Using Large-scale Open Source Data to Identify Potential Forced Migration

Yifang Wei[1], Abbie Taylor[1], Nili Sarit Yossinger[1], Eleanor Swingewood[1], Christopher Cronbaugh[1],
Dennis R. Quinn[1], Lisa Singh[1], Susan F. Martin[1], Sidney Berkowitz[1], Jeff Collmann[1], Susan McGrath[2]
[1]Georgetown University, Washington, DC, USA
[2]York University, Toronto, CA

ABSTRACT

This paper describes initial efforts to use open-source data to capture knowledge about forced migration in Iraq. Our goal is to understand the connection between open-source data and possible leading indicators of forced migration. For our preliminary analyses, we use a corpus of 2.6 million documents. Here we describe the techniques we used and challenges we faced. We conclude with recommendations for those using open-source data for grand-scale data science challenges.

1. INTRODUCTION

A humanitarian crisis can be defined as "any situation in which there is a widespread threat to life, physical safety, health or basic subsistence that is beyond the coping capacity of individuals and the communities in which they reside" [9]. It can be triggered by (1) acute events, including natural disasters nuclear and industrial accidents, 'acts of terrorism' and armed conflict, or by (2) slower-onset processes, including environmental degradation, general violence or political instability. More often than not, humanitarian crises occur due to a combination of these triggers, in addition to underlying stressors, such as poverty and deficient governance [9].

In recent decades, progress has been made in establishing early warning systems to alert the international community, as well as national and local actors, of impending humanitarian crises [8]. For example, tsunami and famine early warning systems monitor and analyze data relevant in anticipating acute and slow-onset crises, respectively, relying on scientific, technological, economic, social and other indicators. Predicting crises involving other triggers, such as conflict and generalized violence, has proven more difficult, although organizations such as the International Crisis Group put out regular alerts of worsening conditions [5]. Lagging behind these systems are effective early warnings of movements of people in response to humanitarian crises. More effective early warning of displacement will help governments and international organizations plan for such movements, as well as directly aid displaced persons before, during and after their exodus.

Patterns of forced migration in anticipation of, during, and following humanitarian crises are notoriously difficult to predict. Because detailed local data is difficult to obtain in a timely manner, we are interested in exploring whether or not open-source, online data can be used to help identify indirect, leading indicators of displacement/forced migration. Indicators relevant to this project include: economic, political, social, demographic and environmental changes affecting movements; intervening factors such as government refugee policies; and community and household characteristics. Parsing irrelevant information from the true indicators, calibrating results, understanding how these indicators change through time, and identifying and removing potential bias, requires large-scale data analysis and potentially, new computational methods for developing meaningful descriptive and predictive models.

To begin understanding the connection between open-source data and possible indicators of forced migration, we have drawn together a team of social scientists and computer scientists to analyze data from Raptor, a vast unstructured archive of over 600 million publicly available open-source media articles that has been actively compiling since 2006. New articles are added at the rate of approximately 300,000 per day by automated scraping of over 22,000 Internet sources in 46 languages across the globe.

Our initial focus is a case study of Iraq - a country located in a region of high volatility. For decades, and particularly in recent years, it has experienced renewed insecurity and displacement [7, 11], allowing for both retrospective and prospective analysis. We are interested in determining whether or not a data-driven approach using open-source data can be combined with domain expertise in a scalable manner to identify possible indirect indicators of forced migration. While many different data sources are considered for the problem at large, we begin by focusing on what we can determine using Raptor. Using over 2.6 million English documents that are broadly related to Iraq, the first step is to ascertain if a clear mapping exists between terms and concepts in Raptor and important events in Iraq and more broadly, the Middle East. Once we feel confident that this relationship exists and can be extracted efficiently, we can use the mapping to understand which concepts and changes in concepts can serve as leading indirect indicators of forced migration. In this work, we describe our approach for identifying this mapping and present the techniques we have used to better understand the strengths, weaknesses, and biases associated with open-source, big data analysis.

2. INTERDISCIPLINARY COMMUNITY

To address massive global issues, we advocate approaches that include a team of researchers from multiple disciplines. Without interdisciplinary insight, it becomes difficult to (1) fully understand the problem; (2) understand the data and the gaps; and (3) analyze the data effectively. The subject matter experts understand the factors that contribute to forced migration at the macro, meso, and micro levels, while computer scientists and statisticians understand how to mine, and analyze mass amounts of data. The coauthors on this paper are a subset of the researchers working on this problem. Currently, our research community consists of more than 25 researchers, technicians, policymakers and humanitarian practitioners from around the world. We mention this important collaboration because we want to emphasize that an approach using a team of computer scientists that understand how to process big data but lack subject matter expertise, will likely miss important and possibly even obvious, insights that domain experts are able to spot. Unfortunately, detailed subject matter expertise is not scalable. Therefore, one challenge becomes using expert knowledge judiciously to maintain the overall scalability of the solution.

3. APPROACH

Here we highlight the methodology and algorithms used to determine whether or not Raptor is a viable source for data that can be used as indirect indicators of forced migration. All of the described steps have been investigated in the database, data mining and information retrieval communities.

3.1 Data Extraction

In order to understand the data set, we processed each document in the corpus by extracting named entities and then stemming words and extracting unigrams, bigrams, and trigrams. We used the Stanford Named Entity Recognizer to extract named entities [3] and the Snowball implementation [4] of the Porter stemmer to stem words. The stoplist was custom developed. To reduce the number of distinct concepts, we also merged synonyms into a single concept using Wordnet [10]. Once this document processing was complete, approximately 4 million unigrams and 3.7 million named entities were maintained in our concept list. The number of concepts was particularly high because of the large number of foreign names and words identified.

In the corpus containing 2.6 million documents, there were documents from over 1600 different sources, including many from the Middle East. While the majority of articles were still from newspapers outside of the Middle East, there were a large number of articles from countries in the Middle East, including the United Arab Emirates, Iran, Turkey, Egypt, and Saudi Arabia. Only 6,600 articles came from Iraqi sources and even fewer came from other countries in the Middle East, e.g. Syria. Domain experts did not find this result unexpected since there has been a history of censorship of controversial issues such as forced displacement in many of these countries, particularly Iraq [11]. This means that future work will need to consider less traditional forms of media in areas where censorship of traditional media is high.

3.2 Domain Knowledge

Manually building models and ontologies is very labor-intensive. Because our goal is a scalable process, one challenge is to identify the most useful manually collected data for this task. We decided to collect three data sets that could be easily provided by domain experts: a small set of ground truth documents; a domain vocabulary; and a timeline of events in Iraq that are perceived by subject matter experts to be directly or indirectly relevant to forced migration.

The ground truth document set consists of twenty English articles that are considered early indicators of forced migration. While a larger ground truth document collection would be useful, manually identifying a large number of articles reduces the overall scalability of our approach. Therefore, we limited the number to twenty.

The subject matter experts also developed a relevant domain vocabulary list. This vocabulary list was divided into ten domains of knowledge, e.g. demographical, displacement, economics, governance, etc. Each domain listed between ten to fifty concepts relevant to the domain in the context of Iraq. Again, we limited the amount of manual labor, requesting a few hundred relevant words when the number of concepts in the entire corpus is over 7 million.

Finally, the subject matter experts created a timeline that contained very brief explanations of potentially relevant events, e.g. acts of violence, civil unrest, political developments, weather-related events, religious holidays, developmental initiatives, etc. This timeline served two purposes. The primary purpose was to provide computer scientists necessary background about the types of events that were relevant and needed to be identified in Raptor. Second, it served as a possible set of relevant concepts that could be used during the noise reduction process.

3.3 Noise Reduction

While we anticipated that the amount of *noise* in the corpus would be substantial compared to the amount of *relevant signal* related to forced migration, we verified this empirically. We considered the frequency of concepts and the frequency of concepts that appear together. We compared the frequent set of concepts in the corpus to concepts generated by our domain experts and concepts in our ground truth documents using the FP-Growth frequent itemsets algorithm [1] with a support of 0.05. We found that the types of concepts appearing frequently were very general and too broad to be used effectively. Example concepts in this frequent set, included year, time, news, people, 2013 and report.

Therefore, to reduce noise we employed the data provided by the subject matter experts. We used the concepts from each of the three types of domain data as *seeds* and identified documents that had a sufficient number of these concepts in them. The intuition was that relevant documents would have similar concepts as those in one of the lists provided by the subject matter experts. As an additional comparison, we also considered a list that contains only locations in Iraq. We pause to mention that even though it may seem logical to maintain all the documents in which any relevant concept from the three expert generated lists or the location list are found, doing so resulted in little reduction in the number of maintained documents.

Table 1 contains the comparison of these approaches. The first column specifies the number of concepts required to be in the document for it to be maintained. The next four columns show the number of documents retained (not considered noise) based on the number of concepts identified using different expert knowledge seeds. It also shows the number of ground truth documents retained in the final document set. Recall that there were initially twenty ground truth documents. The final column shows the overlap in the documents retained by each of the first three different seeding options (we exclude the location seeded comparison in the overlap since the number of retained documents is significantly lower than the other three methods). As an example, the first row of the table states that if only one of the concepts in the different concept list exists in a document in the corpus, the document is maintained. In the cases of the event, domain, and ground truth concept lists, over 2.2 million of the 2.6 million documents are maintained. There is very little reduction in the document set. In the case of the location list, only 96,000 documents are retained. In all cases, all twenty ground truth are still in the reduced corpus. From the table, we see that all the seeds do not lead to the same amount of noise reduction. The location seeds remove the most noise, followed by the event concept seeds (using 34 event concepts, all of the ground truth documents are maintained and the corpus is reduced to 878,386 documents). So while using a small number of concepts generated by domain experts helps us remove a large amount of noise, we still need to use other techniques to find the most relevant documents.

3.4 Topic Modeling

Because we have ground truth documents, we built a topic model using those twenty documents. While we considered a number of different topic models, Latent Dirichlet Allocation (LDA) [6] seems effective on this corpus. This may result because of the varied writing style and vocabulary of the documents in the corpus. Also, using a bag of words model produces topic lists that domain experts consider logical. We used the Mallet implementation [2] of

Table 1: Noise Reduction Comparison

Concept Threshold	**Using Event Concept Seeds**	**Using Domain Concept Seeds**	**Using Ground Truth Doc. Concept Seeds**	**Using Location Seeds**	**Overlapping Documents**
1	2,403,536 - 20	2,284,032 - 20	2,412,006 - 20	48,478 - 20	2,279,112
5	2,257,822 - 20	1,898,790 - 20	2,276,797 - 20	826 - 5	1,898,676
10	2,148,835 - 20	1,385,391 - 20	2,182,860 - 20	1 - 0	1,385,381
50	936,341 - 18	76,883 - 10	1,107,688 - 20	NA	76,883
100	234,030 - 11	2,698 - 0	376,520 - 17	NA	2,698

LDA to generate ten topics each containing twenty different concepts. We chose ten topics to see if the learned concepts overlapped with the ten domains developed by the subject matter experts. We then identified those documents in our reduced corpus that have a large amount of overlap with the topic model generated from the concepts in the ground truth document set. We considered these to be the set of possible relevant documents. As a comparison, we also ran LDA on the reduced corpus and then used Manhattan distance with a threshold of 0.1 to identify the most similar documents to the ground truth documents.

4. PRELIMINARY FINDINGS

While we have a number of interesting findings, we present a few that can be useful for others conducting similar analyses.

Since completely unsupervised methods did not produce as meaningful results, it is important for these types of endeavors to build a team of subject matter experts that can provide guidance. In our case, our subject matter experts developed three types of domain knowledge. Preliminary findings suggest that doing so results in much better precision of relevant documents. Some of the relevant documents found were broadly relevant to forced migration in the region, but were not as relevant to Iraqi forced migration. This was an interesting finding and that suggests the need for a weighting scheme to determine the overall relevance. Further, the number of concepts used to guide the document selection process ranges from between a few hundred to a few thousand depending upon the specific domain knowledge used in the seeding process. To further improve the relevant document set, we are now having experts rank the concepts based on perceived importance.

Location in conjunction with the domain seeds was a consistent piece of information that was present in relevant documents. Further, if we focus on locations as seeds for finding relevant documents, we are able to remove significantly more noise while still maintaining the ground truth documents in the corpus.

When analyzing the topics generated from the ground truth documents, there are two interesting findings. Most of the twenty documents have concepts from multiple topics in them. Second, we found that when mapping topics to domains, the concepts in the topic model covered eight of the ten domains. Considering other languages will likely improve coverage.

As a final finding, the subset of documents that are similar to the ground truth documents for the different corpi created, was a small number ranging from 2 to 112. Domain experts are evaluating their relevance to see if they contain possible leading indicators.

5. DATA SCIENCE RECOMMENDATIONS

While this initial analysis indicates that relevant documents exist in the corpus, there were a number of challenges that we encountered. We now make recommendations for more effective use of open-source data for grand-scale data science challenges.

Data quality: Take time to assess the quality of the data and the data sources. Our initial analysis showed us that our data sources were biased toward countries outside the region of interest. To deal with this, we added more data from sources in the Middle East.

Dynamic data changes: Because all the sources are independently owned, the availability and type of data that is accessible may change over time. Be prepared to spot changes and compensate for them.

Expect missing data: While we found that some of our domains had ample coverage in terms of concepts, there were a few that were not well covered. While every attempt should be made to obtain full coverage, it is better to design algorithms that still work in the presence of this missing data.

Processing power: Regular standalone servers do not have the processing power to handle parsing and analyzing big data. Expect to setup or purchase time on a distributed cloud infrastructure. In our case, we used five nodes of a 28 node distributed cluster, where each node had 32 GB of memory.

This initial analysis shows the promise of using open-source data for identifying movements of people. Specifically, we find forced migration-related documents for our Iraq case study do exist within our Raptor data set. Not all the documents contained leading indicators for identifying movement of people when a humanitarian crisis occurs. However, a clear mapping that can be extracted exists between data in Raptor and important concepts related to internal and cross-border movement relevant to Iraq.

Acknowledgments

This work was supported in part by NSF Grant #SMA-1338507.

References

[1] Find frequent item sets with fp-growth algorithm. http://www.borgelt.net/doc/fpgrowth/fpgrowth.html.

[2] Machine learning for language toolkit. http://mallet.cs.umass.edu/.

[3] Stanford named entity recognizor. http://nlp.stanford.edu/software/CRF-NER.shtml.

[4] The english (porter2) stemming algorithm. http://snowball.tartarus.org/algorithms/english/stemmer.html.

[5] H. Adelman. Difficulties in early warning: networking and conflict management. 1998.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[7] D. Chatty and N. Mansour. Unlocking protracted displacement: An iraqi case study. *Refugee survey quarterly*, 30(4):50–83, 2011.

[8] J. L. Davies and T. R. Gurr. Preventive measures: an overview. *Preventive Measures: Building Risk Assessment and Crisis Early Warning Systems*, pages 1–14, 1998.

[9] S. F. Martin, S. Weerasinghe, and A. Taylor. *Humanitarian Crises and Migration: Causes, Consequences and Responses*. Routledge, 2014.

[10] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[11] J. Sassoon. *The Iraqi Refugees: The New Crisis in the Middle-East*, volume 3. IB Tauris, 2008.