Predicting Amendments via Right to Information Query Log Analysis

Nayantara Kotoky [*1] **Vijaya V. Saradhi** [*1]

Abstract

Amendments to laws are necessary to keep up with the changing needs of the society. Such a process is largely manual, and takes feedbacks from the society for the introduction of an amendment. The Right to information (RTI) Act 2005 gives Indian citizens the opportunity to interact with the government. The present paper discusses how analysis of the RTI system can assist in mining feedbacks that can suggest potential amendments, and the process and challenges associated with solving this problem. Extracting latent patterns from the RTI query-reply process via learning algorithms is the main task at hand; representation of the RTI data, identifying applicable learning models and most importantly interpreting the results in the context of amendments is at the heart of this research work.

1. Introduction to the Problem

Amendments are any change brought to the laws, statute or legislative system. When we talk about amendments, the idea is to change or upgrade any law that does not fit well into the existing social system. This is a key factor in the implementation of democracy, where the people's concerns or convenience is put upfront. There have been 101 amendments to the Indian constitution from 1950 to 2016 (at the rate of 1.5 per year). Although it is a frequent process in India, there are no specific rules that states what exactly qualifies for an amendment. These amendments are proposed by observing the execution of the laws in the society and taking this feedback into account for driving the amendment into effect. This process is largely manual, and rests in the hands of the officials responsible for framing laws. Statistics over time are taken into account while proposing amendments. In view of the above observations, the problem that our present work aims to address is '*whether learning algorithms can predict or propose amendments*'. We can imagine a machine that takes citizen's concerns, issues and feedbacks, and outputs pointers for amendment to laws. Given a source where we can observe citizen's attitudes and opinions, learning algorithms can help mine relevant information that can be used as cues for driving amendments. The Right to Information (RTI) Act, 2005 (RTI, 2005) provides us with this opportunity.

The RTI act allows Indian citizens to obtain information from any public authority. Citizens can put their queries in an RTI application, accompanied by a sum of Rs. 10 as RTI fee, and submit their application to the Public Information Officer (PIO) who is responsible for replying to their queries. Such a system allows citizens to interact with the government machinery, and maintain accountability in the functions of the government. Each government institution has in its possession a collection of the RTI queries received by it, and its associated log data. This log data contains information regarding citizens' queries and concerns with the institution, and is a worthwhile podium to witness their communication with the government bodies.

With regards to performing analysis of RTI query log data for proposing amendments, the present work puts forward the following research questions:

1. Can machine learning help us find amendments?
2. What constitutes a potential amendment?
3. What patterns in the RTI query-reply process help us find amendments?
4. How to represent the RTI data?
5. What learning models can help us in extracting latent patterns that can lead to amendments?
6. How to validate the results?

We attempt to address the research questions during the course of the paper, and present our initial efforts towards this goal.

2. Motivation

There has been previous use of RTI query-reply data to bring amendments. Infact, the RTI act itself has been

[*]Equal contribution [1]Indian Institute of Technology Guwahati, Assam, India. Correspondence to: Nayantara Kotoky <nayantara.kotoky@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

amended by observing and taking into account RTI statistics. The first example of such an amendment is the inclusion of Indian Postal Order (IPO) as a method of RTI application fee payment. The typical methods used for paying the application fee of Rs. 10 were demand drafts, bankers cheque or cash. IPOs with their easy availability and having the least service charge were easily adopted for the fee payment. It was observed that many RTI applications were rejected by institutions citing that IPOs are not a mode for such payment. Over time, this caught the government's eye and an amendment was made to the RTI act that specified IPOs as another mode for RTI fee payment (ipo, 2005). We perceive that 'repeated rejections' played a role in bringing an amendment. In another instant, political parties used to receive RTI applications enquiring about their source of funding. They used to reject these applications on the grounds that political parties are not government institutions and hence do not fall under the scope of the RTI act. This repeated rejection of RTI applications under the same ground led to the introduction of another amendment, where it was specified that political parties do not constitute a government body (pol, 2005).

In both the examples, we see that 'repeated rejection' was used as a cue for introduction of amendment. It is imperative to think that RTI query-reply process shall possess more such information that when identified can be used as drivers for amendments. Given a database of RTI queries and associated statistics, learning algorithms can uncover and identify further latent patterns that can be used for predicting or proposing amendments.

3. Literature

In order to perform query log analysis on RTI data, we first look into studies performed on the RTI act and its effects on the society. We also look into different studies done on political data and analysis of different query-logs to understand the modelling methods and their objectives .

3.1. Execution of the RTI Act

Noronha (Noronha, 2001) has evaluated the influence of the state RTI act (an earlier version of the RTI Act 2005) in the Indian Union Territory of Goa. According to the author, the RTI act is progressive in its views but is laden with poor execution at the ground level. The study also concludes that part of the shortcomings lie on the demeanour of the public officers in the institutions to put efforts in handing out information. Kulkarni (Kulkarni, 2008) identifies two key issues existing in information retrieval via RTI in the Indian state of Maharashtra. The study finds that (a) information about PIOs is not published by all the institutions (b) replies to the RTI questions are raw in nature and not easily understandable. Studies from the perspective of implementation of the RTI act is done in (Lemieux & Trapnell, 2016). They list key descriptors such as demand for information, institutional capacity, oversight etc. that can indicate effectiveness of implementation. RTI specific indicators to understand the effectiveness of implementation of the RTI act have been suggested in the literature. These are (i) input oriented: involves the necessary government machinery like appointment of PIO and organization of the records to facilitate information. (ii) output oriented: takes into account institute-specific summaries such as number of queries received, number of queries replied, number of queries rejected (on varying grounds), swiftness with which the replies are served within stipulated time (iii) outcome oriented: measures the impact of the RTI law on the society (Lemieux & Trapnell, 2016). The above studies reveal that the act has not been executed nor understood at a desired level. However, much of these studies are qualitative in nature, and leaves a huge opportunity of research to figure out the underlying characteristics and issues in depth with respect to the act.

3.2. Modelling the political domain

In the domain of politics, (Gerrish & Blei, 2012) developed a probabilistic model for legislative data to identify lawmakers' voting patterns in specific political issues. They have used topic modelling using Latent Dirichlet Allocation to divide the voting bills into specific subjects (issues) and applied the ideal point model (based on Item Response Theory) to predict each lawmaker's stance in a one-dimensional ideal point trait for each of the different issues. (Martin & Quinn, 2002) introduced Bayesian inference for dynamic item response model to investigate preference change for all justices in US Supreme Court over time. They identified the justices' stance over specific cases, and compared their change in stances across the time axis. (Nguyen et al., 2015) took text of bills, votes of lawmakers and language of debates of the lawmakers to arrive at a joint model of predicting Tea-Party Republicans (political group with specific ideology) characteristics of all the lawmakers. The paper introduced Hierarchical Ideal Point Topic Model, which provides a rich picture of policy issues, framing, and voting behaviour using votes, bill text, and the language that legislators use when debating bills. Poole and Rosenthal (Poole & Rosenthal, 1985) analysed a variant of voting patterns, namely, roll call data for legislators' votes. They took US voting data where choosers are representatives of the law or senators, and the choices are binary, that is, yes or no. They developed a unidimensional model of probabilistic roll call voting, and the methods can be applied to the analysis of voting in popular elections and other forms of political choice behavior. Tae Yano (Yano, 2013) proposed a predictive model of the bill survival using Congressional bill data corpus by employing discriminative

log linear models.

3.3. Analysis of web search queries

Web query analysis is a popular area of research. Study has been done to find user goals from web queries by Lucchese et al. (Lucchese et al., 2013). They clustered queries of different users by proposing query similarity function based on common terms and common semantics (unsupervised). They also used supervised learning by extracting temporal, log sequence and other query features to classify whether the queries have similar task goals. Beitzel et al. (Beitzel et al., 2004) analysed temporal aspects of web queries and identified trends across a day and a year based on different categories of queries. They divided the query logs in temporal sessions (monthly and yearly) and used Pearson's correlation to find overlap of queries for different time sessions. Traditional information retrieval mostly depended on simple term matching between queries and documents. However, it has been observed over time that understanding the meaning of the query is important in improving the precision of a search result, like certain keywords have more relevance in a given query and synonyms need to be identified. Attempts to find such hidden semantics in the queries have been made by (Deerwester et al., 1990).

Apart from web search queries, other forms of queries have been analysed. Question-answer (Q/A) systems do not retrieve documents, but give brief, relevant answers in short text. Semantic information in questions and answers classification is studied in (Moschitti et al., 2007). (Hovy et al., 2002) has presented a new topology to support construction of Q/A systems. Test questions are used to determine the qualification of individuals or behaviour of events, often used for surveys and aptitude measurement. Preston et al. (Preston et al., 2015) used Nominal Response Model to understand the relationships among family members. They surveyed families with people of different age groups and used the item's discrimination parameter to determine which items are most appropriate to judge family relationships depending on different ages of test takers. Rubio et al. (Rubio et al., 2007) used Graded Response Model (GRM) to assess personality traits in Computerised Adaptive Testing technique. They considered 28 items with six graded responses. The authors compared two methods for developing a common metric for the GRM and studied differences between separate calibration runs and concurrent calibration using the combined data.

It is noteworthy that each work consists of different methods of modelling and analysis of the data, which is guided by the type of data, the representational purpose of the data and the interpretability of the results in the context of the purpose of analysis.

4. Problem Statement

The objective of this work is to collect the RTI queries and associated response-statistics (date of reply to RTI queries, queries that have been rejected and their grounds of rejection) of institutions across India, model thus collected text data and identify latent patterns in the RTI database. We are concerned with identifying specific latent patterns that can assist in predicting or proposing amendments. In order to attempt this problem, we briefly enumerate the steps involved and issues to be addressed in undertaking the task of analysing the RTI query log.

4.1. Data Collection

An "RTI database" is created as part of our research. Our dataset consists of the RTI applications that have been posted to all public educational institutions by the citizens of India. RTI data is not found online, but have to be collected from each individual institution. The data is collected by filing an RTI application of our own asking for the data required, namely, all the RTI applications received by the institution, date of reply of each query and the rejected queries with their grounds of rejection. Till date, we have filed RTI applications to a total of 360 institutions and have collected data from 52 institutions.

4.2. Data Modelling and Representation

Representation of the RTI text data is the first step in the proposed analysis. However, traditional text modelling methods like vector space model, latent semantic index etc. do not assist in identifying latent patterns that are useful for our proposed objective. Hence we need to look towards other computational models that not only represent the RTI data effectively, but also extract relevant information that are indicative of amendments. *Cluster analysis* can help in grouping RTI queries into common topics, and determine outliers which do not fit into specific topic groups. The outliers are suggestive of unique queries, and thereby unique concerns of citizens. Both of these can reveal patterns that provide scope for finding flaws and inconsistencies in our rules and regulations, which are cues for potential amendments. *Temporal analysis* may identify trends in the frequency of RTI queries, revealing relation between query categories (topic) and the time of the year they are posted. Certain categories are frequently queried throughout the year compared to others. If such queries break pattern, that becomes an interesting point of analysis and presents scope for potential amendments.

4.3. Identifying relevant latent patters

There is no specific definition in the literature that states what patterns qualify for amendments in the legal system.

Given the RTI data, not all latent patterns shall lead to the proposal of amendments. The task at hand is to find those specific results that can be interpreted in the correct context as drivers for amendments. Transparency of an institution, implementation efficiency of laws etc. are other patterns that we are interested in. Our initial analysis of RTI query-reply data is aimed towards identifying and quantifying these two patterns. For this objective, we have two points of focus: (a) select learning models that find these latent patterns, (b) interpreting the results in the context of amendments. These parameters are difficult to quantify with the help of popular text analysis or query log analysis models. Hence our choice of learning models is guided not just by its representation methods but by the information that it discovers from the RTI query-reply data.

5. Challenges

While undertaking this research work, a few challenges that we need to address are as follows.

5.1. Data pre-processing

The RTI data has reached us in many forms, like emails, photocopies, images etc. The first challenge in using the RTI data is digitizing the acquired data to a machine readable form. From the raw data, our work requires the extraction of RTI queries, date of application of the RTI queries, date of reply of each query. In addition, RTI queries in the local languages need to be translated to English. Besides English and Hindi, India has 21 other recognized languages. This is reflected in the RTI data received by us.

5.2. Query-text representation

Web queries are represented with models like 'tf-idf' since the aim is to match them with relevant documents and retrieve those documents, whereas queries in a question-answer system are commonly represented as a 'parse tree'. RTI queries fall into neither of these categories. RTI queries are longer than web queries, contains natural language and often ambiguous and requires a human (PIO) to interpret the information requirement. RTI data are different from survey questions and hence models like longitudinal analysis, regression analysis etc. commonly used for survey analysis are not directly applicable. Once again we look beyond the traditional methods of text representation and instead venture to find an atypical learning approach that is an amalgamation of statistics, probability and query analysis methods.

5.3. Learning Models

Finding pointers for potential amendments cannot be classified as a traditional supervised or unsupervised learning problem, nor does it conform to the popular models applied for query log analysis discussed in the Literature section 3. The outcomes of applying learning methods directly are not applicable for our work since the objective of those models are not in alignment of what we are attempting to find. Therefore we need to look for an unconventional learning model that effectively represents the RTI text data and identify information relevant for our target. Our requirement lies more in using learning algorithms to learn parameters that in association with other information and in proper context (social and legal constructs) can be employed for suggesting amendments.

5.4. Validation

One of the bottlenecks in the advancement of social data analysis is the challenge in accumulating data. Most of the data are private to each individual; hence collecting them for analysis is difficult. This also means that once some analysis is done, it cannot be reproduced for validation due to privacy and third party issues. In our work, validation of the model is challenging. Waiting for an amendment to take place after communicating the results of our work is not feasible.

6. Preliminary Efforts

Keeping in view the objective of finding pointers for amendments, we provide an initial attempt for representation of the RTI queries, and extract two latent patterns, namely, transparency of institutions and effectiveness of implementation of the RTI Act across India.

Transparency of government offices is a burning issue in our country. This was also the main cause behind the advent of the RTI Act. Hence as a first step in the RTI query log analysis, we want to understand parameters like how transparent the institutions are in replying to RTI queries, and whether the RTI act has been implemented uniformly across institutions in India. An act shall be amended due to its inefficiencies, hence finding effectiveness of implementation is a relevant parameter in this context. We quantify these two parameters based on: 1) whether all institutions in India are equally transparent, based on the reply frequency for all institutions; 2) whether the RTI Act has been implemented with equal efficiency across India, based on the reply frequencies across different types of query-categories. We have taken the reply rates of RTI queries for all institutions for different query-categories, and represented the data as a matrix of institutions in rows and query-categories in columns. Each ij cell in the matrix is the reply percentage of that institution i against the query-category j. The parameters 'transparency' and 'effectiveness of implementation' are quantified by performing psychometric analysis on the reply-matrix. This work can be

found in (Kotoky & Saradhi).

7. Conclusion

The paper proposes the problem of finding amendments via learning algorithms. The RTI query-reply process provides us with an opportunity to observe feedbacks from citizen's interaction with government bodies, and have been used for introducing amendments to the RTI act. The paper discusses the tasks required for extracting latent patterns from the RTI queries and its associated statistics. An RTI database is being created for this purpose. The research work demands finding appropriate learning models to represent and analyze the RTI data to extract specific latent patterns that can be interpreted as drivers for amendments.The outcomes of such analysis can assist in making a more robust process for introduction of amendments.

References

The Right to Information Act 2005. http://righttoinformation.gov.in/rti-act.pdf, Accessed: 2017-06-13, 2005.

Inclusion of IPOs. http://ccis.nic.in/WriteReadData/CircularPortal/D2/D02rti/10_9_2008-IR26042011.pdf, Accessed: 2017-06-13, 2005.

The right to information (amendment) bill, 2013. http://www.prsindia.org/uploads/media/RTI%20%28A%29/RTI%20%28A%29%20Bill,%202013.pdf, Accessed: 2017-06-13, 2005.

Beitzel, Steven M, Jensen, Eric C, Chowdhury, Abdur, Grossman, David, and Frieder, Ophir. Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 321–328. ACM, 2004.

Deerwester, Scott, Dumais, Susan T, Furnas, George W, Landauer, Thomas K, and Harshman, Richard. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

Gerrish, Sean and Blei, David M. How they vote: Issue-adjusted models of legislative behavior. In *Advances in Neural Information Processing Systems*, pp. 2753–2761, 2012.

Hovy, Eduard, Hermjakob, Ulf, and Ravichandran, Deepak. A question/answer typology with surface text patterns. In *Proceedings of the second international conference on Human Language Technology Research*, pp. 247–251. Morgan Kaufmann Publishers Inc., 2002.

Kotoky, Nayantara and Saradhi, Vijaya V. Right to information query modelling via graded response model.

Kulkarni, Ashwini. Governance and the right to information in maharashtra. *Economic and Political Weekly*, pp. 15–17, 2008.

Lemieux, Victoria Louise and Trapnell, Stephanie E. Public access to information for development: a guide to effective implementation of right to information laws. *Directions in Development. Washington, D.C*, 2016.

Lucchese, Claudio, Orlando, Salvatore, Perego, Raffaele, Silvestri, Fabrizio, and Tolomei, Gabriele. Discovering tasks from search engine query logs. *ACM Transactions on Information Systems (TOIS)*, 31(3):14, 2013.

Martin, Andrew D and Quinn, Kevin M. Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political Analysis*, pp. 134–153, 2002.

Moschitti, Alessandro, Quarteroni, Silvia, Basili, Roberto, and Manandhar, Suresh. Exploiting syntactic and shallow semantic kernels for question answer classification. In *Annual meeting-association for computational linguistics*, volume 45, pp. 776, 2007.

Nguyen, Viet-An, Boyd-Graber, Jordan, Resnik, Philip, and Miler, Kristina. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in t. In *Association for Computational Linguistics*, 2015.

Noronha, Frederick. Goa : Perils of knowing. *Economic and Political Weekly*, Vol. 36(Issue No. 32), 11 2001.

Poole, Keith T and Rosenthal, Howard. A spatial model for legislative roll call analysis. *American Journal of Political Science*, pp. 357–384, 1985.

Preston, Kathleen Suzanne Johnson, Parral, Skye N, Gottfried, Allen W, Oliver, Pamella H, Gottfried, Adele Eskeles, Ibrahim, Sirena M, and Delany, Danielle. Applying the nominal response model within a longitudinal framework to construct the positive family relationships scale. *Educational and Psychological Measurement*, pp. 0013164414568717, 2015.

Rubio, Víctor J, Aguado, David, Hontangas, Pedro M, and Hernández, José M. Psychometric properties of an emotional adjustment measure: An application of the graded response model. *European Journal of Psychological Assessment*, 23(1):39–46, 2007.

Yano, Tae. *Text as Actuator: Text-Driven Response Modeling and Prediction in Politics*. PhD thesis, Carnegie Mellon University, 2013.