Information Extraction and Interactive Visualization of Road Accident Related News

Halima Akhter
Graduate, Department of
Computer Science, Stamford
University Bangladesh
213/3 Tejkuni Para Tejgaon
Dhaka – 1215, Bangladesh

ABSTRACT

This paper describes a strategy of extracting information from raw data and visualizing them in web browser. Raw data are collected from newspaper. These raw data are in English language. By implementing text mining process specific information extracted and this process explained clearly. Derived information is specifically on road accident related news but raw data contains all kind of news. One of the significant parts of this process is to visualize retrieved information in an interactive way in web browser. Interactive geographical map, bar chart, bubble chart have used for visualizing data. Interactive data visualization helps people to understand data easily, let them take actions on the data, give comfort to their eyes as data visualization is a combination of art and science. The whole process contains two major parts – information extraction, interactive visual representation of extracted information. People can understand easily by interactive visualization at a glance. Interactive data visualization can make people more concern about road accident. There is also an evaluation part which performed in two ways, one is by measuring the accuracy of information by standard information extraction process and another way to evaluate is by taking feedback from user. It is done by reading some of the raw data files of some days and matching the similarity with the retrieved data of the same days. Some limitations of this process have been described and some improvement which can be take to extend the process and will help to go deeper level have been described in the final part.

General Terms

Computer Science, Data Mining.

Keywords

Text mining, Information Extraction, Natural Language Processing, Visualization, Accident.

1. INTRODUCTION

Application of Natural Language Processing (NLP) [2] is the main job of this process. Information extracted from raw data using text mining process via NLP and then visualized in web browser. Data visualization helps a user to understand easily any information instead of reading the whole text. Day by day, it is becoming popular in online. Extracting information from any kind of corpus and visualizing can save lot of times. In this paper, only road accident related information from various newspapers are extracted. It describes a process of retrieving information from raw data and visualizes them. Raw data collected from newspapers, which is in English language. One of the significant parts of the whole process is to visualize extracted information in an interactive way in web browser. Interactive geographical map, bar chart, bubble chart have used for visualizing data.

The formation of this paper is divided in different chapters. In chapter 2 total steps of the whole process explained. Chapter 3 explained about information extraction from raw data. Derived information's visualization process described in chapter 4. Extracted information's evaluation is explained in chapter 5. Chapter 6 explains the utilization and improvement of this process. Finally, chapter 7 is about conclusion of this process.

2. STEPS OF THE PROCESS

This whole process is an advance solution to extract data and visualize them in an interactive way. This process is made of some simple computer programs used two platform for data extraction and data visualization. Therefore, together the total process is capable to help people in their busy life.

First, need to retrieve informative data from raw data for visualizing data. This data extraction is done by performing text mining process. Analyzing raw data for road accident related data and filtering this information by text mining process.After extracting data, some web programs interactively visualize road accident related information in web browser. Geographical maps, bubble chart, bar chart are used for interactive visualization.It is easy to understand something from a view. Visual data can make a right concept of any complicated statistical data. Thus, day-by-day data visualization is becoming popular in online. It does not only make an easy way but also saves times in our life.

2.1 Raw data

Authentic raw data are used for this process. More than 900 day's news reports from different newspapers in Bangladesh were collected as raw data. The language of this news is English. Each of the news reports is about each of the individual days of three years. Therefore, this is very favorable to have a huge amount of authentic data. Each individual text file contains more than 6000 words because each of them contains all kind of news. Thus, a complex text analytical strategy is required. The following paragraph is a part of a raw text file which is used for this process.police arrested a nephew of former bnp lawmaker shahidul islam on thursday night.teenage boy, girl commit suicidea schoolboy and a domestic help allegedly committedsuicide at two places in the capital yesterday.health hazards...guard stabbed to death in mirpurextortionists stabbed a security guard to death at west kazipara in the city's mirpur area yesterday.four of a bridal party killed in road mishap four members of a bridal party were killed and 30 others injured in a road accident on the bogra-rangpur highway yesterday.precarious tree posiotion...

This is a part of a newspaper.

2.2 Designing the process

The total process is designed with two major parts – information extraction, interactive visual representation of information.

Implementation of Natural Language Processing (NLP) for extracting information is the first part of this process. Raw Text data are analyzed here for specific information.

Extracted information is visualized in an interactive way [1] in web browser.

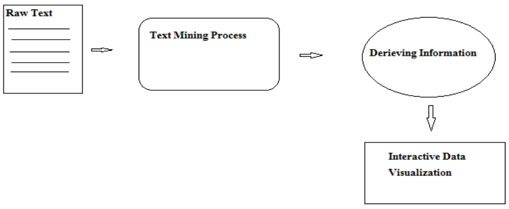


Fig 1: Designing the process.

In this process, some computer programs are used for information extraction. Text mining process is actually done by these computer programs. These programs analyze those raw data of various newspapers to find especially road accident related information. To do this they maintain some algorithms and a train data set.

After extracting specific information, some other web programs are used to visualize information in an interactive way in web browser. For pictorial presentation, Data visualization is used. Data Visualization makes extracted information not only easy to understand but also live.

2.3 Interactive Data visualization

In today’s world, people are getting busier day-by-day. Sometimes they miss what they have to know or will be helpful if they know. For this reason, a pictorial presentation of data is required. Data visualization helps to not only for pictorial representation but also for making it easy to understand at a glance. Interactive data visualization makes the visualization live. It enables direct actions on data and users can take these actions on data.

There are many kinds of data diagrams in data visualization. Some of the examples of data visualization are maps, bar chart, bubble chart, stream graph, scatter plot, tree map. Animations can be built in these visualizations programmatically and thus it can be made live. Just not only by watching the visualization but also by clicking on some object user can know about its information. Here in this process there is a map of Bangladesh is used and it has it’s json data base of all of its regions. So, when a user click one of the regions it shows the number of total death by road accident and also show colors for high/low rates of road accident.

3. IMPLEMINTING TEXT MINING PROCESS

Information extraction is a task of Natural Language Processing. For information’s accuracy, detecting right entities is crucial. Text mining is the process, which turn text into data by analyzing. This part is consists of sentence segmentation, tokenization [3], parts of speech tagging [4], entity recognition, relation recognition. Some training set for comparison and detecting road accident related information of sentences were used. Raw texts are given as input for analyzing and detecting specific entity’s information. Raw texts have been made more specific by applying classify text method of NLP. A dictionary with relevant word is used against raw text data and then Naïve Bayes classifier algorithm implemented for classifying road accident related data. This method helps to make groups with similar entities.

Naïve Bayes is a conditional probability model. An expression can be calculated for P (label | features). Where, P is for probability, features mean the specific set of features, label means the particular label that have those specific set of features.

The equation is :

P(label | features) = P(label, features) / P(features)

Since, raw data were not only about specific data so to classify raw data there are 400 hand written examples as training set of road accident related incident. If a particular incident of road accident is not found in the training set then that particular incident is taken in the training set. Information extraction is an important part of text mining process. Structured information extracting is the main job here. Entities and relation between entities is detected here. Some significant steps are performed to complete IE. The architecture of information extraction is given below.

At first, raw text is split into sentences using sentence tokenization and each sentence is subdivided into words using word tokenization. To make easy named entity recognition step, each sentence goes through part of speech tags. In named entity recognition step, possible entities are searched in each sentence. Named entity recognition step is more complicated. Noun phrase chunking technique is used for entity recognition. The last step is relation extraction, for finding specific patterns between pairs of entities in the text. These relation patterns are used to build structured information.

In the final state, due to some lack of information in some incident for mistakes, some information has been filled manually instead of lacking in information.

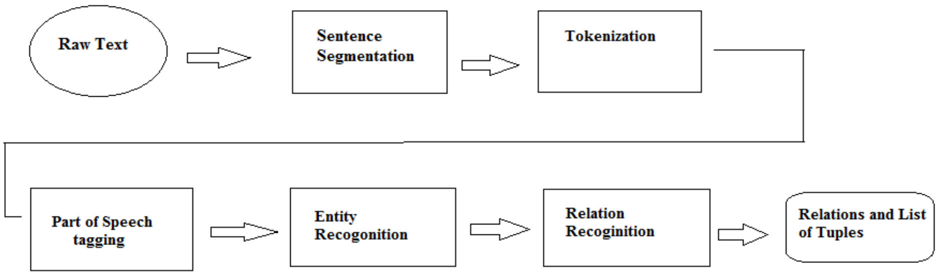


Fig 2: Architecture of Information Extraction.

4. DATA VISUALIZATION

Data visualization is a combination of science and art. It is involved with visual representation of data; this data refers meaningful information which can be some attributes or variable. To communicate information clearly and efficiently to a user by information graphics is the main target of data visualization. An effective data visualization can help an user for analyzing and unification about information data.

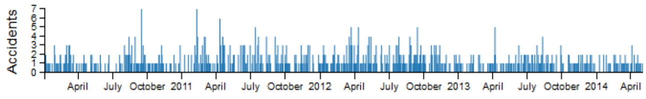
Here, data visualization is used for encouraging eye to compare different pieces of data, showing a large number of data in a small place, coherent large information, serving a clear purpose. Web browser is the media to show these designed information in an interactive way.

In this process, only road accident related information is extracted and visualized. Entities of road accident related information are:

- Date of the accident
- Related Vehicles name
- Accident type
- Number of death by the accident
- Location of the accident

There are many ways and many kinds of tools for data visualization. In this process, a library named dc.js (Dimensional Charting Javascript) is used. It allows highly efficient exploration on large data; it provides an easy way to data visualization and analysis interactively in web browser and on mobile device.

Here are some screenshots of road accident related data visualization.In figure 3, the plain screenshot of map of Bangladesh and a bar chart of years 2011 to 2014 is showed. The color is involved with number of accident. In the map, deep blue presents high occurrence of accident, light blue presents low occurrence of accidents. In the bar chart, monthly occurrences of accidents are visualized.



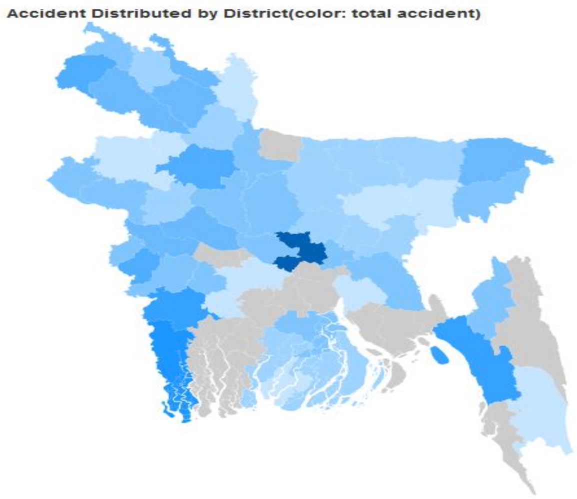


Fig 3: Data visualization by Map.

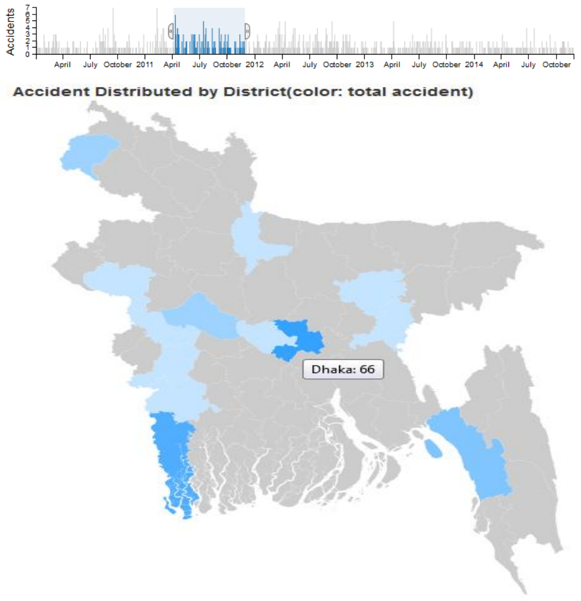


Fig 4: Data visualization interactively using location.

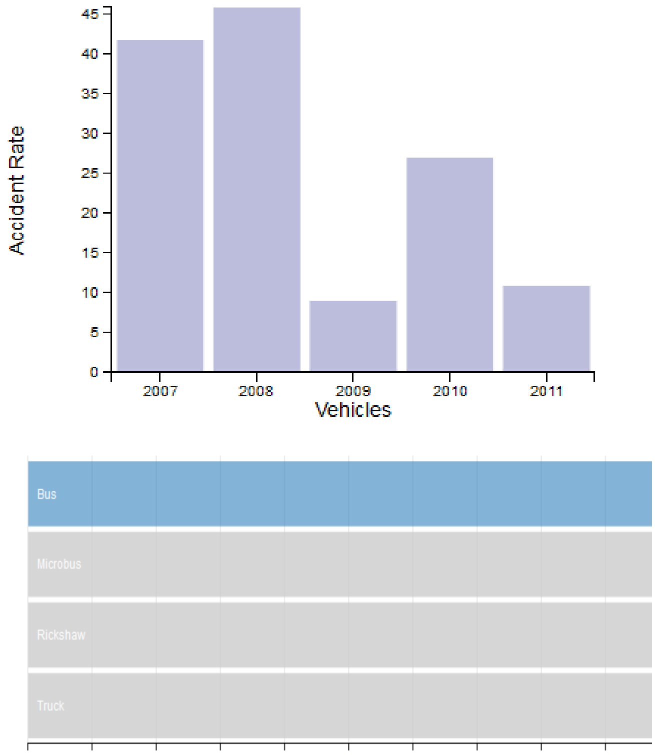


Fig 5: Number of different types of vehicles fall in accident.

The bar chart and the map are connected internally. In figure 5, dragging the cursor on the bar in the year of 2012 and the map changes automatically, shows the result only for the year 2012. In addition, when the cursor put on the location in the map it shows the number of accidents. In the third figure, there are a row chart for vehicles type and a bar chart for presenting number of accident. Suppose, clicking in vehicle BUS's row of the row chart then it will change the bar chart for showing the number of accident occurred only by the vehicle BUS. All these changes occur interactively.

5. EVALUATION

Keeping the accuracy score high of extracted information is significant. That is why evaluating the performance of this whole process is crucial. Evaluation of this process is done by two ways. The first one is, measuring the accuracy of the derived information by standard information extraction process. The second way is using feedback from users of the whole process. Some previously unseen texts is used for evaluation part which is part of the unused raw text.

Extracted information of road accident related news was considered accepted if correlated words are found or implicitly mentioned in texts. This method is used for evaluating location of the accident, vehicles name, accident type.

6. UTILIZATION AND IMPROVEMENT

This process can be utilized in online especially for news presentation. Because of data visualization, it is easy to compare different pieces of data, showing a large number of data in a small place, coherent large information, serving a clear purpose. Online is the mass media, showing this information via web browser is very effective. We can see some renowned newspapers using visual data representation. However, this process can be improved more. This total process is taken as a prototype of information extraction and visualization. Now only just the raw data, which are in a text file format have been used. It can be made more dynamic by processing raw text like accessing text from web automatically. English text is only used in this process but in further improvement, language Bengali can also be added. One significant thing is that, only road accident related information is extracted and visualized because of making people concern about road accident in Bangladesh, but there are so many important news topic like road accident about which people should also be made concerned. Educational news can be included as like road accident news. It is really important in the third world country like Bangladesh to be conscious about death statistics in each day. This total process can be utilized for social life and it can be initialized from our social responsibility.

Some problems faced when evaluating the extracted information. This problem can be resolved in further improvement of this process. Adding human resource can be very helpful.

7. CONCLUSION

A strategy of information extraction and data visualization has been described in this paper. The raw text was collected from some newspapers in Bangladesh. Extracted information saved in a csv file and automatically convert to visual presentation. Interactive data visualization helps people to understand data easily, let them take actions on the data, give comfort to their eyes as data visualization is a combination of art and science. This is just a prototype of extracting information and visualized data interactively in web browser. In the future, this process can be extended to a deeper level. Some new news topic can be added for information extraction for making people concern.

8. ACKNOWLEDGEMENTS

Special thanks go to Nazmus Saquib for his kindly guidelines; and especially for providing very useful and timely suggestions. Without his insight, resources and information, it would have been very difficult to write this paper.Finally, I also want to thanks to my family and my friends for their inspiration to move forward.

9. REFERENCES

[1]. Interactive Data Visualisation, “Interactive Data Visualisation”,https://en.wikipedia.org/wiki/Data_visualization, September 10, 2015.

[2]. Natural Language Processing, “Natural Language Processing”, https://en.wikipedia.org/wiki/NLP, September 10, 2015.

[3]. Richard Johansson, Anders Berglund, Magnus Danielsson, Automatic Text-to-Scene Conversion in the Traffic Accident Domain, Lund University.

[4]. Steven Bird, Ewan Klein, Edward Loper, Natural Language Processing with Python , Highway North, Sebastopol, O’Reilly Media, 2009.

[5]. Steven Bird, Ewan Klein, Edward Loper, Natural Language Processing with Python , Highway North, Sebastopol, O’Reilly Media, 2009.