Towards a General Framework for Data-Driven City Comparison and Ranking

Vishalaksh Aggarwal
IBM Research - India
vishalaksh@in.ibm.com

Biplav Srivastava
IBM Research - India
sbiplav@in.ibm.com

Srikanth Tamilselvam
IBM Research - India
srikanth.tamilselvam@in.ibm.com

ABSTRACT

Knowing about cities that one lives in or wants to visit is of much interest to citizens, tourists, businesses, investors and governments. Open government data provides us this opportunity since data about various domains like crime, traffic and health are being made available by the government. In this paper, we present our approach of using open data from multiple agencies and domains in comparing and ranking cities in a developing country. The framework relies on vocabulary based data normalisation to overcome data collection noise and easily scales with new domains.

Keywords: City Comparison, Inconsistent Data, Data Preparation, Open Data, Clustering, Visualization
Category of Submission: Demonstration
Demo URL: http://city-explorer.mybluemix.net/
Demo Status: Prototype ready

1. INTRODUCTION

A valuable piece of information for citizens, tourists, businesses, investors and governments is to know how good a city is in itself and in comparison with others can be valuable. The traditional way to know this is by surveys. However, such results have many problems like limitations of sample size and possibility of survey bias.

This is where open data can help. Open data is the practice by organizations and governments to make their data amenable to reuse. For cities, governments around the world are making data available about various domains like crime, accident and health[1].

A number of data-driven approaches for exploring cities are coming up. The approach of Global City Indicators (GCI) [2] is to define a set of indicators grouped around themes like *governance*, *people* and *safety*. The indicators are expressed using terms from an ontology[3] and its quantitative values are calculated to help compare cities. City Data[4] is a new initiative where data is collected from mutliple sources and then organized for rapid discovery. Unfortunately, it works only for US cities and technical details are not public.

In this paper, we present the City Explorer app which generates insights about cities and their comparison using open data. What sets it apart from other approaches is that it works directly with noisy data prevalent in developing countries when multiple decentralized agencies produce data. The city data comes from multiple domains (like crime, accidents and health), is organized by districts and spans multiple years. We uses data cleaning and vocabulary based normalization to prepare grounds for city exploration across domains and cities.

Our contributions are that we:

1. formalize characteristics of Indian open data, a first in literature.
2. demonstrate multi-dimensional exploration of city performance based on open data across domains and time
3. perform vocabulary-based city data normalization across domains
4. demonstrate unprecedented multi-dimensional comparison of all cities where data allows.
5. provide a scalable framework which will work for more domains and data from more countries

The rest of the paper is organized as follows: we begin with a background and formalization of Indian open city data and then present our solution approach. We then discuss the salient points with examples and give pointers to future work.

2. DATA CHARACTERISTICS

India has over half a million villages and hundreds of towns and cities. However, the recognized unit of territory is a *district*. The full list of districts in India along with their standard names and unique identifier can be computed from the official controlled vocabulary [5]. (The vocabulary itself has list of states and list of districts for each state). We will refer to it as D^* and call it Normalized District Names (NDN). Its size is 721 (i.e., $\mid D^* \mid$).

We use data about Indian districts from India's open data portal[6]. They correspond to the domains of crime[7], accidents[8] and health services[9]. The data is about districts and each city can have 1 or more districts. We use the terms *city* and *district* interchangeably.

2.1 A Formalization of Data Used

The input data is a set S of 3-dimensional vectors Vs. Each vector V_i represents data of a domain i. We use the domains of crime (V_c), accidents (V_a) and health (V_h). The dimensions (x,y,z) of V_i represent districts, domain attributes and year, respectively.

A reported data for a domain V_i is (x_j, y_k, z_l). Here, $x_j \in D^{V_i}$. We will use D^{V_i} to refer to the set of districts in a V_i and call them

Domain	$\mid D^{V_i} \mid$	$\mid A^{V_i} \mid$	$\mid Y^{V_i} \mid$
Crime (V_c)	807	30	12
Accident (V_a)	50	4	1
Health (V_h)	629	5	1

Table 1: Statistics about data used.

Domain	$\mid V_i \mid$	$\mid V'_i \mid$
Crime (V_c)	8597	6843
Accident (V_a)	50	40
Health (V_h)	637	539

Table 2: Statistics about data filtered based on district name analysis.

Data District Names (DDN). The data for the domains come from different agencies and do not use the normalized district names. y_k refers to a domain attribute from the set of attributes A^{V_i} which can vary from domain to domain. z_l refers to a year. The datasets may have different range of years and we use Y^{V_i} to refer to years in V_i.

2.2 Data Challenges

The challenges with the data are:

1. Inconsistent naming of districts across all datasets. We found that $D^{V_i} \nsubseteq D^*$, for V_c, V_a and V_h, as one would have expected, indicating district names are not in the master list.
2. Inconsistent availability of data for different districts across domains.
3. Inconsistent availability of data for years. Even in the same domain, data for different districts need not be for the same years.
4. Inconsistent labeling of missing data with NA, - or blanks.

We overcome the inconsistent naming problem by using a standard district name vocabulary[5]. Since the data came from multiple sources, there were disparity in districts,state names like 'Kolkata', 'Calcutta' and 'New Delhi', 'Delhi' to name a few. To resolve them across datasets, we used heuristics[10] to arrive at potential matches which were then manually verified. If any of the following happens, it is a potential match: (1) If a district's name was contained in another's name, (2) the similarity in names between two districts is above 0.4 as measured by cosine similarity, (3) The rest were then manually verified to resolve disparity. To handle inconsistent data availability, we restrict comparison to only the cases where districts and years are common.To tackle inconsistent labeling of missing data, we simply assumed all the data which could not be parsed into numbers as missing data.

3. SOLUTION APPROACH

Our solution is a two step process where in the first step input data V_i is normalized and filtered based on D^*. This offline process is then followed by the online process where selected districts X^{V_i} for the years Y^{V_i} are pair-wise compared and displayed. We also demonstrate that when data exists, we can also do all-pair comparison to gain meaningful insights about cities in India.

The resulting City Explorer system is shown in Figure 1. It is cloud-based and available online[11].

3.1 Data Preparation

In this offline process, the controlled vocabulary services[4] is accessed by REST API to get the latest state information, namely unique state code and state name. Currently it has information on all 36 Indian states and union territories. It is available in both XML and JSON format. For our solution, we rely on JSON format. It also exposes APIs to get district and taluk (another administrative terriotrial unit) level information for each of the state. Our experiments are based upon state and district level information alone. District level information are merged with its state details and this we refer to as Normalized District Names (NDD) D^*.

Each district name j in V_i (in S) is matched with D^*. Table 3 shows result of exact match comparison while Table 4 shows substring match. The second approach increased matching values substantially, but had errors that needed human intervention. For example, substring match equated 'PATNA' city with 'VISAKHAPATNAM' and 'ASANSOL DURGAPUR' with 'DURG' though they are not the same in reality. Such errors accounted for roughly 1% of the districts. Each matched records were analyzed by two annotators to remove such wrong matches. We also found synonymous places with different spellings like 'Bengaluru', 'Bglr', 'Bangalore' which we did not include in our current matching. Only districts matching D^{V_i} are retained for further processing while the rest are discarded. Table 2 shows details of domain wise retained data D^{V_i}. (V_h originally had many synonymous districts).

3.2 Pair-wise City Comparison

Comparing a pair of cities (districts) c_1 and c_2 means one wants to compare two corresponding vectors $v_1^i(c_1, i, j)$ and $v_2^i(c_2, i, j)$ for each domain i and year j. The notation can be suitably modified if a city is being compared to its own performance in a domain but in a different year.

For each of the city chosen for comparison, its relative ranking for each of attributes A^{V_i} of vector V_i is computed for the same year Y^{V_i}. The left side of Figure 1 shows barometer ranking of both the compared districts which represents relative ranking of the selected districts with respect to other districts on the two extremes for each of the attribute, in this case highest and lowest number for attribute murder. The positioning on the linear scale is based on the relative number of crimes. Likewise, the color coding is also based on relative number of crimes. The district with lowest number of crime is marked in green color (lowest rank), while the one with highest is marked in red (highest rank) and the one with the average of two numbers in yellow is positioned in the middle. The number of crimes are shown below the name of the district. The selected district is written below the scale while the extremes are written above the scale. The Figure in brackets signifies its relative rank out of the total number of districts in scope for that year and domain. (A minor note is that if a category, like health facilities, is of reverse semantics (where more is better), the signs of values are reversed before processing.) In the right hand side of Figure 1, the line series chart displays the distribution of an attribute for the years Y^{V_i}.

To get a composite view of the two cities across all attributes in a domain, we calculate a dominance score between the cities as defined below with δ_t^i. The score is asymmetric and measures the percentage of times a city dominates the other with ϵ accounting for the cases where their ranks are the same. Figure 2 shows the scores calculated for two cities in the system for each domain.

$$\delta_t^i(c_1, c_2) = \frac{\#(rank(c_1) \prec rank(c_2)) * 100}{\mid A^{V_i} - \epsilon \mid} \quad (1)$$

We now calculate the overall dominance score across all domains as defined below with δ_t. For the experiments, w_i was 1, thus

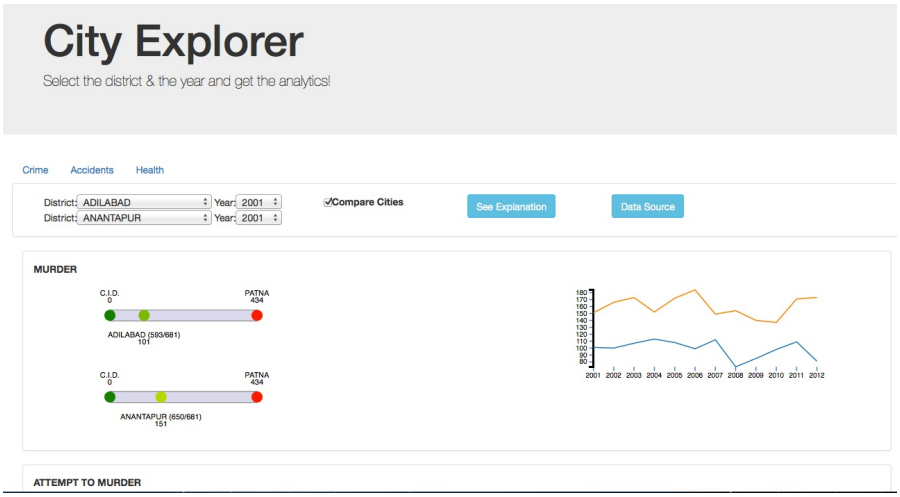

Figure 1: A City Explorer View.

DatasetNamel	# Total Records	# Matches	# Non Matches
Crime	807	485	302
Accident	50	35	15
Health	629	396	133

Table 3: Exact matches between DDNs and NDNs

DatasetNamel	# Total Records	# Matches	# Non Matches
Crime	807	621	186
Accident	50	40	10
Health	629	544	85

Table 4: Partial matches between DDNs and NDNs

weighing all attributes equally across all the domains.

$$\delta_t(c_1, c_2) = \frac{\sum_i w_i.\delta_t^i(c_1, c_2)}{\sum_i w_i} \quad (2)$$

We now define the dominance relation $\succ_t$ between two cities for year t iff:

$$(c_1 \succ_t c_2) = \begin{cases} true, & \text{if } \delta_t(c_1, c_2) \succ 50\% \\ false, & \text{otherwise} \end{cases}$$

In Figure 2, Chandigarh $\succ_t$ Patna and conveys that Chandigarh broadly dominates Patna across the considered categories and their attributes.

3.3 Comparing All Cities

Once we can compare a pair of cities, we also try to compare all cities for which data is available. A problem we faced was that not all cities (districts) release data for all the domains. Hence, the analysis was restricted to only 17 districts that the data was available and only for 2012.

In Figure 3, dominance relationship for all districts are shown, where the district had published data for all the domains under consideration, i.e., crime, accident and health. Here, an edge from district A to district B represents the dominance of A over B. We notice that Indore dominates all cities (source node) and Dhanbad is dominated by all (sink node). Hence, they correspond to the best and worst cities based on available data. We are not aware of any prior work giving this insight.

Figure 2: Comparing a pair of cities.

4. DISCUSSION AND FUTURE WORK

We presented a general data-driven approach of comparing cities using open data in a developing country. The data mining contributions relate to:

1. Data preparation - handling missing data and district normalization
2. Clustering - coming up with a general city comparison framework across categories and attributes
3. Interactive visualization - that can handle missing data and time

The data is noisy and non-uniform. As a result, the insights presented are preliminary and likely to improve / change with better data publishing practices. Specifically, we note a data-based limitation of the current results. The tier-1 cities in India are: Delhi, Mumbai, Kolkatta, Chennai, Bangalore and Hyderabad. However, none of them figure in the list of 17 cities for all-city comparison. The reason is that an official mapping of cities to districts is not available and we must create one manually.

The work can be extended in many ways: (a) We presently looked at simple weighted aggregation function for comparing cities but many others can be considered. (b) The controlled vocabulary for India has information for only districts. One can extend it to cities and villages, and then report results at these level of granularities. (c) For India, one can build a map between cities and the districts they contain to support drill-down, roll-up of comparisons and results. (d) One can normalize city data based on their population to

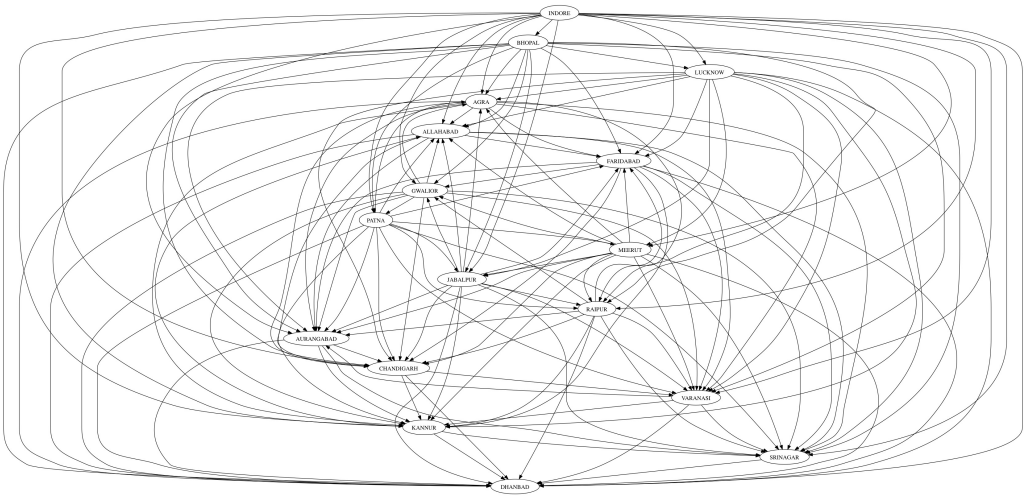

Figure 3: Dominance relationship among districts based on 2012 crime, accident and health data.

provide a more balanced comparison. (e) One can extend to data from more countries [12].

Apart from the general methods developed, we addressed a few UI challenges in building the online system. We believe they are important for any useful system in this space.

1. Making the selectable input options intuitive: We had to fetch a unique list of districts from dataset and arrange them alphabetically so that the user can easily locate the desired district in the dropdown. Also, once the district was selected, we had to first find out the list of years for which data was available for the selected district from the dataset and display it in the second dropdown.

2. Responsiveness to user selections: we have to continuously watch for any new selection of district or year that user makes from the dropdown so that the charts are re-rendered upon the selection.

3. Make the visualizations responsive: The visualizations should not overlap each other upon the change of size of the browser window. Upon decrease of width, the 2nd visualization comes below the 1st one so that both of them are still visible and the user only has to scroll downwards.

5. CONCLUSION

In this paper, we presented a general approach for comparing and ranking cities using open data. We took India as the case study and considered data from different domains from multiple agencies and domains. The framework is general-purpose and can give novel insights about a city with respect to its past, relative to its peers and as a whole group.

6. REFERENCES

[1] "Open data barometer report, second edition 2015," http://barometer.opendataresearch.org/, accessed 6 May 2015.

[2] GCI, "Global city indicators," in *At http://www.cityindicators.org/Deliverables/ListAccessed 6 May 2015*, 2015.

[3] M. Fox, "Foundation ontologies requirements for global city indicators," in *The AAAI 2014 Workshop on Semantic Cities: Beyond Open Data to Models, Standards and Reasoning, Quebec City, Canada*, 2014.

[4] "City data," http://www.city-data.com/, accessed 6 May 2015.

[5] "Controlled vocabulary services," http://vocab.nic.in, accessed 6 May 2015.

[6] "Indian open data," http://data.gov.in, accessed 6 May 2015.

[7] "District-wise crime under various sections of indian penal code (ipc) crimes," https://data.gov.in/catalog/district-wise-crimes-under-various-sections-indian-penal-code-ipc-crimes, accessed 6 May 2015.

[8] "Road accidents profile of selected cities — open government data (ogd) platform india," https://data.gov.in/catalog/road-accidents-profile-selected-cities, accessed 6 May 2015.

[9] "District-wise availability of health centres in india," https://data.gov.in/catalog/district-wise-availability-health-centres-india, accessed 6 May 2015.

[10] A. Singhal, "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

[11] "City explorer," http://city-explorer.mybluemix.net/, accessed 9 July 2015.

[12] "Us open data," http://data.gov, accessed 6 May 2015.