Where Should We Protect? Identify Potential Targets of Terroristic Attack Plan via Crowdsourced Text Analysis

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

The increasing volume of text datasets is challenging the cognitive capabilities of expert analysts to produce meaningful insights. Large-scale distributed agents like machine learning algorithms and crowd workers present new opportunities to make sense of big data. However, we must first overcome the challenge of modeling and guiding the overall process so that many distributed agents can meaningfully contribute to suitable components. Inspired by the sensemaking loop, collaboration models, and investigation techniques used in Intelligence Analysis community, we propose a pipeline to better enable collaboration among expert analysts, crowds, and algorithms. We modularize and clarify the components in the sensemaking loop so that they are connected via clearly defined inputs and outputs to pass intermediate analysis results along the pipeline, and can be assigned to different agents with appropriate techniques. We instantiate the pipeline with a location-based investigation strategies and experimented with crowd workers on Amazon Mechanical Turk. Our results show that the pipeline can successfully guide crowd workers to contribute meaningful insights that are helpful to solve complicated sensemaking challenges. This allows us to imagine broader possibilities for how each component could be executed: with individual experts, crowds, or algorithms, as well as new combinations of these, where each is best suited.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Authors' choice; of terms; separated; by semicolons; include commas, within terms only; required.

INTRODUCTION

Information explosion from modern technology has spurred increased interest in sensemaking to help people gain insights and suggest effective actions from the big data. "Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically" [13] by modern computer technology alone, understanding and acting on the information to solve real-world problems requires human computation as well. Making sense of a large amount of text data requires understanding natural languages, which is considered AI-hard [30]. Observations suggest evidence marshaling and synthesis are particularly difficult [29]. To get the "big picture" by looking at many pages of text, the analyst relies heavily on memory to connect the dots. Crowdsourcing and algorithms present new opportunities for large-scale sensemaking, but we must first understand how sensemaking work can be modularized to allow powerful and diverse techniques to be used where they can contribute best. The existing models and theories of sensemaking process have been applied to various domains using different types of analysis agents including individual and groups of experts, crowd workers, and machine learning algorithms. Most of the techniques focus on some particular parts of sensemaking process or provide ideal inputs to non-expert agents, which still requires a considerable amount of work from expert analysts. We aim to release this assumption and execute the whole sensemaking loop with non-expert agents, by modularizing different components of the process with more unified inputs and clearly defined outputs that allows combinatorial and flexible usage of them and enabling multiple paths with a mechanism to decide where to proceed next. Such a pipeline can compare or combine different agents on each component to make full use of the complementary strength of human and computation. Having more specific goals for each component enables backtracking and re-evaluation of the analysis progress. Additionally, with multiple agents involved in the analysis process, we can overcome stereotypes, bias, and group-think by re-examining intermediate and overall results. With this motivation, we aim to answer the following research questions in this study: What are the information needs (inputs) and intermediate results (outputs) at different stages of text analysis? How well can crowd workers perform in analytical tasks at different level: finding relevant documents that are helpful for the investigation, extracting information pieces from relevant documents that are useful for later information synthesizing, organizing information pieces in a meaningful way that help generate in-depth hypotheses, infer from available evidence information to draw plausible hypotheses, and transform winning hypotheses into a well-structured presentation?

To answer these questions, our pipeline further modularizes the stages in the Representation Construction Model, clarify

CHI'18, TBD

ISBN 123-4567-24-567/08/06…$15.00

DOI: http://dx.doi.org/10.475/123_4

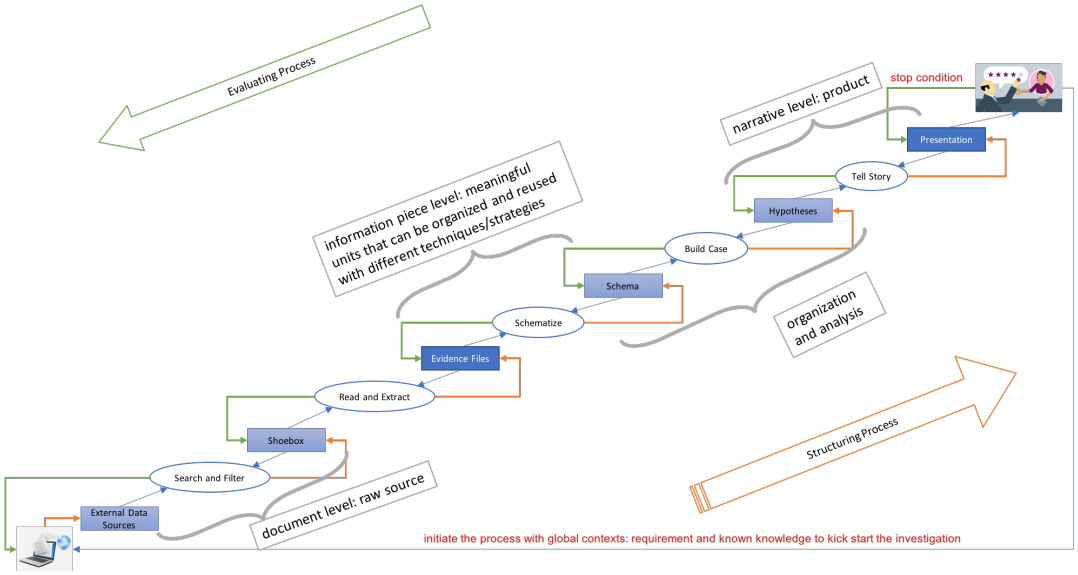

Figure 1. Crowdsourced Text Analysis Pipeline

the requirements (inputs and outputs), such that they can be assigned to appropriate agents, including individuals, crowds, or automated techniques, and completed with a variety of methods. As a proof of concept, we choose crowdsourcing to demonstrate the utility of the pipeline in this study, using a dataset from intelligence analysis training material. We found that the crowds can provide meaningful insights and analysis at each step, offer multiple perspectives on the same given input, and refine previous workers' result to better follow the instructions. This provides evidence that the more formalized model enables new possibilities for modularized development.

Our contributions are as follows. First, we disassemble and tailor the process in the context of intelligence analysis on textual evidence documents. More specifically, we define each component with its functionality, input, and output as well as the relationship among them. Second, we conducted an empirical case study demonstrating the feasibility of the pipeline. Third, we suggest that the pipeline can be used to open the sensemaking process up for more researchers to contribute.

There are also many benefits as by-products of this work. Applying the pipeline is less demanding in terms of data confidentiality since each component loop only takes part of the data. Furthermore, executing a component requires less expertise and opens the tasks up to more possibilities including novice crowds or algorithms[20] . We can compare or combine different agents on each component to make full use of the complementary strength of human and computation. Having more specific goals for each component enables backtracking and re-evaluation of the analysis progress. Additionally, with multiple agents involved in the analysis process, we can overcome stereotypes, bias, and groupthink by re-examining intermediate and overall results.

RELATED WORK

Our proposed pipeline expand on prior research efforts on sensemaking models, the human-debugging paradigm, and application of those theories in supporting expert data analysts in both visual analytics and crowdsourcing community.

Sensemaking Models and Theories

The notion of sensemaking in the field of human-computer interaction (HCI) was framed in the early 1990s [23] as "the process of forming and working with meaningful representations in order to facilitate insight and subsequent intelligent action". The sensemaking process is complicated in that it iterates on multiple intertwined stages, employs combinatorial and individually-variable reasoning heuristics, and differs according to specific goals of problem-solving.

Pirolli and Card [21] proposed the Representation Construction Model (the sensemaking loop) as two iterative processes: information foraging and synthesis. Each process contains smaller loops at different levels of data granularity. Expert analysts generate a theory from raw data by going through a bottom-up process, and trigger re-evaluation on previous intermediate analysis results or re-analysis on previous data resources in a top-down process. The Data-Frame Theory [17] developed in the macrocognition approach of psychology, were also adapted [19] [25] [1] and applied [9][2] in HCI research. The Data-Frame Theory focuses on iteratively developing meaningful representations (frames) that explain external reality (data). Building on the ample literature on frames and similar concepts, Klein et al. synthesized the concepts as "a structure for accounting for the data and guiding the search for more data".

However, the models are usually general and do not have clear boundaries between each component, which will lead to vague or varying interpretations and applications in real-world problems. Empirical user studies [6] have revealed the

importance of conceptual connections and domain knowledge. Applying these theories to tackle real-world problems requires experts to decide the data formats and processing methods to use in each step case-by-case, depending on the broad definitions given in the theories, their prior experience, the data under analysis, the goal of investigation and other situational constraints.

Our pipeline modularizes the sensemaking loop and leverages expert guidance to control the flow of executions of different steps by crowds and/or algorithms. For each step, we apply data-frame theory to monitor the execution status, and suggest if the current step is ready to pass to the next step, or waiting for better input from the previous step, or requires improvement in task design. Experts can give feedback on final results to trigger re-execution of some steps for improvement. If the pipeline observes repeated execution in a particular sub-loop, experts will be notified and give more guidance. We will introduce more related work on expert-guided crowd work and differentiate our focus in later sections.

Human Participation in Computational Process

Quinn et al. examined the focus and distinctions between Human Computation, Crowdsourcing, Social Computing and Collective Intelligence [22], offering a nice taxonomy of online human participation in the computational process. Our study sits in the intersection of all the four fields. Both humans and computers bring strength in information processing, the latter mainly assisting the former with superior working memory and lower-bias environment [8]. Crouser et al. reviewed and identified the patterns in existing affordances representative of the study of human-computer collaborative problem-solving, understanding which forms the basis of a common framework for this domain of problems. Visual analytics community has seen a shift of âĂŸhuman in the loopâĂŹ philosophy for visual analytics to a âĂŸhuman is the loopâĂŹ viewpoint [10], which supports existing interactive process in situ. In the theories presented in the previous section, researchers should first decide if a problem would benefit from a collaborative technique, then which tasks to delegate to which party, and when. Given these two answers, different systems can then be compared to solve the same problem. Modeling human cognition and sensemaking is essential for developing computer-aided information processing and knowledge visualizations to address today's complex problems.

Human involvement is also important to improve the performance of existing computational AI systems. Parikh et al. proposed the human-debugging paradigm [20] to explore how the human vision process could be decomposed into a pipeline in order to identify computational bottlenecks as well as opportunities where new automated techniques could make the most impact. One of the challenges they have identified is defining inputs and outputs for each component in complicated big problems, that should be equivalent to both human and machine. Following this paradigm, the following research questions need to be answered in the context of text analysis: What are the information needs (inputs) and intermediate results (outputs, also serve as inputs for the following stages) at different stages of text analysis? What are the strategies to decide the sequence of analysis tasks? When should the analysis be conducted bottom-up and when top-down?

Sensemaking in Visual Analytics

Visual analytics community straddles both foraging and sensemaking loops in its efforts to assist both individual and groups of users in investigating and hypothesizing on complex and dynamically changing information. Individual analysts can receive a rich range of assistance from evidence to hypotheses and with multiple data types with visualization tools like Jigsaw [26]. Multiple visualizations of reports and the entities within them, as well as the connections that exist in between, allows people to interact with the views and explore possible new avenues of examination. Integration of visualization with shared accessibility and discussion enhance collaborative complex problem-solving in pairs and small groups. Timelines are often used in such visual analysis tools to represent temporal relationships within the data being investigated [4]. Bier et al. identified key aspects of the design, featuring flexible, shared information structure and visualization among experts, and a notification system that finds entities of mutual interest to multiple analysts. However, synchronous collaboration among small groups is pretty much restricted to information seeking and organization up to entity-level analysis.

Prior research has also studied collaborative sensemaking and have identified several suggestions for designing collaborative visual analytics tools. Bradel et al. explored how a large, high-resolution display as a workspace in a co-located setting helps to externalize information to the display in meaningful schemas during pairwise collaboration to make sense of large text dataset [5], and addresses problems like common ground, communication, hand-offs, coordination and attention shifting in teamwork, which is shared among most, if not all collaborative work. Dispatching simpler sensemaking tasks to multiple agents may help to solve some of the problems of attention shifting and mental model interfering. By expert-driven task distribution and aggregation, the coordination can be guided thus more efficient. The expert guidance also naturally offers multiple views on the problem, given the rich pool of strategies and methodologies developed in different domains and individual experience of experts.

Crowdsourced Sensemaking

The crowdsourcing community has seen success in integrating intelligence power of bigger crowds in complex problem-solving. Starting from low-level data processing tasks like image labeling [14], named entity extraction and merging [28], the crowds have accomplished increasingly complicated and interdependent tasks like article editing as word processor [3], and even writing short stories [15]. Crowdsourcing is different from traditional teamwork in that it aims to collect and aggregate distributed and asynchronous work among strangers. Researchers have strive to address the tension between local micro tasks and the global view of the whole dataset [12] [27], leveraging efforts of previous users [11], balancing structure, flexibility and expert guidance [16] [7], and so on. Nonetheless, asynchronous and higher level analysis is still bottlenecked by communicating insights and reasoning even among

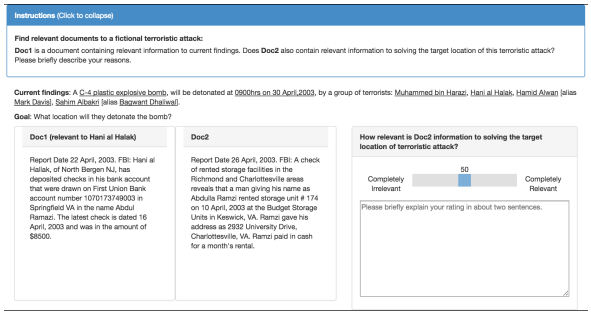

Figure 2. Step 1 Task Interface: Rate document relevance by given global context and a relevant document

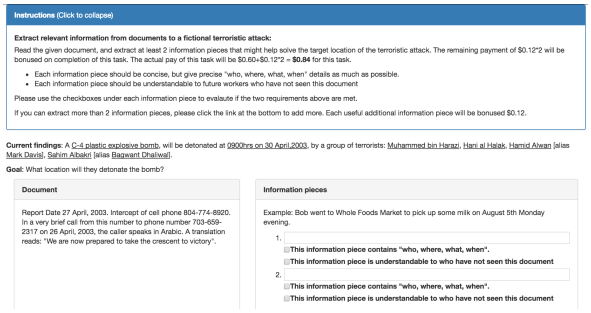

Figure 3. Step 2 Task Interface: Extract important information pieces from relevant documents

few analysts. As is acknowledged in their work, the crowdsourced approach is most valuable where experts generate a lot of valuable information that is unstructured and redundant. Soylent [3] introduce Find-Fix-Verify crowd programming pattern to increase crowd's quality of work; TurKit applied crash-and-rerun programming model [18] to propose iterative tasks on MTurk; IdeaGens [7] explored guiding complex collaborative tasks like brainstorming with Crowds by dividing the crowd into ideation and synthesis tasks. In this research project, we take a step further to explore how crowdsourced approach can be applied to raw textual documents collected on the field that is not pre-processed by expert analysts, and guide the workflow by contextualized feedback.

CROWDSOURCED TEXT ANALYSIS PIPELINE

We developed a pipeline that transforms raw external documents into a presentation of analysis result with the five steps proposed in the sensemaking loops. Each step is modularized and clarified to assign specific tasks to the crowd. Multiple crowd workers asynchronously collaborate in each step to analyze the output from the previous step. Their higher level insights are then passed to the crowd workers in the next step.

When customers (the government or other agencies) initiate an investigation process, they usually have collected some raw data (external data sources for the pipeline), the current findings (a global context for the whole pipeline), and have an investigation purpose (investigation strategy) in mind. Our pipeline shares the global contexts with each step, and feed the external data sources to the first step to start the text analysis.

Step 1: Organize Raw Documents by Relevance

The first step of our sensemaking pipeline takes *Data Input* from external data sources: all collected documents that haven't been investigated yet. The documents are rated by their relevance to the global context (investigation purpose); those rated above a certain threshold are considered *Data Output*: documents pertinent to the investigation purpose. Step 1 remove useless information (noise) at a document level; later analysis will situate in this focused information space.

Break down Step Data Input and create individual tasks

In order to make better use of the crowd intelligence, we first preprocessed the documents algorithmically to find "directly relevant" documents that explicitly mention key elements from the global context. The remaining documents might still be relevant but not obviously. For example, if the investigation purpose is to find potential target for a terroristic attack, a document about "Taliban" activity is not retrieved as "directly relevant" but should be considered relevant. Another possibility is "directly relevant" documents might shed light on some important clues that make other documents relevant. For example, the global context contains a terrorist named "Hamid Alwan". One directly relevant documents mention that Hamid Alwan uses a fake name "Mark Davis". Documents about "Mark Davis" is not "directly relevant" but become relevant because of this document. Following this logic, we pair each one of the remaining documents with one "directly relevant" document that has the most number of overlapping entities with it and ask the crowds to rate its relevance with that extra context. Each pair of documents is presented as one HIT (Human Intelligence Task, the unit of tasks on Amazon Mechanical Turk), and is assigned to three crowd workers.

Aggregate responses of individual tasks into Step Data Output

Each of the remaining document gets three ratings from three different crowd workers. Given a threshold T, each rating can be categorized as "relevant" or "irrelevant". Documents with 2 or 3 "relevant" (majority vote) together with the "directly relevant" documents will be the Data Output of Step 1.

Step 2: Extract Information Pieces from Documents

The second step takes the Data Output of Step 1 as *Data Input*, and extract the core information pieces (mini-event with who, what, where, when). The information pieces should be important to the investigation purpose, and will be the *Data Output* of Step 2. Step 2 remove noise within relevant documents; analysts do not have to read through all the words for a piece of information.

Break down Step Data Input and create individual tasks

For the purpose of scalability, we create one HIT for one document to extract the information pieces inside. The task is to identify a most likely potential target location (if any), and extract N most important information pieces. According to our experience in pilot studies, setting $N = 2$ will encourage workers to focus on extracting comprehensive and important information and keep the workload low. In order to avoid redundancy as well as control the quality of work, we ask the first crowd worker to extract information pieces, and a second

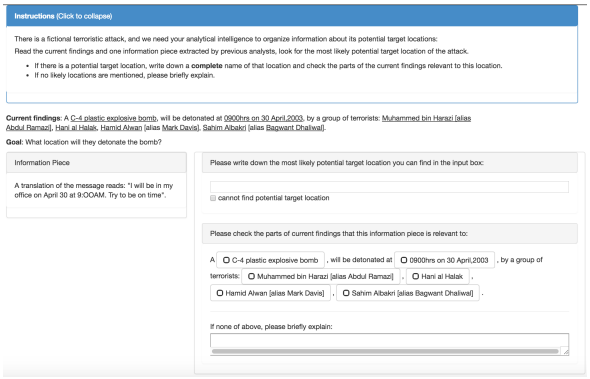

Figure 4. Step 3 Task Interface: Identify potential target locations and tag information pieces

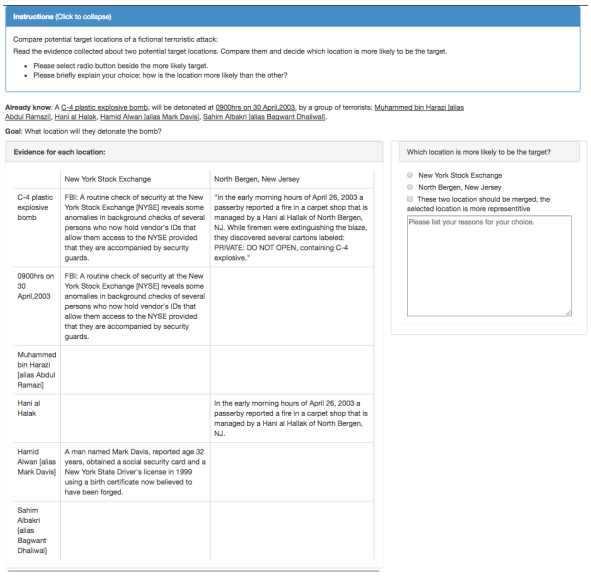

Figure 5. Step 4 Task Interface: Compare and select the most likely potential target locations

worker to "peer review" [24] the first worker's work. The second worker can edit or add to previous work if necessary. If the second worker modified the first worker's work, a third worker will be hired to check the latest work. The process continues until no new modification is made.

Aggregate responses of individual tasks into Step Data Output
Each of the relevant document will have several important information pieces reviewed by at least one more crowd worker. Our mechanism guaranteed that the information pieces are unique and cover all documents. The list of all information pieces along with potential target locations will be the Data Output of Step 2.

Step3: Schematize Information Pieces

The third step takes the Data Output of Step 2 as *Data Input*, and tag the information pieces by appropriate strategies. The tagged information pieces can be organized into a schema to draw inference from multiple perspectives. The schema will be the *Data Output* of Step 3. Step 3 organize new findings and prepare for new discovery.

Break down Step Data Input and create individual tasks
For the purpose of scalability, we create one HIT for one information piece. The tags vary by different investigation purposes and their appropriate strategies. In this paper, the investigation purpose is to find the most likely potential targets of terrorist attack and their connection to other findings of the attack. We have two types of tags: predefined tags and crowd-identified location tags. The predefined tags are the key elements in the current findings, and the crowd identifies potential target locations if any.

Aggregate responses of individual tasks into Step Data Output
Each information piece will have three voting for each predefined tags and three crowd-identified location tags (could be empty). We use majority vote for predefined tags: each information piece will get a tag selected by two or all three workers. Crowd-identified location tags are a bit trickier: some people cannot find any locations, some give different locations or locations at different levels of details. In order to balance the quality of work and preserve valuable insights, we first normalized the location tags identified (removing punctuation marks and transform all letters to lower case), then extract the shared parts (e.g. "First Union Bank Springfield, VA" and "Springfield, VA" have shared parts "Springfield, VA"). When taking a majority vote, longer location tags vote both for themselves and the shared parts in the majority vote. This will give a list of agreed potential targets. Finally, the Data Output of Step 3 is a list of potential target locations and their profiles. The profile contains the information pieces from where the location is identified, along with other information pieces that have the same tags as the source information pieces.

Step4: Develop Hypotheses from Schema

The fourth step takes the Data Output of Step 3 as *Data Input*, draws inference and develops hypotheses by comparing and/or combining different inference from different parts of the schema. The hypotheses will be the *Data Output* of Step 4. Step 4 discover and justify new knowledge.

Break down Step Data Input and create individual tasks
We first rank the location profiles by the number of tags, then break the tie by ranking by the number of information pieces. After that, we employ a single elimination tournament strategy to compare pairs of potential target locations. If there is an odd number of potential targets, the one ranked in the middle is waived from the first round of competition and automatically promoted to the second round. The process continues until only one winning location is left. Each pair-wise game constitutes a HIT assigned to three workers. Each worker is asked to select the most likely location and give a brief explanation. Locations selected by two or all three workers win the pair-wise game and are promoted to the next round of competition.

Aggregate responses of individual tasks into Step Data Output
The final winning location and its profile is the Data Output of Step 4.

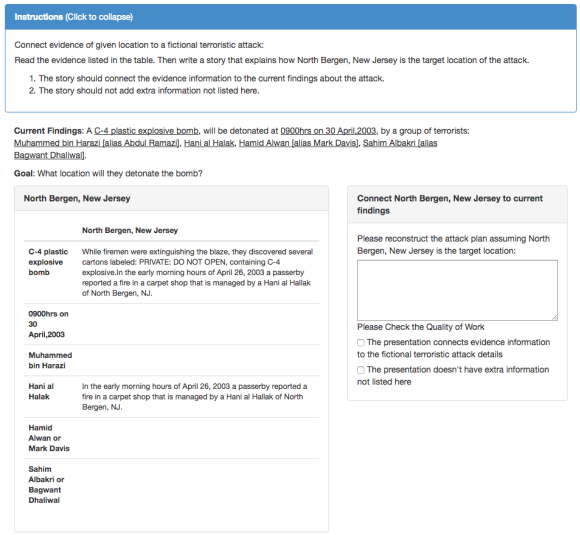

Figure 6. Step 5 Task Interface: Create narrative explanation for final analysis result

Step5: State Conclusion

The final step the Data Output from Step 4 as *Data Input*, explain the best hypotheses as clear and actionable analysis results to the customer. The explanation presents the best hypotheses, justify the plausibility by supporting evidence and logical connections to current findings. This will be the *Data Output* of Step 5, and the final product whole pipeline to be reviewed by the customer.

Break down Step Data Input and create individual tasks

The Data Input only has one winning location profile, thus no need to break down. We use the same peer review mechanism as Step 2: the first crowd worker read the winning location and profile information, create a presentation of analysis result. The second worker is given the same material to review the first worker's work. The reviewing process stops when no modification is made to the presentation.

Aggregate responses of individual tasks into Step Data Output

The finalized presentation (reviewed by at least one crowd worker) is the Data Output of Step 5, as well as the final product of the pipeline.

CASE STUDY

In order to validate the pipeline concept and experiment with the pipeline components, we deployed the tasks described above on Amazon Mechanical Turk to analyze a small dataset from intelligence analysis training material.

Dataset Preparation

We used *the Sign of Crescent* dataset from intelligence analysts training materials, which contains 41 fictional intelligence reports regarding three coordinated terrorist plots in three US cities. This dataset is widely used in visual analytics community. Each plot involves a group of at least four suspicious people. Each report contains one paragraph whose length ranges from 33 to 210 words. For the purpose of our experiment, we picked 13 documents, 10 are relevant to one of the terrorist plots, 3 are irrelevant. We provide the names of terrorists involved, the type of bomb they have, and the time when they plan to detonate the bomb. The investigation purpose is to find the most likely potential target of this terroristic attack.

Procedure

We deploy the pipeline with the same bottom-up workflow of both datasets on Amazon Mechanical Turk (AMT). We used majority vote mechanism for rating and voting tasks (Step1, Step3, and Step4), peer review mechanism for narrative creation tasks (Step2 and Step5). We implemented task interfaces for each task of each step with consistent design and layout. After accepting the HIT, workers would see instructions explaining the background context, task requirements, and purpose of the task. Once they have finished the task and click the "Submit" button, there will be a brief validation on their output depending on the task. If their work passes validation, the results will be submitted to MTurk system for the authors to review and approve/reject.

Participants

We recruited crowd workers from AMT, not require that workers are Masters nor did we set additional qualifications Workers must meet to work on our HITs. When workers accept a HIT, they are randomly assigned to a piece of input for a given step. For majority vote tasks (rating, tagging, comparing), we assign 3 workers for each task; for create-review tasks (extract, narrate), we assign 2 or more workers as needed. Each worker was unique and assigned to only one HIT to mitigate learning effects or collusion. Crowd workers who quit an accepted HIT without submitting it were not allowed to resume the unfinished work or take a new HIT.

Implementation Details

We used the interface provided by Amazon Mechanical Turk (AMT). To create a HIT, requesters on AMT create a new project under the "Create" tab, specifying the project name, contents to be displayed to workers, including the HIT title, description, keywords, reward per assignment, number of assignments per HIT, expiration date, auto-approval time (in case requesters forgot to review the results), and worker requirements. Each HIT can take a CSV file as data input and will produce a CSV file that aggregates crowd's result for all assignments. The task interface was implemented using HTML, CSS, Bootstrap, JQuery and related library for managing layout and aesthetic presentation, interactive interface, and result validation. The tasks are embedded in the MTurk system and the results are saved in CSV files on MTurk systems. Crowd workers with a web browser will be able to log in to Amazon Mechanical Turk and take on tasks.

Result

We recruited a total of 124 crowd workers to analyze the dataset: 18 workers in Step 1 (time spent in minutes: mean=10.9, median=4.3, std=15.6), 22 workers in Step 2 (time spent in minutes: mean=17.3, median=10.7, std=17.2), 78 workers in Step 3 (time spent in minutes: mean=12.1, median=3.1, std=16.0), 5 workers in Step 4 (time spent in minutes: mean=9.9, median=7.3, std=9.0), and 2 workers in

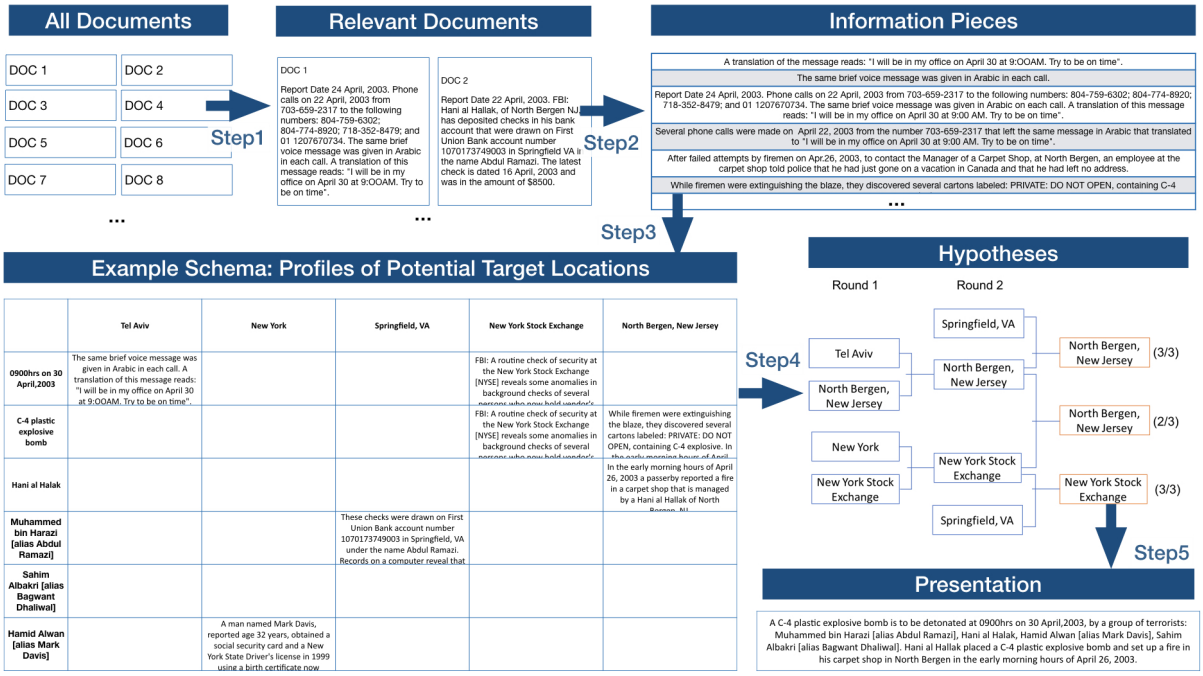

Figure 7. Overview: investigating potential target location of terroristic attack with the pipeline

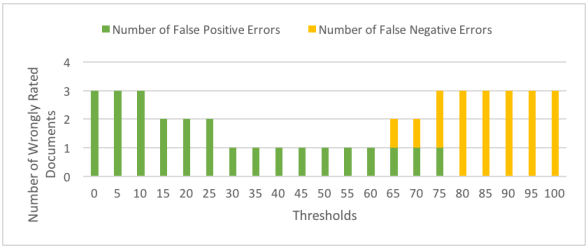

Figure 8. Number of errors in relevance judgment on documents by different thresholds

Step 5 (creating worker took 47.4 minutes, reviewing worker spent 1.5 minutes).

Thresholds and Accurate Document Relevance Judgment

In Step 1, 7 documents out of 13 directly mentioned one or more key elements. The remaining 6 documents (3 relevant and 3 irrelevant) are each rated by three crowd workers in a 0-100 scale. We picked a neutral range of 40-60 that generate 11 relevant documents (with one extra irrelevant document, leading to precision=90.1%, recall=100%) as Step 1 Data Output.

We also experimented with different thresholds with a step of 5, and calculated the different numbers of false positive (an irrelevant document considered relevant) and false negative (a relevant document considered irrelevant) errors in the final relevance judgment of documents (Figure. 8).

Quality and Usefulness of Information Pieces

In Step 2, a total of 26 information pieces are extracted from the documents. Each crowd worker extract at least two information pieces from the one document they are given, and can choose to add more for a bonus of $0.12 per extra information piece. After reviewing all the information pieces, we found that the crowd-extracted information pieces all follow the requirement of using complete sentences. The review process helped solved issues like misspelling, incomplete name reference, and missing elements (who, what, where, when). In terms of the usefulness to identify target locations, the crowd successfully extracted one information piece that mentions the target location, and 17 other information pieces about the the key elements. On the other hand, not all important information pieces were extracted. One important information piece that connects the actual target location to one of the terrorists didn't get extracted. Some information pieces about the relationship and roles of terrorists also didn't get extracted. This could be improved by re-executing Step 2 with further information request from upper steps, which is left as future work.

Accuracy of Predefined Tags and Location Tags

After Step 3, 18 out of all the 26 information pieces are tagged, excluding the four information pieces from the irrelevant document. We closely examine the tags, and found that the crowd tend to give information pieces more tags than needed. Some workers just selected every tag, some selected nothing when they cannot identify any location tags. We believe this is because the workers were confused and didn't understand the task.

Five location tags were created by majority, with two of the tags are the overlapping parts of location tags from different workers. One interesting finding is two different workers both identified a location "Tel Aviv" in the information piece "The same brief voice message was given in Arabic in each call. A translation of this message reads: "I will be in my office on April 30 at 9:OOAM. Try to be on time"." One of the workers even give very specific information "the lacation (location) is israel (Israel) at "Mike's Place", a restaurant in Tel Aviv". This indicates that we might connect the dots from the documents under investigation to the on-going events in real world.

Location Profiling and Selection

For each location tag recognized in Step 3, we organize the source information piece, the key elements tagged to this location, and other information pieces with the same tags, to form the profile of this location tag. Then we rank the location tags by amount of evidence (first by number of key elements, then by total number of information pieces), to prepare for the first-round competition. The five location tags are: *"Tel Aviv"*, *"New York"*, *"Springfield, VA"*, *"New York Stock Exchange"*, *"North Bergen, New Jersey"*, in order of least amount of evidence to the most. Hence, *"Springfield, VA"* is exempted from the first-round competition, and *"Tel Aviv"* is compared with *"North Bergen, New Jersey"*, *"New York"* is compared with *"New York Stock Exchange"*. The crowds picked *"North Bergen, New Jersey"* (3/3) and merged the profiles of *"New York"* and *"New York Stock Exchange"*, picking the latter as the more representative location tag (2/3), the remaining 1 crowd worker picked *"New York Stock Exchange"* as the winner. Now we have three location tags left, this triggers the final round: we ran comparisons between all three different possible combinations of two location tags. The final winning location is *"North Bergen, New Jersey"* (the last place the bomb was stored before transferred to the target location), the second place is *"New York Stock Exchange"* (the real answer), losing with only one vote. Even though the crowd narrowly missed the actual target, the winner result is the second crucial location to investigate.

Information Aggregation and Narrative Presentation

Using the profile of *"North Bergen, New Jersey"*, workers from the last step created a narrative that connects the evidence to current findings, to justify the likelihood of this place being a potential target. We observe that the workers copied the given current findings to start the story, then situate the evidence of North Bergen in the context to finish the story. We also found that the worker creating the story spent a very long time on the task, while the worker reviewing the story only spent 1.5 minutes. Despite that the first worker might have some idling time before submitting the task, we suggest that the creating task takes extra time because the incompleteness of the evidence given in Step 5, which makes it more challenging to connect the dots.

DISCUSSION

Although the crowds ranked the correct target location as second possible, their answer is the second important location where the terrorists store and transfer the bomb. In addition, the crowd showed their analytical intelligence in each step to identify important information, filter out irrelevant noises, connect the dots to draw meaningful hypotheses, and logically present a narrative story to explain their analysis results. We also learned how novice, transient crowd workers tends to work on such micro analytical tasks, which can help us better design future tasks.

False Positive Tolerance of the Pipeline

The crowd picked one irrelevant documents and extracted some useless information pieces in Step 1 and 2, however, the tagging mechanism in Step 3 guaranteed that only useful information pieces are tagged and put into profiles of potential target locations. This makes the pipeline tolerant of extra useless information, gold panning the important information step by step. Another benefit is that when investigating more complicated dataset, expert analysts might have more than one schematizing strategy to organize the information pieces; a useless information piece for one strategy might be useful to another. A comprehensive pool of information pieces provide more possibilities in later steps to draw alternative hypotheses thus leading to a more thorough and convincing analysis.

Feedback and Refine

In our case study, some important information pieces didn't get extracted by the crowds, resulting in incomplete profiles of the potential target locations, which we believe contributed to the narrow miss of correct target. This brings up the importance of evaluating and iterating each step to get more useful information. Our pipeline supports evaluating and passing feedback to refine previous work. Expert analysts evaluate the final presentation of the analysis results, and point out the weaknesses and improvement needed. The feedback on presentation is then passed back to Step 5, and new workers are recruited to address the feedback. If the feedback cannot be addressed given the previous input, the new workers then provide their new feedback to explain what information is needed in the previous input to improve the previous output. If the new workers in Step 5 provided new feedback instead of addressing the expert feedback, a revisit of Step 4 will be triggered to address the new feedback. The same process continues until the feedback can be addressed, which will push the refined output to the upper steps to address earlier feedback. Experts can also choose to review the intermediate results in steps other than Step 5 to decide which step needs to be revisited and provide more specific feedback for that step.

Crowds Analytical Intelligence and Patterns

We observe in the information foraging stage (Step 1-3), the crowds are more inclined to include the information if they are not sure about it. One of the crowd worker from Step 1 who rated the irrelevant document with a score 79 explained *"They are both using false documents"*. In Step2, we gave the instruction that ask the crowds to focus on more important information pieces, and we found that this led the crowds to extract only the important information from the documents. In contrast, in our pilots where we didn't have this piece of instruction and decide the number of information pieces requested from workers by number of words in documents, there are many incomplete sentences and useless information in the

result. In Step 3, some workers would give more tags than mentioned in the information piece. For example, the information piece *"These checks were drawn on First Union Bank account number 1070173749003 in Springfield, VA under the name Abdul Ramazi.* is tagged to be related to "C-4 explosive" by all three workers. We hypothesize that the crowds might infer that the money mentioned is used to purchase or transfer the C-4 explosive bomb.

In the synthesizing stage (Step3-5), we found that the crowds provide diverse perspective and connects their own knowledge to the dataset. In Step 3, the workers created a location tag "Tel Aviv" which is not mentioned anywhere in the documents. This connects the information from the documents to the knowledge of the crowds, and broadens the aspects from which to investigate the dataset. Similarly in Step 4, when comparing two location profiles, workers considered both the evidence given and their experience with the location. One worker who picked *"New York Stock Exchange"* over *"Springfield, VA"* and explained that *"The New York Stock Exchange is a specific, high value target for terrorists because a bomb attack there would likely cause many casualties and have a negative effect on the US economy. Springfield, VA is a very broad target and besides the fact that one of the terrorists lives there there isn't much evidence than an attack will take place there."*. In Step 5, we deliberately prevented the crowds from adding their own knowledge to keep the final presentation strictly follows the evidence from the documents. This will make the final product objective enough for expert analysts to evaluate.

REFERENCES

1. Simon Attfield and Ann Blandford. 2009. Improving the cost structure of sensemaking tasks: Analysing user concepts to inform information system design. *Human-Computer Interaction–Interact 2009* (2009), 532–545.
2. Simon Attfield and Ann Blandford. 2011. Making sense of digital footprints in team-based legal investigations: The acquisition of focus. *Human–Computer Interaction* 26, 1-2 (2011), 38–71.
3. Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM* 58, 8 (2015), 85–94.
4. Eric A. Bier, Stuart K. Card, and John W. Bodnar. 2010. Principles and Tools for Collaborative Entity-Based Intelligence Analysis. *IEEE Transactions on Visualization and Computer Graphics* 16, 2 (March 2010), 178–191. DOI:http://dx.doi.org/10.1109/TVCG.2009.104
5. Lauren Bradel, Alex Endert, Kristen Koch, Christopher Andrews, and Chris North. 2013a. Large High Resolution Displays for Co-located Collaborative Sensemaking: Display Usage and Territoriality. *Int. J. Hum.-Comput. Stud.* 71, 11 (Nov. 2013), 1078–1088. DOI: http://dx.doi.org/10.1016/j.ijhcs.2013.07.004
6. Lauren Bradel, Jessica Zeitz Self, Alex Endert, M Shahriar Hossain, Chris North, and Naren Ramakrishnan. 2013b. How analysts cognitively âĂIJconnect the dotsâĂİ. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*. IEEE, 24–26.
7. Joel Chan, Steven Dang, and Steven P. Dow. 2016. IdeaGens: Enabling Expert Facilitation of Crowd Brainstorming. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW '16 Companion)*. ACM, New York, NY, USA, 13–16. DOI: http://dx.doi.org/10.1145/2818052.2874313
8. R. Jordon Crouser and Remco Chang. 2012. An Affordance-Based Framework for Human Computation and Human-Computer Collaboration. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec. 2012), 2859–2868. DOI: http://dx.doi.org/10.1109/TVCG.2012.195
9. Emanuel Felipe Duarte, Edson Oliveira, Filipe Roseiro Côgo, and Roberto Pereira. 2015. Dico: a conceptual model to support the design and evaluation of advanced search features for exploratory search. In *Human-Computer Interaction*. Springer, 87–104.
10. Alex Endert, M Shahriar Hossain, Naren Ramakrishnan, Chris North, Patrick Fiaux, and Christopher Andrews. 2014. The human is the loop: new directions for visual analytics. *Journal of intelligent information systems* 43, 3 (2014), 411–435.
11. Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 247–256.
12. Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2258–2270. DOI: http://dx.doi.org/10.1145/2858036.2858364
13. Marti A. Hearst. 1999. Untangling Text Data Mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 3–10. DOI:http://dx.doi.org/10.3115/1034678.1034679
14. David R Karger, Sewoong Oh, and Devavrat Shah. 2013. Efficient crowdsourcing for multi-class labeling. *ACM SIGMETRICS Performance Evaluation Review* 41, 1 (2013), 81–92.
15. Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S Bernstein. 2018. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Design Thinking Research*. Springer, 79–104.

16. Aniket Kittur, Andrew M Peters, Abdigani Diriye, Trupti Telang, and Michael R Bove. 2013. Costs and benefits of structured information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2989–2998.
17. Gary Klein, J K Phillips, E L Rall, and D A Peluso. 2007. A data-frame theory of sensemaking. (01 2007).
18. Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2009. Turkit: tools for iterative tasks on mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 29–30.
19. Zhicheng Liu and John Stasko. 2010. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE transactions on visualization and computer graphics* 16, 6 (2010), 999–1008.
20. Devi Parikh and C Zitnick. 2011. Human-debugging of machines. *NIPS WCSSWC* 2, 7 (2011), 3.
21. Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. (2005), 2–4. https://analysis.mitre.org/proceedings/Final_Papers_Files/206_Camera_Ready_Paper.pdf
22. Alexander J. Quinn and Benjamin B. Bederson. 2011. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1403–1412. DOI: http://dx.doi.org/10.1145/1978942.1979148
23. Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. ACM, New York, NY, USA, 269–276. DOI: http://dx.doi.org/10.1145/169059.169209
24. Gregory D Saxton, Onook Oh, and Rajiv Kishore. 2013. Rules of crowdsourcing: Models, issues, and systems of control. *Information Systems Management* 30, 1 (2013), 2–20.
25. Kamran Sedig and Paul Parsons. 2013. Interaction design for complex cognitive activities with visual representations: A pattern-based approach. *AIS Transactions on Human-Computer Interaction* 5, 2 (2013), 84–133.
26. John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: Supporting Investigative Analysis Through Interactive Visualization. *Information Visualization* 7, 2 (April 2008), 118–132. DOI: http://dx.doi.org/10.1145/1466620.1466622
27. Vasilis Verroios and Michael S Bernstein. 2014. Context trees: Crowdsourcing global understanding from local views. In *Second AAAI Conference on Human Computation and Crowdsourcing*.
28. Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. 2012. CrowdER: Crowdsourcing Entity Resolution. *Proc. VLDB Endow.* 5, 11 (July 2012), 1483–1494. DOI: http://dx.doi.org/10.14778/2350229.2350263
29. William Wright, David Schroh, Pascale Proulx, Alex Skaburskis, and Brian Cort. 2006. The Sandbox for Analysis: Concepts and Methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 801–810. DOI:http://dx.doi.org/10.1145/1124772.1124890
30. Roman Yampolskiy. 2013. Turing test as a defining feature of AI-completeness. *Artificial intelligence, evolutionary computing and metaheuristics* (2013), 3–17.