Enabling Compliance of Environmental Conditions

Utkarsh Dwivedi
IBM Research India, New Delhi
utkdwive@in.ibm.com

Anirban Dasgupta
IIT Gandhinagar
anirbandg@iitgn.ac.in

ABSTRACT

Industrial projects in India have to agree to specific sets of environmental conditions in order to function. Lack of compliance with these conditions results both in irreversible damage to the local environment as well as conflicts among the industry and the local community. Our aim is to provide a system that raises general awareness in the local community about the environmental conditions in vogue among the nearby industries so that compliance violations can be reported early on. We outline work in progress to mine the text of the clearance conditions and build a searchable mapping system that can answer various queries about these conditions.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Environmental compliance, Latent Dirichlet Allocation, text mining.

1. INTRODUCTION

Over the last few decades, India has seen a rapid growth in industrial development. The number of new industrial projects that are being setup each year has in fact been accelerating. The challenge, therefore, is to reconcile the economic and social needs of industrialization with the resulting environmental impact. Different government regulatory agencies (e.g. ministry of Environmental Affairs) measure and check environmental impact. For each new industrial site or infrastructure expansion, the appropriate authorities conduct an Environmental Impact Assessment (EIA). The concerns raised by the regulatory authorities, local citizenry and civic bodies (e.g. village or city administration) are codified in terms of "Environmental Clearance" (EC) documents. These documents form a legal contract that the corresponding industry has to abide by. Ensuring compliance to the conditions stated in the EC is the obvious next step needed to maintain the environmental status quo. Currently, such compliance checks are done by the same regulatory authorities, who are severely resource constrained. As a result, compliance to these conditions is a severe issue. A recent report [3] by the non-profit organization Namati, in collaboration with local organizations in Mundra in Gujarat, details rather extensively the inadequacy of compliance in most industrial projects there and its effect on the local fishing and other communities. Citizen complaints to the authorities are an important mechanism to trigger compliance monitoring and enforcement. As per the OECD report [1], in Maharashtra, between April 2004 and March 2005, citizens filed 761 complaints with respect to air, water, solid waste and noise pollution. However, citizen complaints may not be productive (or not useful in a legal proceeding) if they are unrelated to the exact clearance conditions. Thus a potential solution to the compliance issue could be found if the larger community could be made aware of the environmental issues and the various restrictions that have been imposed on the industries functioning in their locality.

In order to both encourage citizen complaints by spreading awareness of the clearance conditions in effect, and to ensure that such complaints are actually related to environmental clearance conditions, it would thus be useful if we could make the EC documents accessible to as broad a population as possible.

2. PROBLEM AND TOOLS

The Ministry of Environmental Affairs (MoEF) does an admirable job of keeping the EC documents in the public domain. However, there are still a number of barriers for a general user in comprehending and utilizing the information stored in these EC documents. These are the following:

- EC documents are available publically, but they are not easily searchable, and extracting any information from them requires significant amount of effort.
- The technical language of these documents inhibits an easy comprehension of what the conditions are. For each industrial site, there is actually a collection of documents, each being addendums to the original. Each such document contains a collection of clearance conditions.
- The user is typically not able to get a global view of the clearance conditions associated with a particular region or location.

3. SOLUTION

We collected the EC documents and identified their locations using a regular expression based strategy. Next, we applied text mining tools to model the collected clearance conditions from all documents. This identifies the latent topic for each clearance conditions, and can then be used to cluster conditions as well as identify similar conditions to a given one.

3.1 Collecting and preprocessing the data

Environmental clearances are available at the MoEF website http://environmentalclearances.nic.in (13614 clearances granted till July, 2014). An EC document contains a file number, addresses of the MoEF and concerned company, subject with location and name of project. It begins with an introduction to the project, and conditions of the clearance if granted, subdivided into general and specific conditions. This text suffered from irregularities in following a common notation for clearances, file number formats. We programmatically downloaded html files, scraped them, and violations were extracted using REGEX. We extracted the file number for indexing, location of a project and

ACM DEV 2015, December 1–2, 2015, London, United Kingdom.

ACM 978-1-4503-3490-7/15/12.

DOI: http://dx.doi.org/10.1145/XXXXXXX.XXXXXXX

the clearance conditions. Then the data was stripped of stopwords, and stemmed to build a vocabulary. Finally, we made a corpus of 24 manually chosen and 600 scraped EC docs, yielding 700 and 8500 clearance conditions respectively. To make sense of this text, we used 20 EC's, a subset of this corpus for further analysis.

4. USING LDA FOR CLUSTERING

We used Latent Dirichlet Allocation (LDA) technique for making sense of the compliance conditions, using online batch based approximation, which is faster than a full corpus based approach [2]. This allows us to define similarity measures between conditions and cluster them. We use the open source Python implementation provided by the authors [2].

5. RESULTS

We found best parameters for LDA empirically and implemented the recommended TFIDF inspired term score i.e. Hellinger distance grouping for the sense of similarity between two documents. These results (Figure 1, 2) are for a 13 topic run of LDA over a corpus of 700 conditions from 24 clearance documents.

5.1 Document Similarity

To allow us to determine the efficacy of the treatment of these conditions by LDA, we made the corpus searchable using an open source searching and indexing library, Whoosh. When a query of a particular clearance would be entered, two of the most similar conditions would be returned using the similarity score [2]. Below is an example of a query and its response. The following is an example of a query violation:

"Separate funds shall be allocated for implementation of environmental protection measures long with item-wise break-up."

For the above query, the following violations were returned as the nearest:

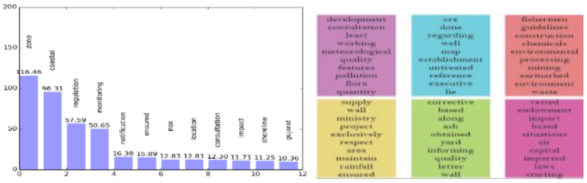
"Special package with implementation schedule for free potable drinking water supply in the nearby villages and schools shall be undertaken in a time bound manner."

"A special scheme for upliftment of SC & ST population in the study area shall be formulated and implemented in a time bound manner…"

We are currently working to further improve this clustering. We are also working to setup a search index using keywords e.g.'*forest'*, '*encroachment'* etc. allowing an easier way to explore these documents.

5.2 Clustering and keywords

We then clustered conditions using these topic distribution vectors, using K-means method with K determined empirically as K = 10. We extracted top 10 keywords in these cluster of clearance conditions using TFIDF. Figure 1 shows term distributions over for a topic. Figure 2 shows top words of some clusters that made most sense.



Figures: (1) Topic term distribution (2) TFIDF ranked word descriptors for topic clusters.

6. LOCATION MAPPING

To map these documents we made use of the geo information in the subject line. Figure 3 shows the results over a map of Gujarat. The following screenshot gives an idea of how an overview of the data is superimposed on a map.

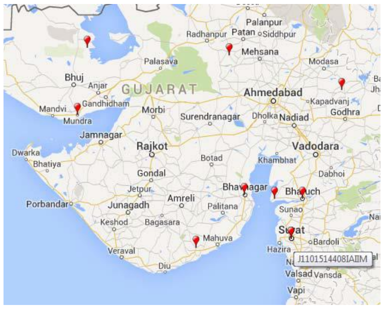


Figure 1: Automatic location mapping results

7. CONCLUSION

This project is work-in-progress and we are currently working on developing better indexing and visualization methods for the EC documents. We plan to use MTurk to transcribe scanned pdfs and construct a rich repository of the regulatory information in India. On the other hand, in collaboration with Namati, we have also built an Ushahidi based web platform (http://env-compliance.in) to collect the compliance related complaints by the local environmental activists. We are currently working to integrate the compliance report collecting portal with the document analysis and visualization site and build a one-stop portal where interested users can analyze existing environmental compliance conditions as well as upload reports and evidences for violations of these. We plan to conduct a user assessment of this system, to see that if relevant search results are being returned, and whether this helps users identify the EC conditions that are most appropriate for a specific violation.

8. REFERENCES

[1] Environmental Compliance and Enforcement in India: Rapid Assessment. http://www.oecd.org/environment/outreach/37838061.pdf.

[2] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In advances in neural information processing systems, pages 856–864, 2010.

[3] Namati, Mundra Hitrakshak Manch (Forum for the Protection of Rights in Mundra), Machimar Adhikar Sangharsh Sangathan (MASS), Ujjas Mahila Sangathan. Closing the enforcement gap:Findings of a community-led ground truthing of environmental violations in Mundra. http://www.namati.org/wpcontent/uploads/2013/ 10/Kutch-proofed-0.1-merged.pdf