and find a new patch. Information foragers often have their own information patches. These could be information sources or websites. They may also be clusters of inter-related documents that contain a rich (and lucrative) set of information. This concept of an information patch is another inspiration for the technical approach in this thesis, namely, using clusters of highly-related information pieces that serve as means to facilitate information exploration.

Another key paper that studies successful information and intelligence analysis processes was authored by Richards Heuer, which asserts the need to mitigate cognitive bias in the information used for sensemaking[2]. One method for

mitigating bias is the generation of alternative and competing hypotheses, which requires a wide breadth of information gathered and analyzed. To the extent feasible, the analyst should assess the full information space, filtering and identifying information that is both consonant and dissonant to his/her beliefs and the hypotheses themselves. White, et al, assert that the very tools available to modern information seekers and analysts, such as search engines, may exacerbate cognitive biases because they find information that closely matches the query terms provided by the information seeker[6]. These papers further motivated this thesis to provide an automated means to assist bias mitigation through the presentation of all the topics contained within the information space. By having access to the full breadth of information topics from a corpus, the analyst should be able to quickly assess the information space and identify those topic that are relevant to the various hypotheses and which are consonant or dissonant to their beliefs.

2.2 Information Retrieval

Information Retrieval (IR) is the subfield of computer science concerned with the efficient storage, search, and discovery of large amounts of information. Many of the research areas within IR are applicable to this thesis, the first of which is *exploratory search*, which addresses how people find information when the problem is poorly defined, when they are unfamiliar with the domain in question, or when multiple perspectives must be analyzed[5]. Modern tools often facilitate information browsing behavior by following hyperlinks in documents which can lead to serendipitous discovery of information. This process, while easy for information seekers and can help in refinement of search goals, is often inefficient and not appropriate for fact-finding or exploring a wide breadth of the information space[40,41]. However, the literature strongly favors development of approaches for goal and task-oriented exploration of the information over serendipity[42]. To

maintain a higher-level of efficiency and value in the exploration process, quantification and assessment of the *relevance* of information is needed for exploratory search tools. This research has motivated this thesis to identify methods for quantifying relevance of information topics presented to the user.

Another important research area within exploratory search is concerned with the presentation and navigation of the information space. Some researchers assert that visualization and user interfaces are as important, or perhaps more important than analytic strategies. Visualization helps information seekers understand where they are in the domain, and how to identify and navigate the various decision points to explore related information. These kinds of tools and interface should facilitate information understanding over just finding and ranking information[9,10,43]. Although this thesis is not focused on user interfaces, this research motivates the AIF goal of providing a simultaneous presentation of all topics to the information seeker, so that navigation of those topics will be more tractable, as opposed to following semantic network links as if they are hyperlinks.

Topic modeling is widely researched area in IR, where the intent is to generate a set of topics contained in a corpus, where a topic is a pattern of term co-occurrences. Most approaches leverage a probabilistic model for identifying the statistically significant patterns of word (term) use and their associations to documents that exhibit similar patterns[44,45]. The leading approach is Latent Dirichlet Allocation (LDA) in which documents are modeled as mixtures of topics that generate terms with probabilities of association to those topics[30]. Therefore, the output of an LDA process is a set of n topics (where n is chosen a priori) containing significant, high probability terms associated with that topic. Despite the popularity and successes of LDA, there has been little analysis on the factors that characterize LDA's performance[46]. Through experimentation, Tang, et al,

demonstrate several factors that affect LDA performance, including the number of documents and the length of the documents. Additionally when there are large numbers of topics, LDA may not converge well. This is confirmed in real-world use cases[47,48]. For information foraging, one challenge that LDA poses is its need for the number of topics as an input parameter. If supplied by the user, this can be an implicit bias source. If iteratively tested with varying values by software, it could affect the resolution of topics generated. This probabilistic approach may also not detect small topics that are statistically insignificant, but may be relevant to the information exploration. But the successes of LDA motive this thesis to use a form of topic modeling for finding the information patches in a corpus as a means to facilitate exploration.

The IR community has explored the use of semantic networks, also called associative networks, for information indexing and retrieval. These semantic networks model knowledge and their interrelationships, similar to the information and intelligence analysts' use of concept maps in the Pirolli and Card study. This subfield of IR, known as *Associative Retrieval*, leverages these relationships between knowledge elements (terms) and documents to identify related, adjacent knowledge that may be relevant to a search. One successful associative retrieval method uses a *spreading activation model* that uses the associative nature of a semantic network as a means for controlling search[38]. This approach activates one or more nodes in the semantic network and propagates an output signal to adjacent nodes and controls how that signal and subsequent nodal activations spread through the larger network[49]. Cresanti posits that spreading activation can be used to retrieve relevant information by identifying other information that is associated with what is already known to be relevant. Because of its common underlying data structure (i.e., semantic networks) for organizing information used for sensemaking, as well as

its ability to identify relevant information, this thesis is motivated to leverage associative retrieval concepts for relevance analysis.

2.3 Graph Analytics

Because this thesis focuses on the analysis of a semantic network, literature on the analysis of graphs is warranted. Two key areas of graph analytics were investigated: centrality and community detection. Graph Centrality assesses which nodes in a network have the most relative importance with respect to the other nodes in the graph (or subgraph). There are several approaches toward computing centrality.

Degree centrality measure the importance of a node by the cardinality of its neighbors[50]. This measure has two disadvantages in that many nodes will share a common degree score and that it only uses local information with respect to each node. The graph as a whole is not important, and thus minimizes its descriptiveness as a measure of importance relative to the graph. Closeness centrality measures the average distance from each node to all other nodes[51]. While this is an improvement over degree centrality in that it uses non-local information and considers the whole graph, this measure does not have much semantic meaning regarding relative importance of nodes. Betweenness Centrality is a widely used method for computing centrality in a graph. This method finds the shortest distance between every pair of nodes in the graph and counts the times every node is traversed in those paths. This method does indeed create a semantically meaningful concept of importance. Additionally, betweenness centrality can be applied to both weighted and unweighted graphs as well as both nodes and edges, making this a versatile metric. Several efficient approaches to betweenness centrality have been developed by Newman[52] and Brandes[34].

Graph community detection, also referred to as graph clustering, are

algorithms that identify topological structures in the graph, specifically, communities of highly connected nodes with looser connections between those communities. Three popular classes of algorithms were investigated to determine an optimal approach to community detection. These classes are: Graph Partitioning, Divisive Hierarchical Clustering, and Hierarchical Maximum Modularity Clustering.

Graph partitioning was originally motivated by electronic circuit board designs that require partitioning the circuits to minimize the physical interconnections between modules[31]. The graph's edges model the physical interconnections on the circuit board, and clusters model groups of components on that same board. This class of algorithms optimizes a function that determines the ratio of the number of edges between modules to the number within modules. This method, although efficient and widely used for various applications, requires a predefined number of clusters to be provided to the algorithm.

The class of Divisive Hierarchical Clustering algorithms approaches community detection and clustering by identifying candidate edges that can be iteratively removed until communities emerge. One of the first and more popular methods is the Girvan-Newman Algorithm[13]. Girvan-Newman uses the Betweenness Centrality (B_c) Metric of the edges to find clusters by removing the edge with the highest B_c score, then recomputing the B_c scores for the entire graph again. This is repeated (hierarchically) until all edges have been removed or a modularity-score stopping condition is achieved, resulting in a dendrogram.

Girvan and Newman first used the modularity metric for assessing the quality of the community structures in the graph, which measures the degree to which a cluster's membership has dense edge interconnections while edge connections between clusters is relatively low[13]. Modularity can also take into account not just the density of edges between nodes to identify communities, but

also incorporate the weights of those edges, thus enabling the graph clustering to identify strongly connected groups of nodes[32]. Although maximizing modularity is NP-Hard, Noack and Rotta identify several efficient greedy and heuristic-based approaches for achieving effective clusterings[33]. This class of algorithms uses a multi-level heuristic of coarsening, which iteratively merges cluster pairs starting from singleton clusters, then refinement, which iteratively reassigns vertices vertices to different clusters.

Chapter 3

Technical Approach Overview

This automated information foraging (AIF) framework is intended to facilitate research in assisting an information seeker in gaining an understanding of the breadth of information in a domain's document corpus, find relevant topics (that may or may not have been known a priori), and discover documents or document snippets that can be exploited for evidentiary substantiation or refutation of hypotheses or decisions. The underlying premise behind this AIF framework is that when documents in a corpus are decomposed into constituent terms and formed into a semantic network, there is inherent value in the topology of that network. The topology will contain graph communities (or clusters) based on the connectedness (and strength of the connections) among the concept nodes, revealing cohesive topics in the corpus. This topology additionally models semantic relationships between corpus concepts and between the topic clusters that assist in exploratory search of the information landscape. The semantic network, however, is only a subset of the overall AIF schema, which is a hierarchical graph that both facilitates breadth-wise exploration of the information domain of a corpus and depth-wise enrichment and exploitation of the documents and documents snippets associated with specific topics within the domain. In addition to facilitating the seeker's exploration-exploitation transitions, this hierarchical graph-based schema provides traceability between the topics, concepts, and the documents that spawned them.

As with any search process, this is not a completely automated process. The human information seeker must guide the process at a few key points to ensure that the automated process satisfies to the seeker's information needs. This is most evident in two areas, specifying the broad domain and in identifying exploration-exploitation transitions. This section describes the schema and the

high-level AIF framework processing (with details provided in subsequent chapters), noting points where human-in-the-loop inputs are required.

3.1 Hierarchical Graph-Based Schema

The overall schema for the AIF framework's data is an undirected graph comprised of four interconnected layers, where nodes in each layer are a homogeneous class of data, as shown in Figure 2. Each of the layers is a distinct subgraph with connections to the layers above and below (as applicable). The AIF framework constructs each layer from the bottom-up. The information seeker, however, will interact with the AIF data from the top-down, starting with exploratory search within the topic layer, then exploiting (i.e., analyzing) the snippets and documents by accessing data further down the graph.

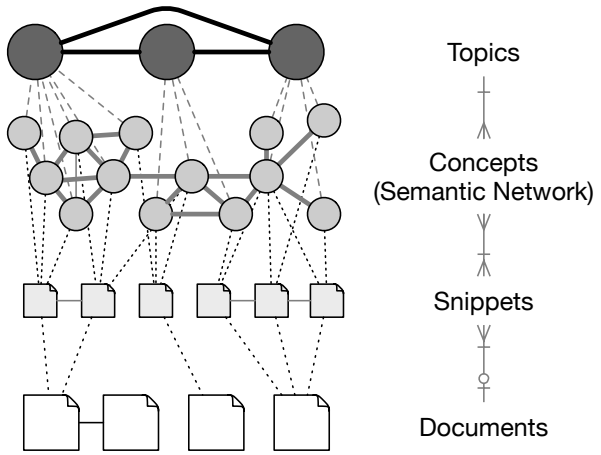

Figure 2. Automated information foraging hierarchical graph-based schema

The bottom layer is the *Documents Layer*, and contains one node per document in the corpus. In this context, a document may include a news article, a web page, blog post, or any other unstructured textual content. Metadata for each

document node contains the document's source (e.g., URL) and unstructured text content.

The next layer, the *Snippets Layer*, contains discrete sections of documents, such as article sections or paragraphs. Snippets are useful in this model because they are typically cohesive (and related) groups of sentences, usually more cohesive than the document as a whole. For each document, there are one or more snippets. From an information retrieval perspective, snippets are document surrogates that allow information seekers to quickly identify information based on utility or relevance[12]. For the AIF framework, snippets are used to limit the edges between nodes in the Concepts layer.

The third layer is the *Concepts Layer*, which is the semantic network of significant terms within the snippets, such as named persons, organizations, locations, as well as significant noun phrases. These terms represent concepts within the domain's corpus. Information analysts often form conceptual schemas of information by generating concept maps to organize information to infer patterns. Similarly, this layer forms a large semantic network (as an undirected weighted graph) of the significant terms in the corpus based on the co-occurrence relationships between terms found in the snippets. These concepts may originate from one or many snippets. The AIF framework will account for the common (and expected) situation when a concept is represented in multiple snippets.

The final (and top-most) layer is the *Topics Layer*, which is formed by identifying weighted community structures within the Concepts Layer's topology. Each node represents one concept cluster (or community), containing links to its constituent concepts in the layer below. The most significant concepts within each cluster are presented to the information seeker as representative terms (i.e., keywords) for that topic. In the current AIF prototype, a concept may only be a

member of one topic, however the framework allows for many-to-many relationships for future research.

3.2 Automated Information Foraging Process

The AIF framework supports the information seeker's exploration and exploitation of the information landscape, and therefore is comprised of tasks assigned to both the human and the automation support algorithms, as shown in Figure 3. The first phase of the automated information foraging process deconstructs a document corpus to its most basic atomic elements, textual terms, and reconstructs and reorganizes those terms to form the semantic network. The second phase of the process analyzes the topology of that semantic network using graph analytic techniques and leverages the structure of the network to facilitate both exploration and exploitation of the information.

The process initiates with the information seeker providing a document corpus to the AIF framework. This can be accomplished in one of two ways. The first is that the seeker can provide an externally-generated document corpus to the framework. The alternative is that the framework can use information retrieval tools and techniques to create the corpus for the user. The critical assumption of the AIF framework is that, regardless how the corpus was acquired or created, it should represent a wide breadth of documents that cover a domain. This corpus should not yet be restricted to subtopics or specific searches. Any search used to create this collection should be very broad and very high-level. For example, if the information seeker needs to explore the information landscape to understand Hurricane Irma, the search terms used to create this corpus should be "Hurricane Irma", not something more specific such as "Damages from Hurricane Irma". This will allow AIF to generate a wide range of topics for the seeker to explore, where one or more topics relating to *damages* would be represented.

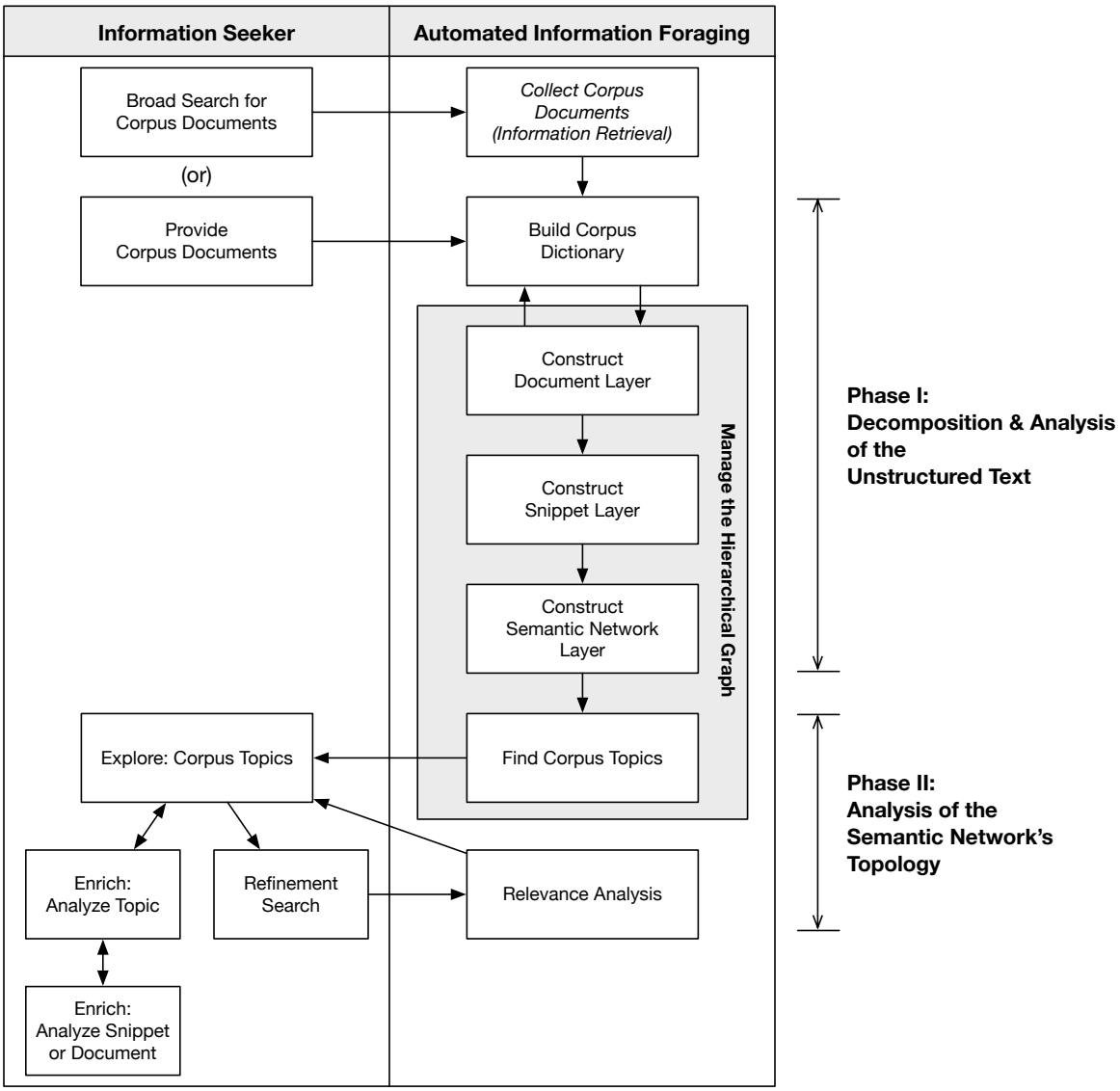

Figure 3. Automated information foraging high-level process

3.2.1 Phase I: text analysis. Once the corpus has been generated, the framework then creates a baseline dictionary by first extracting all of the terms that will eventually comprise the concepts in the semantic network. These terms can be unigrams or n-grams (from noun phrases or named entity recognition), and their corpus-wide frequency (i.e., the number of occurrence of that term) are determined

and stored. The terms are ordered and used as features for modeling the corpus in vector space. The set of possible terms, their ordering in vector space, and their frequencies are stored as a *Corpus Dictionary*.

To construct the Documents Layer, the AIF framework must first clean the data. Text documents, such as news articles and HTML pages, are often noisy, containing advertisements, author or source attributions, multimedia, or links to other documents. These superfluous textual elements often have little or no relevance to the main body of the text, and may skew the semantic network by adding irrelevant concepts or extraneous links. These elements are identified and removed from the text. In many cases, especially with news articles, similar or duplicate content may be published by separate sources. AIF identifies similar documents and removes duplicates when they are encountered. Without this step, link weights in the semantic network may become overemphasized. As documents or textual elements are filtered out, the Corpus Dictionary is updated accordingly, ensuring that the term frequencies are reflected accurately.

Snippets are constructed from the resulting documents in the previous step. Snippets are distinct subdivisions within the document, but the entire document content is represented by a set of snippets. Based on the specific runtime configuration of the AIF, snippets may be subdivided by paragraph boundaries, article sections (which contain one or more paragraphs), or as a whole document (one snippet per document with the entire body of text).

The final step of Phase I is the construction of the semantic network that models the concepts of the corpus and their interrelationships. For each snippet, each term that comprises it are identified from the Dictionary, and the terms are ranked by their salience in the document. Depending on the runtime configuration of the AIF framework, a certain percentage of the most salient terms are then

included as concepts in the semantic network. The assumption in the framework is that if two terms occur in the same snippet, they are related, and a semantic edge will connect them. The strength (weight) of those edges, however, are not fixed but rather are based on a weight function. Only those concepts that reside in a common snippet will have a link in the semantic network, regardless of whether they share a common document. This restriction is intended to limit the effects of large documents that may lack strong cohesion across the entire body of text. When a pair of concept nodes in the semantic network already exist, the weight of that relationship will be increased if that node-node link is discovered in a new snippet. By the conclusion of this step in the process, the semantic network (i.e., the Concepts Layer) will model the semantic relationships across the corpus. Structures in the topology of that network will have emerged organically as document snippets are added to the hierarchical graph.

3.2.2 Phase II: semantic network topology analysis. After the construction of the semantic network, Phase II of the AIF framework's processing will analyze the structure of that network to identify a complete set of topics across the corpus. The assumption in this step is that collections of highly related and cohesive graph clusters (also referred to as graph communities) exists in the topology of the semantic network. A commonly used metric for quantifying cluster quality is *modularity*, which measures the degree to which a cluster's membership has dense (or strong) edge interconnections while edge connections between clusters is relatively weak[13]. Each cluster is interpreted semantically as a topic in the corpus, based on the denseness of the weighted concept term relations. They are instantiated as new nodes in the Topics Layer, with connections to their constituent concept nodes in the Concepts Layer. Modularity clustering can be used for hierarchical clustering, allowing for large, less-cohesive clusters to be iteratively

decomposed further into more cohesive communities.

Because the nodes in the Topics Layer represent textual topics in the corpus, they must be presented to the information seeker as such. The current AIF framework identifies a set of terms to be used as representative terms for that topic (i.e., keywords). However, due to the varying size of the topics' concept membership and the hierarchical nature of the topic, the number of constituent terms may be too large to serve as a useful surrogate for that topic. Therefore, our approach identifies representative terms (derived from the topic cluster's concept nodes) as keywords for presentation to the seeker. The framework identifies representative keywords from the concepts comprising the topic/cluster. One method identifies the most salient concept terms, while another method analyzes the subgraph of concepts in the Concepts Layer that comprise the topic, and scores each concept's betweenness centrality (B_c)[34] for that cluster's subgraph. The complete set of all topics across the corpus are presented to the informations seeker as a hybrid collection of keywords so the seeker can understand the gist of that topic.

The human information seeker, being presented with a set of topics that cover the breadth of the corpus's information space, has the initiative to explore these topics to gain an understanding of the domain. After choosing one or more of the generated topics, the seeker may wish to explore the depth of that topic. Using the hierarchical graph, the AIF capability can present the concepts and edges in the semantic network that comprise only the selected topic, which is a subgraph of the larger network. Using the snippets linked to the constituent concepts, the seeker can choose to exploit this textual content and assess the relevant and/or lucrative information as evidence against decisions or hypotheses.

Each topic is a modular structure when compared against the larger semantic network, as identified by the graph clustering algorithm. However, these

clusters are identified hierarchically, and the selected topic cluster may not be semantically cohesive from the seeker's perspective because it may still cover multiple subtopics. In these cases, the seeker can direct the AIF capability to further decompose the topic into sub-clusters that are more cohesive. Upon decomposing the topic, the new sub-clusters are then added to the Topics Layer and linked with adjacent topics corresponding to the adjacencies in its Concepts Layer nodes.

The AIF hierarchical graph-based schema facilitates the presentation of corpus topics as well as the snippets that are linked to the concepts, providing the seeker with options across the exploration-exploitation spectrum. One can either traverse the breadth of the topics, decompose selected topics into smaller cohesive topics, or explore the detailed concepts comprising the topics. Therefore, both the breadth and depth of information is available to the information seeker. The snippets and their source documents are linked from these concepts to gain access to their raw textual content, enabling the exploitation of their content to enrich the information product.

The information seeker now has access to a more concise and navigable model of the information, generated organically from the topology of the concepts and relationships embedded in the entire corpus. This set of resulting topics, much easier to visualize and navigate than the corpus itself, may still be rather large and wide in its breadth of scope. Depending on the specific structure of the Concepts Layer and the level of cluster decomposition, a potentially large number of topics may be generated, some of which will have lesser relevance to the information seeker. Additionally, documents taken from the web (especially news articles) often have additional unrelated text snippets, such as advertisements and article teasers, causing the generation of topics that are semantically unrelated to the original corpus. For example, a corpus of news articles pertaining to Hurricane Irma may

contain mentions of articles about terrorism. While these topics seem anomalous or tangential to the information seeker, the AIF framework is correctly identifying these semantic concept clusters as distinct topics in the corpus.

As discussed in the Pirolli foraging and sensemaking model, information seekers need to reduce the information space from a "raw" collection, to a relevant collection, then to a lucrative collection of information. The last step of the AIF framework's processing flow assists the information seeker in restricting the corpus-wide set of topics down to a relevant set, and used iteratively to further restrict this down to a lucrative set.

The information seeker can provide the AIF capability with a more refined search criteria, based on his exploration of the topics. Continuing the previous example of *Hurricane Irma*, the seeker discovered an interesting set of topics relating to the health effects of the storm's aftermath. The refinement search provided to the AIF capability can be more specific, such as "health". Since the entire corpus was constructed around the broad search for "Hurricane Irma", there is no need to add this term as a refinement search constraint, as it is implied in the corpus itself.

The AIF relevance assessment step will score all topics based on this refinement search. It employs a spreading activation algorithm (SAA) across the semantic network contained in the Concepts Layer. Each of the concept nodes matching the criteria in the refinement search terms are activated, and that activation is propagated (though at a decreased level) to adjacent concepts, until an activation threshold is met on a per-node basis[38]. Once the propagation ceases, each topic is scored based on the ratio of the sum of its concepts' activation level relative to the number of concepts in the topic. The presentation of topics to the information seeker are then restricted to those topics that contain activated concepts, and ranked by this activation score, where the highest scoring topics are

considered most relevant. The seeker can choose to clear the relevance scoring, reverting back to the complete set of topics, select a new refinement search criteria, or to add a new refinement search that will further activate the Concepts Layer nodes with an additional activation signal, in essence combining multiple SAA propagations.

Chapter 4

The Corpus

The Corpus is comprised of the bottom-most two layers of the Automated Information Foraging (AIF) hierarchical graph-based schema, the Documents Layer and the Snippets Layer. These two layers model the information content in its raw, unstructured textual form. The AIF schema provides for one graph node per unique document (e.g., article, blog post, web page) instantiated within the Documents Layer. Each snippet node represents a discrete subset of the document, such that there is at least one snippet per document, where a snippet can be a paragraph or section of a document. It is from these two layers that the corpus is deconstructed into its atomic elements, document terms. A term can be a single word or a group of coincident words representing a noun phrase or a named entity (e.g., person, organization, location). These terms are then used to construct a corpus dictionary that maintains statistics on each term in the corpus, especially the frequency of a term within the corpus. This dictionary will be used throughout the AIF framework's processing. This chapter describes the corpus deconstruction process and the subsequent creation of the Corpus Dictionary, the Documents Layer, and the Snippets Layer.

4.1 Technical Challenges

There are several technical and practical challenges that this process is intended to overcome. The first challenge involves the common use of *Circular Reporting* in media, where the source of a document may seem like an original source, but may in fact be a copy from another source. In practice, multiple news outlets will often publish the same article under their organization's banner, resulting in many copies of the same article being published throughout the web and print. Although the original author is usually attributed to the material, search

engines and news aggregation databases will have multiple near-identical copies of the same article. There may be minor differences among the copies, such as advertisements and links to related articles, but the core content is usually the same. Circular reporting can significantly skew the term frequency statistics in the corpus dictionary, and therefore, it is desirable to eliminate these copies before the dictionary is finalized. The challenge arises in how to identify near-similar documents.

A secondary challenge is determining how to identify a term from a stream of tokens (i.e., words) in a document. There are several approaches in the information retrieval domain. This challenge is exacerbated because (in Chapter 5) the AIF framework will attempt to resolve terms across multiple snippets that may have differing case, plurality, or tense. This chapter will discuss various approaches attempted, and the methods chosen based on in-practice observations.

4.2 Corpus Creation Through the Web

Phase I of the AIF framework's process, Decomposition and Analysis of the Unstructured Text, begins with one of two conditions: a corpus is provided to the AIF framework, or the information seeker provides a broad search criteria for the automated collection of the corpus. There are numerous methods by which a corpus can be identified and acquired from the web. This section describes the method used by the author. No matter which approach is used, creating a document collection that covers a wide breadth of topics and from as many independent sources as possible will yield the best results, which is a corpus rich-enough for information exploration and diverse enough to mitigate bias.

Due to restrictions in end-user license agreements and application programming interfaces (API), many search engines are not able to provide a search capability that meets the needs of information foraging. However, there is a data

source that indexes an extensive collection of news articles, the Global Database of Events, Language, and Tone (GDELT) Project Global Knowledge Graph index[16]. GDELT monitors and aggregates news from across the globe in over 100 languages, making this an excellent data source for the AIF framework as it draws from a diverse set of authors and organizations with varying perspectives. Access to GDELT data is accomplished through the Google BigQuery API[17], which has a Structured Query Language (SQL) dialect. One of the GDELT tables is the Global Knowledge Graph (GKG) that contains extracted entities (names, organizations, locations) and themes. What is returned, however, is not the documents, but rather the Uniform Resource Locator (URL) of the published document, stored in the *docid* column.

Once the set of candidate document URLs has been retrieved from GDELT, the AIF framework downloads each one from its specified URL. If the document is inaccessible (e.g., HTTP 404, behind a paywall, unallowable), that document is ignored. Once downloaded, the textual content is extracted. For this prototype, the *BoilerPipe* library[18] was used, which uses HTML boilerplate detection approaches to eliminate navigation elements, templates, and some advertisements[19]. This process produces an unstructured text string that contains the main body of the HTML document.

4.3 Data Cleaning

The extracted content, in most cases, requires further cleaning. In most news articles, additional content is present, such as: author, editor, and contributor attribution, the name of the news organization, email addresses, social media hashtags, and URL source website. While this is valuable information for an information seeker to assess source credibility, this extraneous text will skew the corpus dictionary term frequency statistics and must be filtered. In practice, there

are also some common phrases that frequently appear in news articles extracted from websites, such as “click to follow,” “sign up today”, and “related articles.”. The AIF framework maintains a list of common text strings that should be removed when encountered as they will also skew term frequency statistics. The document is scanned for HTML, Javascript, JSON, or XML content that may have survived the BoilerPipe extraction process.

4.4 Document Layer Nodes

For each remaining document, a new *Document Node* in the Documents Layer of the AIF hierarchical graph is created. This node will be assigned several attributes, as shown in Appendix B, so that the process and the information seeker will always have access to the document’s metadata. Once the document node is created in the hierarchical graph, the cleaned text is indexed by the Lucene engine, allowing for later search[20]. The document identifier used as the primary key in the index is the vertex identifier (*vertexId*) of the document node.

4.5 Identifying Terms

Prior to building the corpus dictionary, each document’s cleaned text must be broken down into its constituent terms using natural language processing (NLP) techniques. For this prototype, the Apache OpenNLP library was used[21].
The first step in the term identification process is the detection of individual sentences. Each sentence is then *tokenized* such that each distinct word is added to an array of strings, one word (i.e., token) per array element. Many of the downstream NLP models require tokenized sentences as inputs and is therefore a required step. A set of Named Entity Recognition (NER) models analyze the tokenized sentence to identify persons (i.e., names), organizations, and locations within the sentence. These named entities will become individual document terms. The next step will use a Part of Speech (POS) Chunking model to create a tree

structure that identifies hierarchical relationship between groups of tokens and the parts of speech they form. For example, the sentence, "Hurricane Irma will make landfall and strike the Florida coast on Wednesday" will result in the POS tree shown in Figure 4. Each token is represented as a leaf node of the tree, while the non-leaf nodes represent higher-order parts of speech, such as noun phrases (np), verb phrases (vp), and prepositions (pp), where the codes are defined by the Penn TreeBank specification[22].

The AIF framework will create terms from the noun phrases that are encountered in the sentence tree[23]. The information retrieval community has shown that using noun phrases can often produce improved search results and term clustering results[24] which will play an important role in the generation of the semantic network layer. In this example, the AIF framework will look for the lowest-level noun phrases (np nodes) in the tree, and extract the noun tokens (nn), grouping them into a single term. For the example in Figure 4, the following noun phrases will be extracted: *hurricane irma*, *landfall*, *florida coast*, and *wednesday*. Note that this will eliminate some tokens such as determinants (e.g., "the"), providing a more accurate representation of the concept. Additionally, any terms that were identified as part of a named entity will not be duplicated, as the named entity will take precedence. For any noun phrases consisting of a single token, such as *landfall*, the token will be stemmed, meaning that its root (i.e., stem) will be used in its place. All plurality and tense is eliminated thus making term resolution possible in later steps when creating the semantic network. This stemming process uses the WordNet Electronic Lexical Database[25,26], accessed via the Java WordNet Library[27]. Once every sentence in each document has been decomposed into a set of distinct terms, the corpus dictionary can now be constructed.

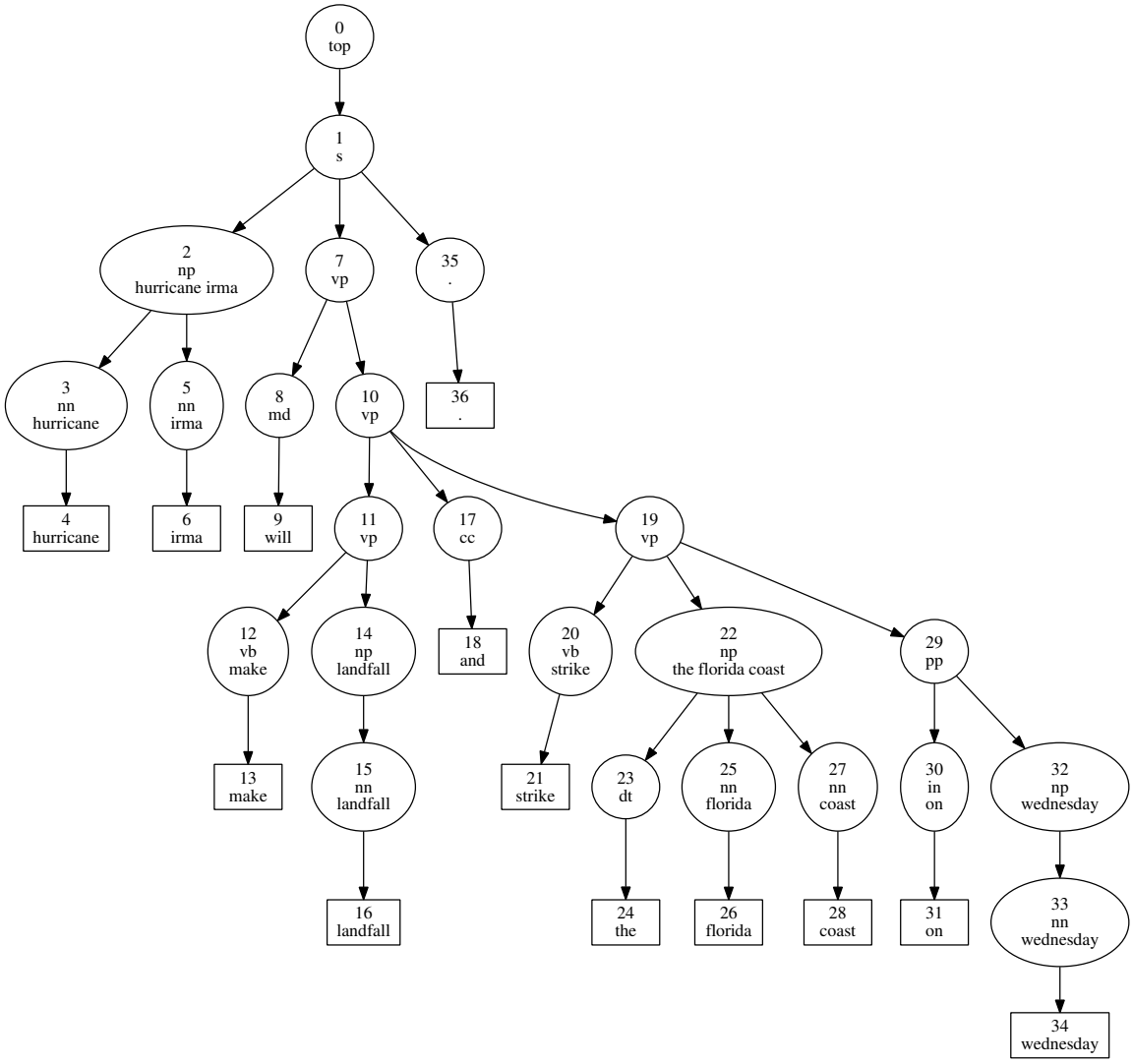

Figure 4. Part of speech trees identify noun phrases

4.6 Corpus Dictionary

The corpus dictionary maintains the complete set of terms and term frequencies used across the document corpus as well as for each individual document (and later, snippets). This is accomplished by creating a series of term vectors, which are ordered lists of values where the index (i.e., key) of each element is one term in the corpus and the value of that element is an integer representing the

term's frequency, as shown in Figure 5. An additional term vector represents the document corpus vector and its integer values are the sums of all frequencies for each document's vector elements. The term vectors and their frequencies are used throughout the AIF framework's processes for comparing documents and quantifying the importance of terms within the corpus and the documents.

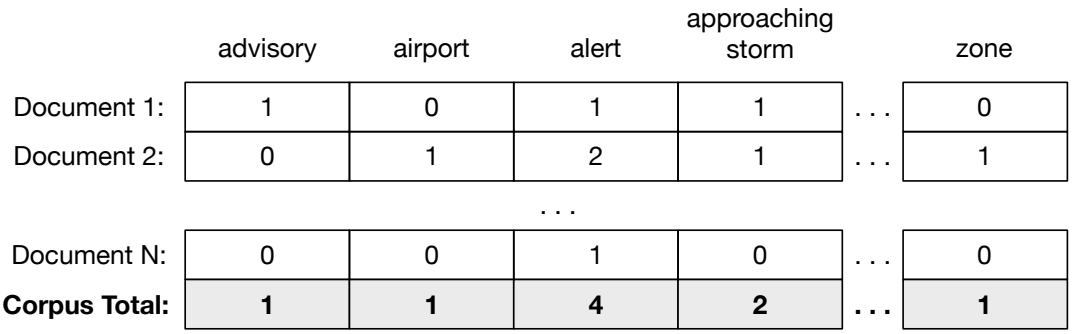

Figure 5. Term vectors in the corpus dictionary

4.7 Document Similarity Filtering

The first use of the term vectors addresses the challenge of circular reporting, where the AIF framework must identify identical or nearly identical documents. Having documents that use the exact (or almost exact) set and frequency of terms add no semantic value to the corpus and can cause their term frequencies to be over-represented in the document corpus vector. The AIF framework will use each document's term vector to generate pairwise document similarity scores using a cosine normalization[28]. This metric takes into account the direction of the vector where a value of 1 means the documents are the same, and 0 means they have nothing in common. This equation is shown in (1), where D is the first document's term vector, and E is the second document's term vector.

$$D = d_1, d_2, \ldots, d_n$$

$$E = e_1, e_2, \ldots, e_n$$

$$sim(D, E) = \frac{\sum_{i=1}^{n} d_i \cdot e_i}{\sqrt{\sum_{i=1}^{n}(d_i)^2 \cdot \sum_{i=1}^{n}(e_i)^2}} \quad (1)$$

If the similarity score between two documents exceeds a threshold (very close to 1.0, such as 0.95), one of the documents is removed from the corpus, and its Documents Layer node is removed accordingly. When this similarity filtering process is completed, the affected frequencies in the corpus dictionary are similarly updated to maintain accurate statistics.

4.8 Snippets

Once the set of document nodes has been finalized and filtered, the AIF framework is ready to subdivide the unstructured text from the each documents into a set of snippets nodes (for the Snippets Layer). There are three ways that the AIF framework can be configured to subdivide the documents. The first method subdivides each document by paragraph, looking for "newline" breaks in the text. The second method looks for distinct sections in the document, separated with an introductory title line. Since the text is unstructured and there are no semantic markings, a title identification heuristic looks for detected sentences (using Apache OpenNLP) that are adjacent to newlines with no trailing punctuation. Sections must contain at least one paragraph, but often contain two or more. The final method is to do no subdivisions at all, using the entire document as a single snippet.

The progression of each method (paragraph to section to whole document) monotonically increases the size of the snippet. As the snippet size increases, more

terms will be linked in the subsequent semantic network. However, there is a cost to having a wider scope for term relations. The level of relatedness of the terms will often decrease with size, as does the semantic cohesion of the entire snippet. In practice, the section-division approach seems to produce the best subjective results, since article authors will tend to organize related paragraph together in sections. When whole document snippets are used, the semantic network tends to be over-connected, causing poor topic clustering, while paragraph-based snippets tend to miss many obvious term-term relationships.

The snippet nodes are created and added to the Snippets Layer, then linked to the document node from which it was extracted. These nodes will be assigned several attributes, as shown in Appendix B, so that the AIF framework and the information seeker will have access to the snippet's metadata.

Chapter 5

The Semantic Network

Constructing the Concepts Layer is the most critical part of this automated information foraging (AIF) framework as it lays the foundation for all graph analytic steps to follow. It is through the construction of its semantic network that the natural language terms in the corpus dictionary take on a semantic meaning to become *concepts* based on their interrelationships with other terms. The semantic network that comprises this layer enables the entire exploration-exploitation spectrum by establishing concept-concept (i.e., term-term) relationships for exploration as well as concept-snippet relationships for exploitation. The structures that organically emerge from the topology of this network allow the information seeker to discover concept clusters that serve as potential information patches[7].

5.1 Technical Challenges

There are two key technical challenges that are encountered during the construction of the semantic network. The first is the selection of terms to use in the semantic network. Having too few concepts will reduce the breadth and depth of concepts that can inform the information seeker and allow organic topological structures to form. On the other hand, having too many concepts will potentially saturate the semantic network with low-value concepts whose interrelationships can cause weakly cohesive topics to emerge. The process by which salient terms are identified as candidate concepts should be configurable at runtime to maximize the value of the semantic network for the information seeker. Having the ability to tailor the salience level of terms can ensure that the AIF approach can work with diverse information corpora.

Another key technical challenge is ensuring that the relationships between concepts in the semantic network accurately reflect the relative strengths of those

relationships in the corpus. The eventual clustering of these concepts into topics depends heavily on the weights assigned to these concept-concept links. Weights that do not reflect the relationship strengths in the corpus will not convey the same semantic meaning to the information seeker during the exploration of the information space.

5.2 Identifying Salient Concepts for the Semantic Network

The first step in creating concept nodes for the semantic network is to identify the salient terms from the snippets. Salience, however, cannot simply use the highest term counts or frequencies as its measure. Doing so may bias the membership of the semantic network toward commonly used terms that may not be indicative of the core and important concepts. Rather, the AIF framework employs the *Term Frequency-Inverse Document Frequency* (TF-IDF) numerical statistic for normalizing term frequencies[29]. The TF-IDF value increases proportionally to the number of times a term appears in the snippet (used in place of a document), but is inversely weighted by the term's frequency throughout the entire corpus. This is a widely-used approach because of its utility in weighing term relevance for documents. The variant of TF-IDF implemented is shown in (2), where f_{tf} = raw frequency of the term in the snippet, N_s = number of snippets, and n_{t} = number of snippets containing the term.

$$V_{tf,idf} = (1 + log f_{tf}) \cdot log(\frac{N_s}{n_t}) \tag{2}$$

As each snippet is examined to find candidate concept terms, every term is scored with a TF-IDF value. After all terms have been scored, they are sorted in ascending order by that score. It has been observed that when the TF-IDF scores

are plotted with the scores on the y-axis and the ordinal on the x-axis, as shown in Figure 6, the scores tend to increase slowly then have a sharp increase. From the inflection point to the rightmost value, the framework considers these terms the most significant for the respective snippet, and must be included as concept terms in the semantic network.

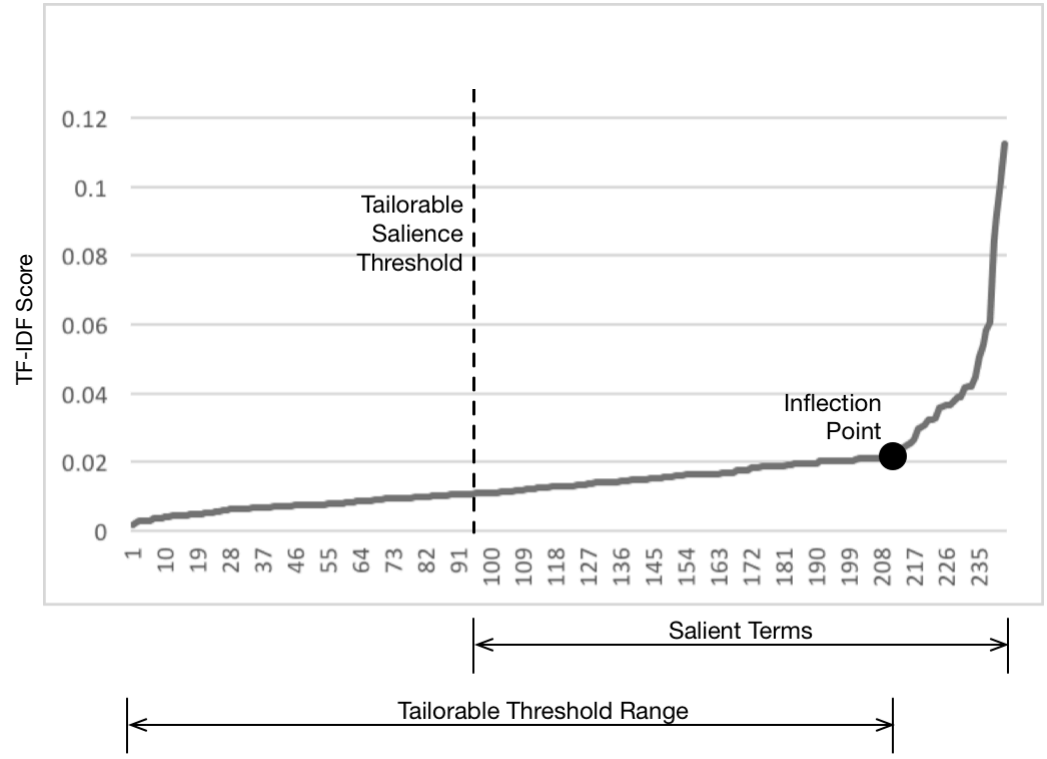

Figure 6. TF-IDF ranking for selecting salient terms

However, instead of only including these most-salient terms, the AIF framework will extend the salience threshold to a lower index in the term series as a means to add more context to set of concept terms. More contextual terms ensures that some lower ordered terms may also be considered salient, thus not inadvertently missing any significant terms. More importantly, having more

contextual terms allows a larger number of inter-snippet relationships, allow for the semantic network to have a higher connectivity. This connectivity is critical for enabling exploration of concepts and for identifying topics in the network topology. The AIF framework allows for the actual runtime salience threshold to be set by the information seeker, allowing the threshold to be tailored to the needs of individual corpora. This threshold can be set to any value between 0 (include to all terms from the snippet) to the inflection point (include only the most significant terms).

The framework identifies the index in the term series at which the threshold is set. If the terms adjacent to the threshold term have the same TF-IDF score, then the threshold is adjusted leftward to incorporate all "ties." Once all salient terms have been identified (all terms to the right of the threshold), each term is instantiated as a concept node in the Concepts Layer of the hierarchical graph. If a node for that term already exists, a new one is not created. However, all snippet nodes that spawned this concept node will be linked to it. These nodes will be assigned several attributes, as shown in Appendix B, so that the AIF framework and the information seeker will always have access to the concept term's metadata. Many of these attributes, however, are added at a later stage in the processing flow.

5.3 Computing the Weights of Concept-Concept Edges

The next step is the generation of edges between the concept term nodes. Since the AIF framework assumes that a snippet is a cohesive set of sentences and terms, all concept term nodes created from a single snippet will be adjacent to each other in the semantic network. The second assumption is that the closer the concept terms appear to each other in the snippet, the stronger their semantic relationship. Therefore, the weight of the edge uses a distance-based scheme that gives a weight of 1.0 for terms in the same sentence, and decreasing value as they get farther apart, where distance is how many sentences away the terms are. The

weight value will always be positive, thus ensuring that all intra-snippet relationships will be nonzero, however the values may be very close to zero as sentence distance increases. The weight is computed with an inverse sigmoid function, where function parameters are able to change the shape and width of the sigmoid curve, as illustrated in Figure 7. An inverse sigmoid function has several advantages. Both the upper and lower plateaus are asymptotic, so that a value in the range of (0..1) is ensured. This function produces a value close to 1 for smaller inter-sentence distances, and decreases as the distance increases. Additionally, by being able to alter the rate of change, inter-term relationship strengths can be tailored. The edge weight (W) is computed as using:

$$W = \frac{1}{m + e^{kx-d}} \tag{3}$$

where m is the maximum weight value (usually 1.0), k affects the steepness of the curve, d is an offset in the x-axis to translate the curve, and x is the sentence distance between two terms (0 means the same sentence, 1 means adjacent sentences, etc). If the same term-pair exists multiple times in the snippet, then the maximum W for that term-pair is used.

If an edge connecting two concepts already exists, the new W for the edge is added to the existing edge weight. Therefore, the structures and weights in the Concepts Layer are an aggregation across all snippets and represent the weight of the inter-term relationship across the corpus. At the completion of the edge creation, all edge weights across the semantic network are normalized as a value between (0..1).

Once the corpus-wide set of concept terms and their relationships are

Figure 7. Inverse sigmoid edge-weight function

instantiated and modeled into a single graph, the semantic network has been created. An example semantic network (from the Hurricane Irma corpus) is shown in Figure 8, in which topological structures can be visualized. These structures represent cohesive communities of concept terms that can serve as "information patches" in the Pirolli and Card Information Foraging Theory[7] model. We will see, however, that the topological structures that will be used as topic clusters (discussed in chapter 6) may not necessarily be evident in this view of the semantic network because the edge weights are not shown. As shown at the bottom of the figure, sets of disjoint subgraphs may also emerge from this process. These disjoint communities model individual snippets that share no concept term nodes in common with the rest of the graph. These disjoint communities often model the tangential snippets common in news articles, but not in all cases. No assumptions can be made yet regarding their relevance, as this will be addressed in chapter 7.

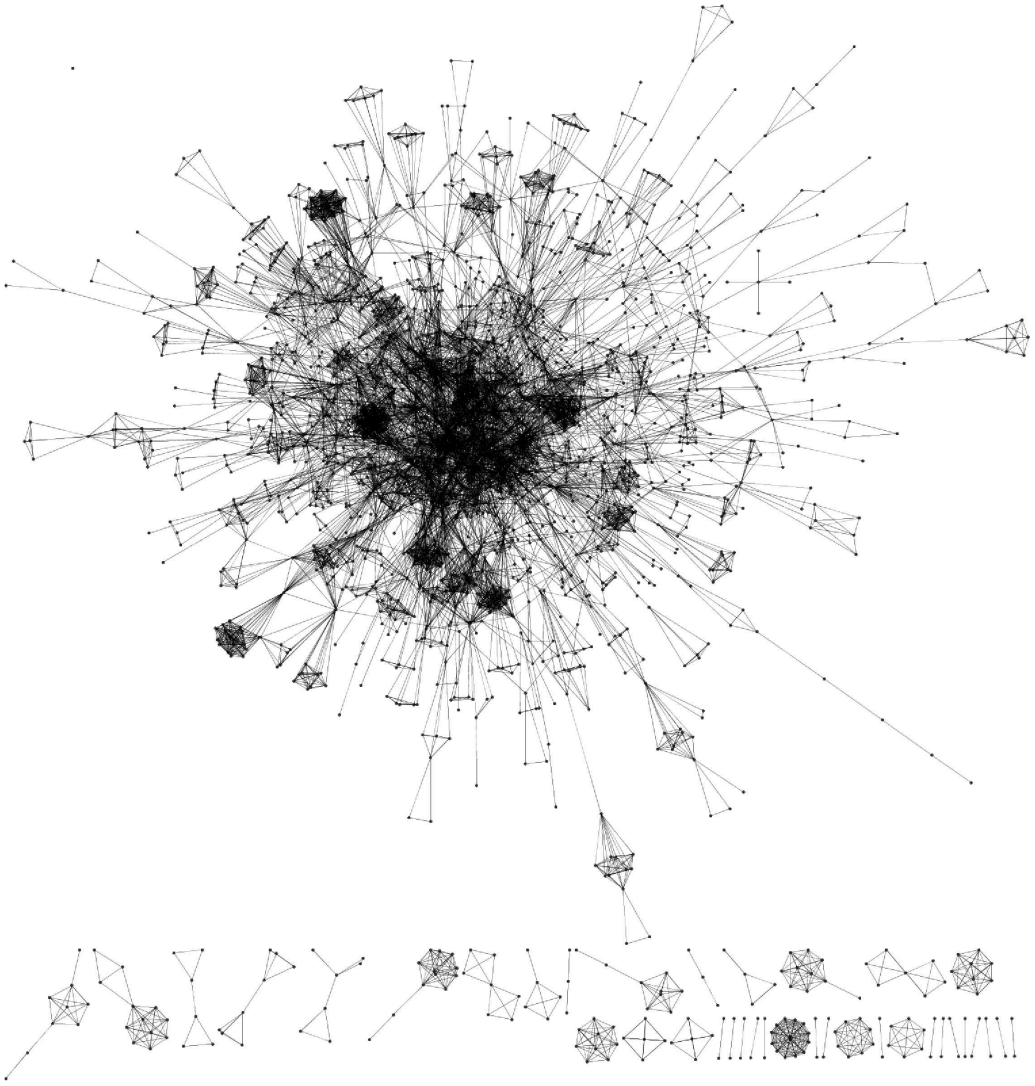

Figure 8. Example semantic network

Chapter 6

Topics

The second phase of the Automated Information Foraging (AIF) framework's processing flow analyzes the semantic network to identify and exploit its topological structures enabling information exploration. This chapter focuses on identification and modeling of *topics* that emerge from the corpus through this topology, and how to present these topics to the information seeker.

6.1 Technical Challenges

Identifying a set of topics from the semantic network topology is a unique approach to topic modeling. Popular methods in topic modeling, such as Latent Dirichlet Allocation (LDA), are statistical models that identify topics based on the likelihood of term co-occurrence[30]. In this approach, there is little or no semantic understanding of the term relationships. Because one of the goals of AIF is to help information seekers discover "hidden nuggets" of information based on semantic relationships, these probabilistic models may not be effective for that end because these hidden nuggets are often have a lower likelihood of co-occurrence, which is why they may be hard to find. Another goal of AIF is the facilitation of the breadth *and depth* of information topics. While methods such as LDA can certainly establish a broad set of topics based on the corpus, these topics may vary is size and scope. If an information seeker chooses to explore a single topic's concepts and semantics, that topic may need to be decomposed into sub-topics, a method not well suited for LDA.

Based on experiments with topic modeling, it was discovered that some corpora's topics can overrepresent (or bias toward) locations. This phenomenon is especially true for news articles, whose reports are often location-based, since the town and city names are prominently mentioned in the text. The effect is that

numerous topics emerge for one or more events that occurred at certain cities and towns. While this is a valid subset of topics, the information seeker may need topics that are more conceptual or event focused. Accordingly, the AIF framework would need methods to de-emphasize locations in favor of these abstract concepts, events, or people. Because the LDA approach does not take semantic analysis into account, the information seeker will need to analyze the topics in more detail to overcome this potential location bias.

6.2 Graph Clustering: Community Detection

The AIF framework seeks to create semantically-sensitive topics by identifying strongly connected sets of nodes within the semantic network, known as *communities*. These community structures (also called graph *clusters*) are a feature of real-world graphs where there is a high concentration of edges within a node group, and a low concentration between such groups[31]. A commonly used metric for quantifying cluster quality is *modularity*, which measures the degree to which a cluster's membership has dense edge interconnections while edge connections between clusters is relatively low[13]. Modularity can also take into account not just the density of edges between nodes to identify communities, but also incorporate the weights of those edges, thus enabling the graph clustering to identify strongly connected groups of nodes[32].

There are several widely used methods for identifying clusters within graphs. The AIF framework imposes several requirements for selection of an algorithm.

- The algorithm must be efficient at scale for real-world corpora that may contain an excess of thousands of nodes and tens of thousands of edges.
- The clusters are identified hierarchically as a means to facilitate information exploration in breadth and depth.

- The algorithm must incorporate edge weights to assess the strength of a cluster rather than simply the density of its edges.
- A predefined number of partitions cannot be used to determine clusters. For information foraging, the number of clusters will be dependent on the corpus.

For the AIF framework, three popular classes of algorithms were investigated to determine an optimal approach to community detection. These classes are: Graph Partitioning, Divisive Hierarchical Clustering, and Hierarchical Maximum Modularity Clustering. These approaches (discussed in Chapter 2) were compared against semantic networks produced by the framework.

6.2.1 Graph partitioning. Graph partitioning optimizes a function that determines the ratio of the number of edges between modules to the number of edges within modules. This method, although efficient and widely used for various applications, requires a predefined number of clusters to be provided to the algorithm. Having the information seeker specify this value could inject potential bias into the process. An automated iterative approach could try ranges of values, but this approach is inefficient and requires a heuristic stopping condition. Other approaches proved more effective.

6.2.2 Divisive hierarchical clustering. The Girvan-Newman approach to Divisive Hierarchical Clustering uses the Betweenness Centrality (B_c) Metric of the graph's edges to find clusters by removing the edge with the highest B_c score, then recomputing the B_c scores for the entire graph again. This is repeated until all edges have been removed or a modularity-score stopping condition is achieved, resulting in a dendrogram.

Betweenness Centrality is a measure of the centrality of nodes (and extended to edges) in a graph based on all-pairs shortest paths[34]. For every node

and edge in a graph, B_c is the number of paths that pass through it. For weighted graphs, B_c measures the sum of weights is used. In the Girvan-Newman algorithm, removing the edges with the highest B_c score essentially removes the most highly traversed edges across the all-pairs paths, thus eliminating the likely inter-cluster edges. Once all edges have been removed or when a modularity threshold is reached, the resulting dendrogram is analyzed to find the hierarchically-generated clusters.

In practice, this algorithm worked very well for small and moderate sized graphs of a few hundred edges. However, with a large graph containing thousands of edges, it proved very inefficient, primarily because after each edge removal, every remaining edge's B_c must be recomputed.

6.2.3 Multi-level algorithms for modularity clustering. Girvan and Newman first used the modularity score for assessing the quality of the community structures in the graph. Rossi, et al. developed a hierarchical maximal modularity algorithm that extends the Noack and Rotta approach by testing clustering significance using random graph generation and recursive hierarchical clustering[35,36]. This method is also useful for graph visualization applications[37] that can hierarchically decompose not just the semantic network as a whole, but also individual clusters' subgraphs while maintaining inter-cluster connections. The approach lends itself to the information foraging problem because information seekers will follow the data both in breadth (across clusters) and in depth (decomposing clusters) as needed, and these inter-cluster links can provide an exploration path such that AIF visualizations and user interfaces can take advantage of this added information. The Rossi algorithm meets each of the AIF requirements for clustering the concepts in the semantic network, and is used for such in the AIF prototype.

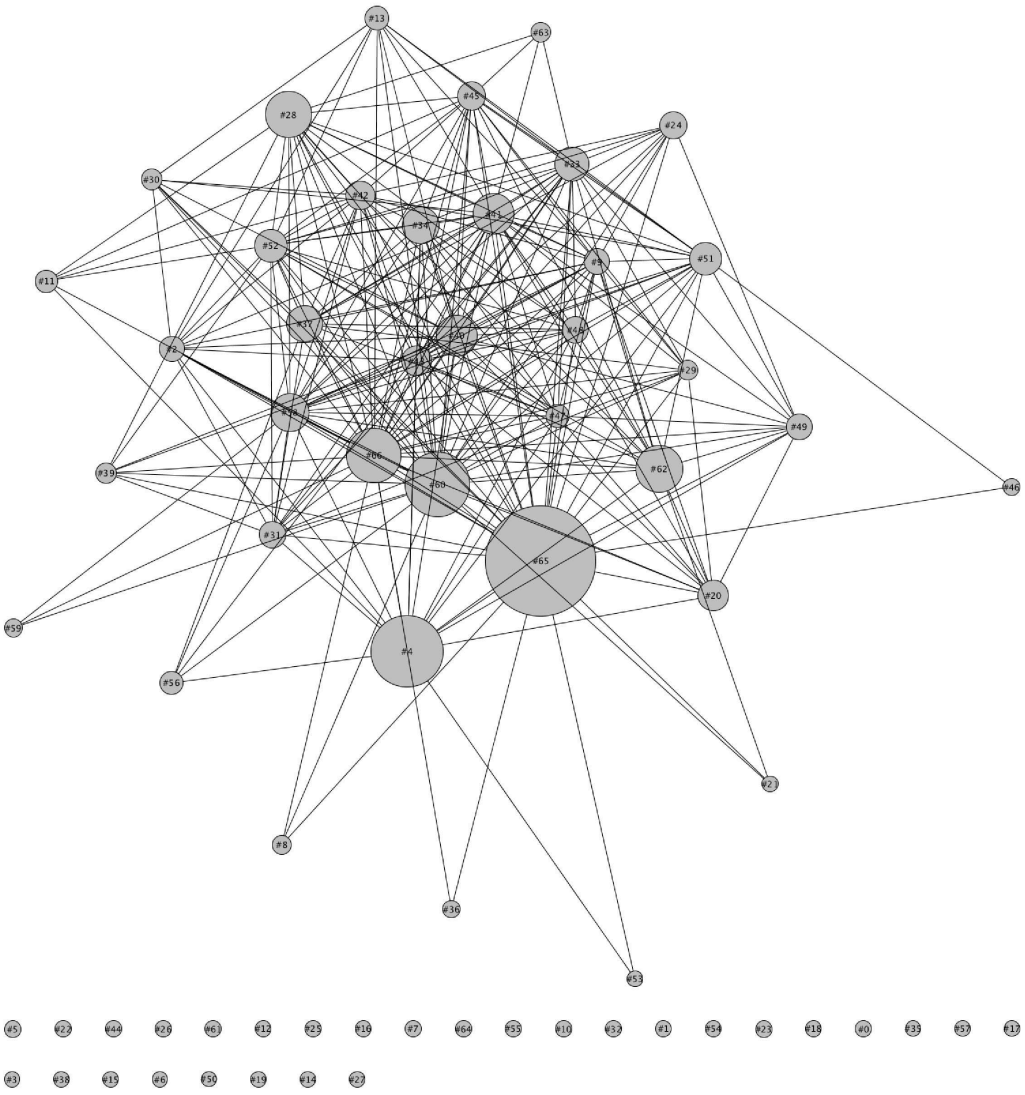

Figure 9. Example topic layer showing concept clusters

6.3 Creating the Topics Layer

For each cluster of concepts identified by the Rossi community detection algorithm, a new node is added to the Topics Layer of the AIF hierarchical graph, and edges are added between the topic node (i.e., cluster node) and each of the semantic network concept nodes that comprises it as well as edges to adjacent topics. Figure 9 shows the resulting Topic Layer after the first-level hierarchical

decomposition has been applied to the semantic network in Figure 8. The size of the topic node indicates the relative size of its cluster (in terms of the number of constituent semantic network concept nodes). Additionally, the disjoint subgraphs in the semantic network are modeled as disjoint topic clusters.

Each cluster is interpreted semantically as a topic in the corpus, based on the denseness and strength of the weighted concept term relations. These topics represent strong relationships of concepts potentially spanning multiple snippets and documents from the corpus. These relationships are formed not from likelihood of co-occurrence, but from actual links in the text, facilitating the possibility of linking concepts with low frequency to related concepts of higher frequency.

Each of the topics is linked to the concept nodes that comprise it. Therefore, the AIF framework can extract the semantic network subgraph, G_t, that represents the topic as shown in (4).

$$G_t = (V_t, E_t) \tag{4}$$

where: T = the set of concept nodes comprising a topic,

$V_t \in T$,

$E_t \in (u_t, v_t)$,

$u_t, v_t \in V_t$

6.4 Cluster Subgraph Analysis

By extracting the semantic network subgraph that represents the topic, the AIF framework provides additional capabilities for information foraging. Just as the entire semantic network's topology was analyzed for structure, so too can each G_t subgraph. For instance, the same community detection algorithm that generated

these topics can be applied to G_t to create a lower-level set of sub-topics for selected cluster nodes. This allows two possibilities: automated decomposition of topic clusters to create a larger set of smaller topics in the Topics Later, and allowing the information seeker to hierarchically explore the "depth" of the topic by examining lower-level subtopics.

The AIF framework analyzes and maintains several additional metrics about the topic's semantic network subgraph. e_{max}, the maximum possible number of edges among concepts in G_t, is shown in (5). e_c, the *completeness* of G_t, as shown in (6), is used to assess the density of edges within G_t as the ratio of actual number of edges to the maximum possible number of edges.

$$e_{max} = \frac{|V| \cdot (|V| - 1)}{2} \tag{5}$$

$$e_c = \frac{|E|}{e_{max}} \tag{6}$$

Another analysis that is computed is the Betweenness Centrality (B_c) for each node in each G_t. By computing B_c within the context of the G_t subgraph, this process identifies the concept nodes that are "most central" for that topic. This centrality metric is only done when the completeness metric is significantly less than 1. When e_c is exactly 1, no B_c can be computed as there is no concept of centrality for a complete graph. When e_c is close to 1, the resulting B_c scores are not statistically meaningful. The B_c scores will be used in section 6.6 to identify surrogate concepts/terms to represent each topic to the information seeker.

The degree of each topic node (D_t), the number of the topic node's adjacencies within the Topics Layer, is computed as a potential ranking order for

presenting topics to the information seeker. Additionally, each of the snippets that are linked to this topic through its concept nodes are ranked (relative to each topic) based on the number of links between the snippet and the topic's concepts. When the snippets are presented to the information seeker, they will be presented in this order, where the most referenced snippets are shown first.

6.5 Automated Cluster Decomposition

The community detection algorithm produces a set of concept node clusters that accurately reflects the topological structures in the semantic network that maximizes the modularity metric of the graph. In practice, this produces a Topics Layer with a wide spectrum of topic sizes. Often, as the topic size increases (based on the cardinality of its concept nodes), the subjective cohesiveness of the concepts in the cluster diminishes because the larger clusters cover a wider range of concepts. These larger clusters' G_t may also have their own topological structures representing subtopics. The community detection algorithm can be executed against G_t to decompose the cluster into sub-clusters. When this is done, the parent cluster node in the Topics Layer can be removed, and the sub-clusters can be inserted in its place.

This thesis did not research methods to identify when a cluster should be automatically decomposed into sub-clusters. This is a research area that would significantly improve the results of the AIF framework's approach. For the results discussed in Chapter 8, the AIF framework executed the community detection algorithm twice, once for the whole semantic network, then again for each topic's G_t. This produces a larger set of topics, but these topics are smaller and more cohesive than a single pass of the community detection algorithm. A more effective approach would be selective decomposition of certain topics based on analysis of each G_t.

6.6 Topic Keywords

The topics represent a collection of concepts from the semantic network associated with the textual snippets and documents from which those concepts were derived. When presenting topics to the information seeker, a representative, or surrogate set, of concept terms should be used in lieu of the full set of snippets. This serves as a summary of the topic, and is referred to by the AIF framework as a collection of "keywords". This approach is familiar to the information retrieval community since this is how LDA topic models represent topics. But for many moderate-sized or large topics, the number of concept terms may be too numerous to be an effective summary, and therefore must be restricted to the most salient keywords. The framework has several methods to choose the representative keywords, as shown in Table 1. The number of keywords for each topic will be restricted to a presentation threshold, specified at runtime.

Table 1

Approaches for Selecting Topic Keywords

Approach	Description
Concept type	Segregate the concepts into their distinct concept types (e.g., location, person, organization, other) and rank them using another metric.
Max TF-IDF	Rank all the concept nodes by their maximum TF-IDF score (there is one score for each snippet).
Centrality	Rank all the concept nodes by their B_c score relative to G_t.

In practice, based on subjective assessment, the Betweenness Centrality method provided the most representative keywords. This makes intuitive sense as the B_c score represents the most central nodes to the subgraph, or those nodes that are most traversed when random walks are performed within the subgraph. Based

on the topology of the G_t subgraph, not every node will receive a nonzero B_c score. Therefore the number of B_c keywords may be zero (for a fully connected G_t), or may be below the presentation threshold. In these cases, the set of keywords may be augmented with the top ranked concepts with the highest maximum TF-IDF scores.

6.7 Manipulating Edge Weights to Affect Topics

Relationships among concepts in the semantic network are constructed based on their co-occurrence in textual snippets. Therefore, the resulting topics (clusters of these concept nodes) will emerge based on the frequency of those co-occurrences and their relative strengths in the corpus. If certain concept terms occur more frequently than others, their is a higher likelihood that they will have more and stronger relationships with other concepts.

In many corpora, certain types of concepts (e.g., locations, persons, organizations) may have over-represented interrelationships. For example, news articles often report on events that occur at a specific location, rather than cover a wide breadth of a topic with contextual information. Other types of news articles (and corpora) may be about a prominent person or organization, and thus that concept node's adjacencies may be overly strengthened.

One advantage of using a semantic network, is that it models these relationships. During the semantic network construction process (chapter 5), the framework identifies concept terms that are *named entities* through the Named Entity Recognition process, and tags each concept based on its entity type. The AIF framework can manipulate the weights of the edges adjacent to specific types of entities (i.e., concepts) to add or reduce bias in the topics. For example, the information seeker may observe that the topics generated by the framework are biased toward locations, and wants to reduce this bias by 20%. The AIF framework can locate all edges in the semantic network adjacent to location-type concept nodes,

then scale their weights by 0.8. The community detection algorithm will analyze the semantic network and create a new topic layer that reflects the changes in concept relationship weights. The scaling can also be used to increase a bias in concept types if the information seeker needs this new view of the information landscape.

Chapter 7

Relevance Assessment

Depending on the specific structure of the Concepts Layer and the level of cluster decomposition, a potentially large number of topics may be generated, some of which will have lesser relevance to the information seeker. Additionally, documents retrieved from the web (especially news articles) often have additional unrelated text snippets, such as advertisements and article teasers, causing the generation of topics that are semantically unrelated to the original corpus. For example, a corpus of news articles pertaining to Hurricane Irma may contain mentions of articles about terrorism and entertainment news. While these topics seem anomalous or tangential to the information seeker, the AIF framework is correctly identifying these semantic concept communities as distinct topics in the corpus. Because AIF is intended to facilitate efficient exploration of the topics, a very large topic set may adversely affect that efficiency. Therefore, the AIF framework needs a mechanism by which topics can be ranked and filtered based on a relevance metric, while still allowing for information and topic discovery. Just as with topic generation, as discussed in Chapter 6, this relevance assessment metric is based on the topology of the semantic network using the inter-relationships among concepts to guide the exploration.

7.1 Technical Challenges

Relevance is often a subjective attribute to the information seeker, and is therefore challenging to quantify. Additionally, the relevance of a topic may not be known a priori, posing a challenge to the AIF framework. The information seeker, however, typically has some preexisting notion of topics that are known to be relevant, or through the exploration of the unranked topics, may have discovered such relevant topics or concepts. The AIF framework will leverage those known or

likely relevant concepts to find other topics associated with them in the semantic network using a form of information retrieval known as *associative information retrieval* (AIR). Given that associative information retrieval techniques operate on the semantic (or associative) network, AIF must employ a method that can quantify and rank the topic clusters, not individual concepts. The relevance assessment approach must be able to leverage what is known to be relevant, while still proving a means to discover other relevant topics not yet known.

7.2 Spreading Activation Algorithm

The information retrieval community has investigated the heuristic rule of using associative retrieval through a technique called spreading activation (SA)[38]. This iterative algorithm consists of one or more *pulses* or signals that begin from one or more nodes in the semantic network and propagate to adjacent nodes at some decreased level, as shown in Figure 10. For any single unit in the network (in the AIF framework, this unit is a concept node in the semantic network), SA first computes the input signal, I_j, as illustrated in (7). The output values of each node can be either a binary (0 or 1) value or a real number (0..1). For the AIF framework, a binary output is used. If I_j exceeds a threshold k_j as shown in (8), the node is activated and its output is propagated to its adjacencies.

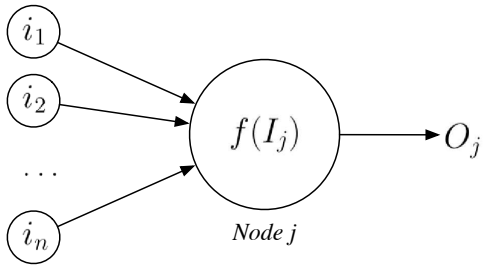

Figure 10. Spreading activation unit

$$I_j = \sum_{i}^{n} O_i w_{ij} \tag{7}$$

where: I_j = the total input of concept node j,

i = the node propagating the signal to j,

O_i = the output of node i to node j,

w_{ij} = the edge weight between i and j,

n = the number of adjacencies to j

$$O_j = f(I_j) = \begin{cases} 0 & I_j < k_j \\ 1 & I_j \geq k_j \end{cases} \tag{8}$$

This iterative propagation continues until the termination condition is met, or all I_j are less than k_j for all j. The value of k_j is configurable at runtime, allowing for the information seeker to tune the threshold to the specific network topology or desired level of information exploration. Because the semantic network is potentially cyclic, the SA algorithm used in the AIF framework also ensures that nodes are not re-activated from these cycles.

7.3 Spreading Activation Constraints

The topological structures in the semantic network that are used to cluster the topics may adversely affect the propagation of the SA signal. Specifically, the modularity-based clustering of the semantic network will encourage intra-cluster propagation but discourage inter-cluster propagation. Therefore, SA algorithms often use application-specific *path constraints* to create preferential paths for propagation[38].

The AIF framework provides a runtime variable that can adjust an edge's input signal, O_i based on whether the i and j nodes are members of the same cluster. This new input signal, $I_j^{'}$, will allow inter-cluster edges to propagate a minimum input signal if the sending node is activated, as shown in (9) given the adjusted output signal $O_i^{'}$ in (10), where $m = [0..1]$. This improves propagation of the signal to new topic clusters because inter-cluster edge weights can often be very small (close to zero).

The runtime configurable variable m, provides the information seeker with the ability to ensure that inter-cluster propagations have a minimum signal value, but they will not fall below the current edge weight. For instance, if m is set to 0.5, then the propagated signal will be either 0.5 or the edge's weight w_{ij}, whichever is higher, if nodes i and j are members of different clusters.

$$I_j^{'} = \sum_{i}^{n} O_i^{'} w_{ij} \tag{9}$$

$$O_j^{'} = f(I_j^{'}) = \begin{cases} 0 & I_j^{'} < k_j \\ 1 & I_j^{'} \geq k_j \wedge i \in G_t \wedge j \in G_t \\ \frac{max(w_{ij},m)}{w_{ij}} & I_j^{'} \geq k_j \wedge i \in G_t \wedge j \notin G_t \end{cases} \tag{10}$$

7.4 Initial Node Activations

Executing the SA algorithm on the semantic network first requires the identification of one or more *seed nodes* to serve as the initial activations in the iterative process. This is accomplished through a *refinement query* where the information seeker provides a search term based on a known relevant concept. The seed concepts can be identified through a priori knowledge or through the

exploration of the topics (prior to or after executing the SA algorithm). The AIF framework will search the existing concepts in the semantic network for matches against the refinement query, and select them as the initial SA nodes. Because the concepts in the semantic network are n-grams and the user has the ability to specify multiple refinement query terms, several concept nodes may be matched as initial seed nodes, causing a wider propagation of relevance.

7.5 Topic Relevance Scoring

Using spreading activation, all nodes in the semantic network will have an activation score, I_j. However, it is likely that a large subset of nodes will have a score of zero because they had no incoming signals. The AIF framework, however, needs to have a relevance score for each *topic* as a quantitative attribute for ranking. The Topic's relevance score, R_t is a ratio of the sum of all its concept nodes' relevance scores in the semantic network to its node cardinality of the subgraph G_t, as shown in (11), where $|v|_t$ is the cardinality of nodes in G_t.

$$R_t = \frac{\sum_{i=1}^{|v|_t} I_i'}{|v|_t} \tag{11}$$

Once each topic is assigned its R_t score, the information seeker is presented with a list of relevant topics sorted in descending order by R_t. Any topics whose $R_t = 0$ can be optionally omitted, providing the seeker with a smaller, manageable set of topics to explore.

7.6 Extensions to Relevance Assessment

Although not yet implemented in this prototype, there are other mechanisms to seed activations into the semantic network. One area that should be investigated is user feedback. As the information seeker selects specific topics, opens, or saves

snippets, these foraging actions could add input signal and/or activate the concept nodes associated with them, respectively. This feedback would create a more comprehensive view of what relevance means to the information seeker.

Chapter 8

Example Use Case

The Automated Information Foraging (AIF) framework is intended to assist information seekers throughout the exploration-exploitation spectrum, facilitating the understanding and discovery of topics within the corpus and to identify relevant and lucrative information documents. To demonstrate the AIF processing flow, this chapter describes using AIF on a real-world example from the information seeker's perspective, identifying topics relating to health dangers from Hurricane Irma (2017). The information domain relating to Hurricane Irma is a good exemplar use case for several reasons. Firstly, this information space has been widely covered in both local and national (United States) news outlets. This coverage is diverse in its frequency of publication from various news outlets, where some documents are one-time reports on specific events and situations, and others contain daily or hourly updates. Secondly, the published documents are representative of typical digital (i.e., web-based) news publications that contain unrelated snippets, advertisements, and exhibit circular reporting. Lastly, it is a domain in which most potential information seekers will have at least some familiarity and knowledge. However, unless they live in a region that is frequently afflicted by such storms, they may have knowledge gaps and pre-conceived ideas (i.e., biases) about important and relevant subtopics.

8.1 Corpus Generation

Initiating the AIF framework, the information seeker provides a broad, high-level domain topic, "Hurricane Irma". The AIF capability queries the Global Database of Events, Language, and Tone (GDELT) Project Global Knowledge Graph index[16] and subsequently downloads article content from the web (where available and permitted).

The AIF framework queried GDELT for the URLs of the first 1000 documents relating to "Hurricane Irma". After eliminating those URLs whose documents are inaccessible (e.g., behind a paywall or receiving an HTTP 404 error), and then identifying and removing duplicate documents, a corpus of 362 documents was created. After cleaning the documents, each was subdivided into snippets by looking for section breaks (document sections with distinct titles and one or more paragraphs), resulting in 1823 snippet nodes in the Snippets Layer. AIF identified the significant terms in the snippets and created a semantic network (in the Concepts Layer) of 2391 nodes and 19125 relationship edges among them.

8.2 Observations From the Generated Topic Set

There are several observations that can be drawn from examination of generated set of topics, as well as from a more detailed analysis of the their underlying concepts and snippets.

8.2.1 Topics for information exploration. The first observation is that topics generated by AIF does indeed cover a wide breadth across the domain. This provides the information seeker with the ability to explore the complete information space in a manner that is much less costly in terms of the time required to read the entire corpus. Some of the topics are expected based on common knowledge of hurricanes affecting Florida, including the following:

- Topics #165 and #250: Electrical and Sewer Utility outages.
- Topic #22: Sparking electrical transformer.
- Topic #24: Downed trees, debris, and closed roadways.
- Topic #91: Sinkholes that have formed.
- Topic #70: Closure of Orlando-area theme parks.

However, one of the goals of AIF is to help information seekers overcome potential biases and extant, a priori knowledge. This example has identified several such topics including:

- Topic #139: Carbon monoxide poisonings resulting from electrical generators.
- Topic #247: Insurance claims relating to hurricanes and flooding.
- Topic #72: Deployment of US Navy ships for Search and Rescue efforts.
- Topic #81: Looting and arrests.
- Topics #208 and #213: Government response to the management and transportation of prison inmates in evacuated areas.
- Topic #17: Destroyed sea turtle eggs on the beaches.
- Topic #1: Owners getting medicine for pets prior to evacuations.

8.2.2 Irrelevant topics. The second observation is that there are several topics that are irrelevant or tangential to even the broad search criteria provided by the informations seeker. Despite being irrelevant, these topics were correctly identified from snippets in the corpus. In most cases, these snippets are links to other published articles embedded in the HTML content of news articles. Such snippets would be challenging to identify, and false-positive matches for these snippets could risk removing relevant information. Examples include:

- Topic #62: International Trade.
- Topic #120: Terrorism.
- Topic #31: Polymath celebrities.

8.2.3 Topic presentation. The third observation is that the keywords are not always indicative of their relevance to the information seeker's broad search. One example is Topic #2, whose keywords are "artifact, loan, historic palm cottage, nancy holcomb". These keywords do not seem to related to "Hurricane Irma". However, when the one examines the snippet that is associated with that topic, the relevance becomes apparent. The snippet states,

> Nancy Holcomb, a historic preservation outreach coordinator for the Naples Historical Society, takes artifacts that are on loan and packages them for safekeeping at the Historic Palm Cottage in Naples, Fla. The Palm Cottage was built in 1895 and has withstood all weather events including Hurricane Donna in 1960.

In this example, the topic was indeed relevant and is likely a topic that would expand the information seeker's understanding of the domain. However, to fully meet the AIF goal of saving time and reducing the analytic burden of analyzing the documents in the corpus, more research is needed to identify the optimal method for relaying the gist of the topic to the information seeker.

8.2.4 Size of the topic set. The final observation is that AIF produced a large set of 342 topics that may be time consuming to browse and explore. This is a large number of topics to analyze, so the user can restrict this to a smaller and more focused set of topics by using a relevance assessment.

8.3 Relevance Assessment

To reduce the wide topic set to a more relevant and lucrative set, the information seeker can direct the AIF framework to execute a relevance assessment against the topics by entering a *refinement query*. For this example, the seeker specifies the terms "health AND dangers" to restrict the results further. The AIF

framework then identified 7 relevant topics and generated a ranked list that topic subset, as shown in Table 2. The resulting set of relevant, focused topics provide the information seeker with the ability to exploit the snippets and documents as evidence against a hypothesis or decision. This set of topics exposes the seeker with specific snippets that include:

- Excessive heat caused by power outages endangering elderly patients in a rehab center (Keywords: rehab, elderly people, tragic deaths).
- Carbon monoxide poisoning risks results from electrical generators, and chemical spills resulting from flooded industrial areas (Keywords: carbon monoxide, poison, chemical plant).
- Mosquito-borne illness (Keywords: mosquito, entomologist).
- Storm surge flooding a neighborhood (Keywords: storm surge, flooding, neighborhood).

These topics include results based on common-knowledge, such as flooding from storm surges. But more importantly, these results also include topics that were non-obvious to an information seeker unfamiliar with the domain, such as carbon monoxide poisoning, highlighting benefits of this approach.

8.4 Benefits and Opportunities

This example illustrates some of the benefits of the AIF framework, namely the ability to both explore topics within the corpus and exploit snippets and documents as evidence against specific refined topics. This example also highlights opportunities for improvement and future research, most importantly, the need to have a more informative title or textual summary for each topic rather than use keywords (i.e., key terms) from the semantic network.

Table 2

Relevant Topics from the Refinement Search

#	Size	Keywords
1	Concepts:17 Snippets:5 Articles:2	location, deputy, advantage, pinellas, access, base, purse, roadway, broward health medical center, broward county sheriff, sarasota county sheriff, marion county, clearwater beach, nicholas rossell, scott israel, lido key, coon key
2	Concepts:7 Snippets:4 Articles:3	wake, slam, flooded neighborhood, dangerous storm surge flooding, destruction irma, heres a look, us national hurricane center
3	Concepts:17 Snippets:5 Articles:4	ant, poison, monoxide, plant, carbon, cut, example, chemical, hurricane harvey, danger, dangerous, crosby, chemical plant, atmosphere, florida health department, turkey point, texas
4	Concepts:16 Snippets:9 Articles:9	mental, entomologist, cycle, people, reapri, treatment, caribbean islands, mosquito, hurricane katrina, dialysis, chronic, illness, disease, diabetes, american journal, kidney disease
5	Concepts:23 Snippets:2 Articles:1	big ask, rightfully, rehab, industry, state health care officials, lax regulatory approach, hollywood rehab center, repercussions, elderly people, tragic deaths, rehab centers, swelter, awaken, repercussion, lax, care industry apologists, broward county tragedy, industry s image, isolated incident, drubbing, apologist, tragedy, drub
6	Concepts:15 Snippets:7 Articles:6	requirement, goal, nurse, committee, representative, slosberg, justice, lawsuit, lawyer, senator, senior citizens, health and human service, stockton, delray beach, nelson
7	Concepts:9 Snippets:7 Articles:6	airport, hiv, health, flight, track, gobeil, delta air lines, american airlines, atlanta

Chapter 9

Experiments

This chapter summarizes a set of experiments intended to further an understanding of how the key AIF framework parameters affect the information seeker's ability to explore the information in a corpus.

9.1 Key Parameters

The AIF framework has several key parameters that affect the configuration of algorithms as listed in Table 3. *Document Segmentation* is the method used to subdivide each document into one or more snippets. The *Concept Extraction Threshold* specifies the top percentile of concepts that were identified within each snippet, as ranked by their TF-IDF scores. *Cluster Decomposition* specifies how many levels of topic clustering is executed. The first pass is for the whole graph. The second pass is for each topic cluster's subgraph G_t, etc. The *Edge Weight Distance* is the k value in equation (3) that affects the steepness of the distance function's curve. As the k-value increases, the curves becomes less steep and therefore raises the edge weight between concepts when their distance falls between the curve's plateaus. Finally, the *Location Edge Weight Adjustment* is used to weaken the adjacent edge weights on concepts that are a location.

Table 3

Key Parameters for the AIF Framework

Parameter	Symbol	Values
Document Segmentation	S	{ article, section }
G_t Cluster Decomposition	C_t	1, 2, .. n
Concept Extraction Threshold	T_c	[0.5 .. 1.0]
Edge Weight Distance k-value	W_k	[0.5e .. 4e]
Location Edge Weight Adjustment	W_l	[0.001 .. 1.0]

9.2 Metrics

For each experiment, a set of metrics is collected and used to assess the experimental run, as listed in Table 4. The metrics provide insights into the size of the Snippet, Concept, and Topic layers of the AIF framework's data store.

Table 4

Metrics for the AIF Framework Experiments

Metric	Symbol	Description
Snippets	N_s	The number of snippets generated.
Concepts	N_c	The number of Concept Nodes generated.
Topics	N_t	The number of Topic clusters generated.
Duration	D	The runtime of the AIF process.
Precision	P	Fraction of relevant topics over generated topics.
Recall	R	Fraction of generated topics over the total relevant topics.

For assessing the utility of the AIF framework for exploratory search, *precision* and *recall* are the most important overall metrics. Precision, computed in (12), assesses the framework's ability to generate topics that are relevant to the information seeker, where P_t is the true positive rate, or the number of *generated* topics that also appear in the answer key. Recall, shown in (13), assesses how completely the topics cover the breadth of the corpus, where N_{tk} is the number of topics in the answer key.

$$P = \frac{P_t}{|N_t|} \tag{12}$$

$$R = \frac{P_t}{|N_{tk}|} \tag{13}$$

9.3 Method

Each experiment uses the same corpus of news articles relating to the *Health Effects and Damages* from Hurricane Irma. This is a 52-document subset of the full Hurricane Irma corpus, because assessing (scoring) the results from the larger, more broad corpus would be very challenging and imprecise due to the scale of concepts and topic clusters produced. A smaller corpus makes the experiment more manageable. Each document in the corpus was analyzed, and if *any part* of the document discussed health or damages, it was included in the experimental corpus.

Through a manual analysis of the corpus, an *answer key* of valid topics was created and used to score the precision and recall metrics for each test. As noted in Chapter 5, the AIF framework's method of creating concept nodes can be significantly improved through Entity Resolution and synonym resolution techniques. Therefore, some of the topics overlap semantically. For instance, the concepts *wind, gust, gale* refer to the same semantic concept, but may appear as separate topics. This is accounted for in the answer key that assesses whether each topic (with respect to precision and recall) are valid. For each experiment run, the appropriate parameters are set prior to execution and the resulting hierarchical graph is analyzed and scored.

9.4 Results and Discussions

All experiments described in this section were executed using an Apple MacBook Pro 2.3GHz Intel Core i7, running macOS 10.13.2. The execution environment was Oracle Java 1.8 using a maximum heap of 1 GB.

9.4.1 Concept extraction threshold. The first experiment compares various Concept Extraction Threshold values using both *section* and *whole-article* snippet segmentation. The threshold value T_c specifies the top percentile of concepts for each snippet (ranked by TF-IDF) that will be added or updated in the

semantic network. A T_c value of 0.3 includes the top 30% of TF-IDF scored concepts. Therefore, the larger the T_c parameter is, the more concepts from each snippet will be added (or have their weights updated).

Table 5 shows that by raising the T_c parameter, the number of concepts, N_c, also increases, as does the processing time, D. However, as the number of concepts increases, the number of topic clusters, N_t decreases. This is attributable to the fact that as the concept nodes increase their degree also increases by virtue of having more connected nodes from the snippets. Additionally, because the number of interconnections per snippet increases, their respective weights are reinforced. This illustrates that as more information is added to the semantic network, stronger communities form in the semantic network's topology.

Table 5

Results of the Concept Extraction Threshold Experiment

Parameters					Counts				Scores	
S	C_t	T_c	W_k	W_l	N_s	N_c	N_t	D	P	R
section	1	0.3	2e	1.0	1823	2391	66	0:05:25	0.97	0.80
section	1	0.5	2e	1.0	1823	4548	50	0:11:08	0.96	0.82
section	1	0.7	2e	1.0	1823	6976	36	0:24:41	1.00	0.80
section	1	0.9	2e	1.0	1823	7786	32	0:35:31	1.00	0.88
section	1	1.0	2e	1.0	1823	8004	29	0:44:50	0.96	0.84
article	1	0.3	2e	1.0	53	1153	19	0:04:03	0.95	0.60
article	1	0.5	2e	1.0	53	2909	17	0:14:25	0.94	0.80
article	1	0.7	2e	1.0	53	5006	17	0:13:52	1.00	0.88
article	1	0.9	2e	1.0	53	6223	17	0:22:36	1.00	0.82
article	1	1.0	2e	1.0	53	7372	16	0:33:58	1.00	0.84

The biggest difference between the section and whole-article segmentation is in the number of snippets (N_s) and topic clusters (N_t) generated. Both have very

high precision scores because with such few topics generated, every one was a relevant topic. Both sets also show high recall scores because multiple answer key topics are covered by a single generated topic. This is inflated because each generated topic cluster is comprised of multiple sub-clusters that are more cohesive and are more aligned to the answer key. However, subjectively, these larger topics are more difficult for information seeker to quickly understand their contents.

9.4.2 Topic cluster decomposition. Because of the inflated recall from the previous experiment, this next experiment focuses on cluster decomposition. Because each topic cluster can contain two or more sub-clusters, if the AIF framework generates a set of sub-clusters for each topic cluster's subgraph, G_t, it is expected that a larger set of smaller, more cohesive topic clusters will emerge.

Table 6 compares the results of a 1-pass and a 2-pass clustering on both section and whole-article segmentation. As expected, there is a drastic increase in N_t for both 2-pass tests. For the section-based segmentation approach, there is also a significant increase in both precision and recall. This intuitively makes sense because second level clustering will produce better clusters. However, for whole-article based segmentation, precision drops and recall remains the same. This leads to the conclusion that when concepts are allowed to generate relationships with other concepts throughout the entire document, the community structures formed are not as cohesive. While this may not be true for all types of source documents, this makes sense when using news articles from the internet. As stated in Chapter 5, news articles are noisy and contain content for irrelevant or tangential information, such as news teasers and advertisements. A noisy document will lead to a noisy, loosely cohesive semantic network.

Table 6

Results of the Cluster Decomposition Experiment

Parameters					Counts				Scores	
S	C_t	T_c	W_k	W_l	N_s	N_c	N_t	D	P	R
section	1	0.5	$2e$	1.0	1823	4548	50	0:11:08	0.96	0.82
section	2	0.5	$2e$	1.0	1823	4548	342	0:28:56	1.00	0.96
article	1	0.5	$2e$	1.0	53	2909	17	0:14:25	0.94	0.80
article	2	0.5	$2e$	1.0	53	4548	133	0:51:51	0.89	0.82

9.4.3 Distance function for edge weights. The next experiment examines the effects of changing the inverse sigmoid curve used in the edge weight distance function described in Chapter 5. The k-value of the function describes the steepness of the curve and extends the width (i.e., distance) between the curve's plateaus. Larger k-values increase the steepness and therefore only close concept distances will have significant weights, while decreasing the k-value will allow more distant concepts to have larger weights.

As shown in Table 7, a larger k-value significantly raised the N_t count and slightly raised the precision and recall. This illustrates that for this corpus, the topological structures in the semantic network are slightly improved by restricting the significant inter-concept relationships. There was no effect on whole-article segmentation scores. Distances were not a discriminator for whole-document segmentation because the distance function curve's plateau is reached well before the end of the document. With documents that greatly vary in length, a distance metric spanning 8+ sentences may be meaningless.

Table 7

Results of the Edge Weight Distance Function Experiment

Parameters					Counts				Scores	
S	C_t	T_c	W_k	W_l	N_s	N_c	N_t	D	P	R
section	1	0.5	$4e$	1.0	1823	6213	76	0:28:35	1.00	0.86
section	1	0.5	$2e$	1.0	1823	4548	50	0:11:08	0.96	0.82
section	1	0.5	$1.5e$	1.0	1823	4549	53	0:15:00	1.00	0.86
section	1	0.5	e	1.0	1823	4548	50	0:14:51	0.96	0.82
section	1	0.5	$0.5e$	1.0	1823	4548	50	0:14:10	0.96	0.82

9.4.4 Location weight adjustment. The final experiment examines the effect of reducing the weights of edges that are adjacent to location concepts. The rationale for this experiment is that based on reporting styles of some news articles, they are often written from the perspective of a single city or town, and that location name is often repeated in the article. Therefore some communities in the semantic network may become biased toward those locations. By providing a means to reduce the weights of edges adjacent to location nodes, the AIF framework can alter the community structures away from these locations for a more semantically meaningful network.

As shown in Table 8, there is no significant change in precision, but the recall score rose slightly. This slight improvement makes intuitive sense because not all topic clusters were location-centric. But even this slight improvement may help the information seeker to discover relevant topics.

9.4.5 Experiment summary. Based on these experiments, the optimal base configuration for the AIF framework is section segmentation with 2-pass clustering. Depending on the corpus, an optional location edge weight adjustment may also improve the results. The most important factor for high precision and

Table 8

Results of the Location Edge Weight Adjustment Experiment

Parameters					Counts				Scores	
S	C_t	T_c	W_k	W_l	N_s	N_c	N_t	D	P	R
section	1	0.5	$2e$	1.0	1823	4548	50	0:11:08	0.96	0.82
section	1	0.5	$2e$	0.1	1823	4548	53	0:10:59	0.94	0.84
section	1	0.5	$2e$	0.01	1823	4548	58	0:10:53	0.97	0.86
section	1	0.5	$2e$	0.001	1823	4548	57	0:10:22	0.97	0.86

recall is having smaller, more cohesive topic clusters. In these tests, automatically decomposing all topic subgraphs into their sub-clusters was a critical step in producing better results for exploratory search. However, the downside to this is a very rapid growth in the number of topics, which can be detrimental to the information seeker. An important area for future research will be the intelligent decomposition of select topics, rather than blindly decomposing all topics in a second-level pass. Doing so will identify cohesive topics while controlling scale.

Chapter 10

Ongoing Research

This thesis provides a framework for continued research in Automated Information Foraging. Although the AIF prototype demonstrates promise in satisfying the goals outlined in Chapter 1, there are several areas that can benefit from ongoing research to enhance upon its usability and capability. This chapter outlines several areas that, if realized, could provide significant improvement.

10.1 Documents and Snippets

10.1.1 Beyond unstructured text. This thesis focused on the use of unstructured text documents comprising the information corpus. Information seekers, however, rely upon multi-modal information including imagery, video, audio, and databases. By incorporating non-textual information into the Snippets and Concepts layers would expand the richness of information available for analysis.

10.2 Semantic Network Construction

10.2.1 Entity resolution. Upon analysis of the semantic network constructed by the AIF framework, several concept nodes are sometimes created for the same real-world entity because of the variations in their naming. For example, President Donald Trump appears multiple times in the semantic network as: "President Donald Trump," "President Trump," "Donald Trump," and simply "Trump". Although these concepts are linked in the network, a more accurate semantic network would employ *entity resolution* techniques to identify these names refer to the same entity, then fuse them into a single concept node. By resolving these entities, the semantic network topology would be more accurate and significantly alter the structures (i.e., topic clusters).

10.2.2 Term synonyms. Similarly, identifying if concept nodes are synonyms and fusing or linking these concepts together may yield a more accurate semantic network. Since the AIF framework is already using the WordNet database for term stemming, its use could be expanded for synonym detection among nodes. Once challenge is that if concept nodes use n-grams (as opposed to unigrams), one or more unique words in the n-gram may be a synonym of another n-gram or n-gram constituent. An appropriate means for resolving this would require further research, but may yield significant improvements to the network's construction.

10.3 Topic Generation

10.3.1 Topic summarization. Perhaps the most useful improvement upon the current AIF framework is a more easily understandable and readable summary of the topic's contents. The current method of using representative keywords from the semantic network subgraph for that topic, G_t, provides an accurate proxy for the topic. This set of distinct n-grams, however, can be difficult to understand quickly. A better method may be to provide a textual summary of the topic's contents. But this textual summary should provide a multi-document summary of the snippets linked to the topic, highlighting the keyword in that summary. A textual summary would be easier to read and understand in a rapid manner. Recent research in multi-document text summarization would be an initial focus area for improving presentation of the topics to the user.

In addition to text as a proxy for the topic's contents, visual cues such as images or word clouds can also be presented alongside the textual summary. These visual cues will also align with the notion of providing an *information scent* to the seeker, as described in Pirolli's *Information Foraging Theory* research[7]. Information Scent assists the seeker is quickly recognizing potentially relevant or lucrative information patches for further exploration.

10.3.2 Auto-decomposition of topic clusters.

Another research area that will significantly improve the AIF framework is auto-decomposition of individual topic clusters. The framework identifies topic clusters hierarchically, where any cluster may contain zero or more sub-clusters. When examining the clusters generated after one iteration of the clustering algorithm, the resulting cluster set contains clusters with zero or more sub-clusters. Having large clusters with multiple sub-clusters can cause a lack of internal cohesion within that cluster, while topic clusters that cannot be decomposed are highly cohesive. The challenge for the AIF framework lies in the means to detect when a topic cluster should be automatically decomposed into constituent sub-clusters by the framework. Research should identify the metrics that can be used to detect when decomposition is warranted, and strive to maintain a balance between the number of topic clusters and their internal semantic cohesion.

10.3.3 Overlapping topic clusters.

The AIF framework currently uses a Max-Modularity Graph Clustering algorithm to detect community structures (i.e., topic clusters) in the semantic network. This approach assigns each node in the semantic network to exactly one cluster. Fortunato identifies several promising techniques for overlapping community detection in graphs[31], such as the Clique Percolation Method[39]. Having concept nodes that span one or more clusters may provide clusters that are more cohesive.

10.4 User Feedback in Relevance Assessment

Ultimately, the AIF framework will be used to assist the information seeker in exploring the information domain. During this exploration, the information seeker will select topics for exploration and snippets for analysis. If the snippets are relevant or lucrative to a hypothesis, the seeker will save the snippets for further detailed analysis or to be used as evidence to refute or substantiate a hypothesis.

Each of these *foraging actions* models important information about the seeker's notion of relevance and can further provide an input signal into the spreading activation used in relevance assessment. This feedback would create a more comprehensive view of what relevance means to the information seeker by incorporating topological structure of the semantic network, a priori knowledge via refinement queries, and human-in-the-loop feedback from information exploration.

10.5 Temporal Differences in the Information Space

By using document metadata (e.g., document publication date), the AIF framework can construct multiple semantic networks where each network models only those documents published within specific dates. This analysis can assist the information seeker in identifying temporal trends in the information space, and how they evolve and relate over time. Some topics may have limited lifespans while other may span larger time frames. The resulting temporal trends can be a valuable tool in understanding the information domain with limited information or in hindsight.

Chapter 11

Conclusions

Chapter 1 outlines a set of goals for an automated information foraging (AIF) framework that facilitates both information exploration and exploitation while mitigating potential biases. Additionally, there is a set of AIF technical goals for accomplishing these information retrieval goals. This chapter revisits these goals and identifies successes and areas for improvement.

11.1 Exploration of the Information

The first goal of the AIF framework is that when given a document corpus covering a specific domain, it analyzes that corpus and provides a presentation of the full breadth of topics that comprise that domain. As shown through the example use case in Chapter 8, the AIF framework can successfully provide a topic list that spans the breadth of the corpus. Rather than generating a set of documents that closely match a search query, as is done in search engines (potentially biasing the results), this approach provides users with a set of topics that can be used to expand their search, potentially identifying topics that they would not have known to search a priori. This kind of information discovery allows the information seeker to explore multiple, alternative hypotheses where it will be apparent what topics are consonant or dissonant with their extant beliefs or biases.

The underlying assumption behind the AIF's ability to realize this goal is that the corpus contains this wide breadth of unbiased documents. The AIF framework does provide a mechanism to accomplish this using GDELT as a news article index.

While the results show the AIF framework's approach to realizing this goal a success, there are two key areas for improvement. The first area is the mechanism used to present the topic contents to the information seeker. The current approach

identifies representative keywords (i.e., n-grams) from the topic's concepts. While this approach is similar to competing topic modeling approaches, such as Latent Dirichlet Allocation (LDA), a series of seemingly unordered terms is difficult to read quickly, and requires mental analysis to potentially understand their relationship to each other. Multi-document text summarization is a likely means to improve this feature.

The second area for improvement is automatic topic cluster decomposition. The current AIF method for clustering topics considers the modularity assessment for the entire semantic network to create a whole-graph set of topic clusters. This generates a Topics Layer with some very large clusters (with sub-clusters) while other topic clusters may be small (and fully decomposed). Unless a balance is achieved between the number of topics and their semantic cohesion and relative size (i.e., number of constituent concept nodes), efficient information foraging is degraded. Over-decomposition will generate a large number of topics, while under-decomposition will result in topics that are not semantically cohesive. After the first iteration of this algorithm, AIF should then decide on a cluster-by-cluster basis whether it should be further decomposed into sub-clusters, rather than a universal iterative decomposition of all topic clusters.

Despite these areas for improvement, the AIF framework provides the means to generate a wide set of topics that can identify both widely reported news events as well as news events and stories only reported by one article. The resulting topics produced by the AIF allow the information seeker to explore the full breadth of topics in the corpus, potentially discovering topics that are not known a priori.

11.2 Exploitation of the Information

One of the key tenets of the Pirolli Information Foraging Loop in the Sensemaking model, is the need for information seekers to find raw information, then filter it down to a relevant set, then to a smaller lucrative information set. The latter two sets are stored for analysis and for use as evidence to substantiate or refute hypotheses. This process of analysis and filtering information is the exploitation of the information. The AIF framework facilitates information exploitation through two key ways. The first method is the Relevance Analysis approach that uses a refinement query to filter the topics down to a smaller highly-relevant set. The terms for this refinement query are discovered through the information seeker's exploration of the information topics and their associated document snippets. The second method is the AIF framework's ability to provide full traceability between the Topics, Concepts, Snippets, and Documents. The traceability facilitates analysis of the source materials in the snippets that spawned the topics, which can be used as evidence against hypotheses and for continued analysis.

11.3 Analysis of the Semantic Network Topology

There are several advantages to modeling the information space via a semantic network. Using community detection / graph clustering techniques, topics are extracted from the underlying structures and communities of concept nodes found in the corpus. This is a form of topic modeling that can allow for a variable number of topic clusters based on concept co-occurrence relationships in the corpus. An additional advantage to using this approach for topic modeling is that even small communities resulting from a single news article mention can be identified as its own topic, which may not always be possible with other topic modeling approaches.

Using spreading activation, relevance scores for each topic are generated

based on a seed query, providing both a filtering and ranking of topics. This approach also relies on the topology of the semantic network and the relative weights between concept nodes to identify potential relevance. Inter-concept relationships as well as inter-topic cluster relationships facilitate exploration of the information via navigation through network adjacencies.

By analyzing the topology of the semantic network subgraph that comprises a single topic, the AIF framework can further facilitate depth-wise exploration of the information. This can provide more detailed analysis of topics and also be used to identify sub-topics. Additionally, significant topic concept terms can be identified as representative *keywords* for that topic by scoring the betweenness centrality of the subgraph nodes representing the topic's concepts.

The hierarchical graph-based schema provides topic-concept-snippet-document traceability, allowing the information seeker to quickly access the documents and document snippets relating to topics.

One significant area for improvement is in the construction of the semantic network. The current AIF framework does not account for concept node terms that can be represented by more than one n-gram. The semantic network can benefit from approaches in Entity Resolution and synonym analytics. By resolving concepts that represent the same or very similar entities, duplicative concept nodes will be reduced, which should in turn cause more cohesive topic clusters to form.

11.4 Summary

The automated information foraging (AIF) framework described in this thesis provides information seekers with the ability to explore the full breadth of information contained within a document corpus. By identifying the natural language terms and forming a semantic network based on the co-occurrence of those concept terms within document snippets, the AIF framework extracts the clusters of

information and presents them to the information seeker, allowing for the discovery of information topics that may have been unknown or dissonant to the seeker. These capabilities help overcome inherent cognitive biases over the information as well as potential biases resulting from the use of search engines. Finally, the framework facilitates an information seeker in exploiting the information by filtering it down to a lucrative set of information snippets that can be used to form, substantiate, or refute hypotheses, an essential requirement for sensemaking.

The AIF prototype provides a starting point for research into Information Foraging techniques. As discussed in Chapter 10, the development of this framework and the subsequent analysis of its products has identified numerous topics for continued research into both the computer science and cognitive science needs to realize the Automated Information Foraging vision.

References

[1] Laney, D. 2001. *Application Delivery Strategies*. Meta Group. Accessed on March 25, 2018 from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[2] Heuer, R.J., Jr. 1999. *Psychology of Intelligence Analysis.* Center for the Study of Intelligence, Central Intelligence Agency, Washington, DC.

[3] Pirolli, P., Card, S. 2005. *The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis.* (2005) In Proceedings of the International Conference on Intelligence Analysis (Vol 5).

[4] Mehlhorn, K. et al. 2015. *Unpacking the Exploration-Exploitation Tradeoff: A Synthesis of Human and Animal Literatures*. Decision 2, 3 (2015), 191215.

[5] White, R., Marchionini, G., Muresan, G. 2008. *Evaluating exploratory search systems.* Information Processing & Management 44, 2 (2008), 433436.

[6] Kayhan, V. O. 2015. *Confirmation Bias: Roles of Search Engines and Search Contexts* Thirty Sixth International Conference on Information Systems, Fort Worth 2015.

[7] Pirolli, P. 2007. Information Foraging Theory. *Information Foraging Theory* (March 2007), 3-29.

[8] Jonas, E., Schulz-Hardt, S., Frey, D., and Thelen, N. 2001. *Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information.* Journal of Personality and Social Psychology 80, 4 (2001), 557571.

[9] Genc, Y. 2014. *Exploratory search with semantic transformations using collaborative knowledge bases.* Proceedings of the 7th ACM international conference on Web search and data mining - WSDM 14 (2014).

[10] Yogev, S. 2014. *Exploratory search interfaces*. Proceedings of the companion publication of the 19th international conference on Intelligent User Interfaces. IUI Companion 14 (2014).

[11] Shapiro, S. 1992. Encyclopedia of artificial intelligence, New York: Wiley.

[12] Arora, P., Jones, G. 2017. *Identifying Useful and Important Information within Retrieved Documents.* Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval - CHIIR 17 (2017).

[13] Newman, M. E. J., Girvan, M. 2004. *Finding and evaluating community structure in networks* Physical Review E 69 (February 2004).

[14] Brandes, U., 2001. *A Faster Algorithm for Betweenness Centrality* Journal of Mathematical Sociology 25 (2001), 163177.

[15] Crestani, F., 1997. *Application of spreading activation techniques in information retrieval* Artificial Intelligence Review 11, 6 (1997), Kluwer Academic Publishers, Norwell, MA, pp. 453-482.

[16] The GDELT Project. Retrieved from https://www.gdeltproject.org/ on January 25, 2018.

[17] Google Cloud BigQuery. Retrieved from https://cloud.google.com/bigquery/ on March 22, 2018.

[18] Kohlschütter, C. *BoilerPipe*. Retrieved from https://code.google.com/archive/p/boilerpipe/ on September 23, 2017.

[19] Kohlschütter, C., Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. Proceedings of the third ACM international conference on Web search and data mining - WSDM 10 (2010). DOI:http://dx.doi.org/10.1145/1718487.1718542

[20] The Apache Software Foundation. *Apache Lucene*. Retrieved from http://lucene.apache.org on September 29, 2017.

[21] The Apache Software Foundation. *Apache OpenNLP*. Retrieved from http://opennlp.apache.org on September 29, 2017.

[22] Taylor, A., Marcus, M., and Santorini, B. 2003. *The Penn Treebank: An Overview*. Treebanks Text, Speech and Language Technology (2003), 522.

[23] Copestake, A. 2004. Retrieved from https://www.cl.cam.ac.uk/teaching/2002/NatLangProc/revised.pdf on January 20, 2018.

[24] Subhashini, R. and Kumar, V. J. S. 2011. *Optimization of Internet Search based on Noun Phrases and Clustering Techniques*. International Journal of Computer Applications 20, 2 (2011), 4954.

[25] Miller, G. 1995. *WordNet: A Lexical Database for English.* Communications of the ACM Vol. 38, No. 11:39-41.

[26] Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

[27] Walenz, B., Barton, G., and Didion, J. *JWNL: Java WordNet Library*. Retrieved from: https://sourceforge.net/jwordnet/wiki/Home/ on September 29, 2017.

[28] Ramakrishnan, R. and Gehrke, J. 2011. *Database Management Systems*, Boston: McGraw-Hill.

[29] Manning, C., Raghavan, P., and Schtze, H. 2009. *Introduction to information retrieval* New York: Cambridge University Press.

[30] Blei, D., Ng, A., and Jordan, M. 2003. *Latent Dirichlet Allocation.* Lafferty, J., ed. Journal of Machine Learning Research 3 (January 2003), 993-1022.

[31] Fortunato, S. 2010. *Community detection in graphs.* Physics Reports 486, 3-5 (2010), 75-174.

[32] Newman, M. E. J. 2004. *Analysis of weighted networks.* Physical Review E 70 (2004).

[33] Noack, A., Rotta, R. 2009. *Multi-level Algorithms for Modularity Clustering* Experimental Algorithms Lecture Notes in Computer Science (2009), 257-268.

[34] Brandes, U., 2001. *A Faster Algorithm for Betweenness Centrality* Journal of Mathematical Sociology 25 (2001), 163-177.

[35] Rossi, F., and Villa-Vialaneix, N. 2011. *Representation d'un grand rseau partir d'une classification hirarchique de ses sommets.* Journal de la Socit Franaise de Statistique, volume 152, number 3, pages 34-65, December 2011.

[36] Rossi, F. *Graph Clustering* Retrieved from: http://apiacoa.org/research/software/graph/index.en.html on October 25, 2017.

[37] Clémençon, S., De Arazoza, H., Rossi, F., Tran, V. 2011. *Hierarchical clustering for graph visualization* (2011) In Proceedings of XVIIIth European Symposium on Artificial Neural Networks (ESANN 2011), pages 227-232, Bruges (Belgium), 2011.

[38] Crestani, F., 1997. *Application of spreading activation techniques in information retrieval* Artificial Intelligence Review 11, 6 (1997), Kluwer Academic Publishers, Norwell, MA, pp. 453-482.

[39] Dernyi, I., Palla, G., Vicsek, T. 2005. *Clique Percolation in Random Networks.* Physical Review Letters 94, 16 (2005).

[40] Marchionini, G., Shneiderman, B. (1988). *Finding facts vs. browsing knowledge in hypertext systems.* IEEEComputer, 21(1): 70-79.

[41] Shneiderman, B., Plaisant, C. 2005. *Designing the User Interface 4th Ed.* Person/Addison-Wesley.

[42] Pirolli, P. 2006. *Analysis of the Task Environment of Sense Making.* SIGIR'06 Workshop, August 10, 2006, Seattle, WA, USA.

[43] White, R., Kules, et al. 2006. *Supporting exploratory search: Introduction.* Communications of the ACM, 49(4): 36-39.

[44] Lafferty, J., Blei, D. 2009. *Topic Models.* Text Mining Classification Clustering Applications. 10. 71-93.

[45] Steyvers, M., Griffiths, T. 2007. *Probabilistic Topic Models* In Landauer, T., McNamara, D., Dennis, S., et al. *Handbook of Latent Semantic Analysis.* Psychology Press. ISBN 978-0-8058-5418-3.

[46] Tang, J., Meng, Z., et al. 2014. *Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis*. Proceedings of the 31st International Conference on Machine Learning, PMLR 32(1):190-198, 2014.

[47] Hong, L., Davison, B. 2010. *Empirical study of topic modeling in Twitter.* In Proceedings of the First Workshop on Social Media Analytics, SOMA 10, pp. 8088, New York, NY, USA, 2010. ACM. ISBN 978-1- 4503-0217-3.

[48] Mimno, D., McCallum, A. 2007. *Organizing the oca: learning faceted subjects from a library of digital books.* In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 376385. ACM, 2007.

[49] Preece, S., 1981. *A spreading activation network model for information retrieval.* PhD thesis, University of Illinois, Urbana-Champaign, 1981.

[50] Wasserman, S., Faust, K. 1994. *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge, UK, 1994.

[51] Bavelas, A. 1950. *Communication patterns in task-oriented groups.* J. Acoust. Soc. Am, 22(6):725730, 1950.

[52] Newman. M. 2001. *Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality.* Phys. Rev. E, 64(1):016132, 2001.

Appendix A

Abbreviations and Symbols

The abbreviations and symbols used throughout the document are listed in this appendix.

Symbol	Description
AIF	Automated Information Foraging
AIR	Associative Information Retrieval
API	Application Programming Interface
B_c	Betweenness Centrality
GB	Gigabytes
GDELT	Global Database of Events, Language, and Tone
G_t	Semantic Network subgraph comprising a single Topic
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
NER	Named Entity Recognition
NLP	Natural Language Processing
SA	Spreading Activation
SQL	Structured Query Language
TF-IDF	Term Frequency - Inverse Document Frequency
URL	Uniform Resource Locator
XML	Extensible Markup Language

Appendix B

Data Dictionary

The meta-data attributes for all nodes in the hierarchical graph-based schema are listed in this appendix. The nodes in the hierarchical graph are assigned a class designating the layer in which it resides.

Layer(s)	Key	Description or Value
All	class	Layer: {document, snippet, concept, topic}
All	vertexId	Unique Identifier for this node.
document	text.raw	The original downloaded content.
document,snippet	url	The URL from which the document was downloaded.
document,snippet	text	The filtered, cleaned text.
concept	name	The concept term's name.
concept	type	ngram, location, person, organization
concept	partition	The topic-cluster identifier this concept node belongs to.
concept	bc	The betweenness centrality score within its topic cluster subgraph.
concept	tfidf.max	The maximum TF-IDF score for this concept term across its snippets.
concept	activation	The current activation score for the relevance analysis
topic	partitionId	The cluster identifier this topic.
topic	keywords	List of keywords for this topic.
topic	completeness	The completeness metric for this topic's G_t
topic	degree	The degree of the topic node in the Topics Layer