Data Analysis to Determine Business Effectiveness Based on User Reviews And Ratings

Anonymous

ABSTRACT

1. OVERVIEW

Nowadays online search has become one of the most important sources of our information. This information is an accumulation of user generated experiences which we call reviews. We can find reviews of a lot of items, which we use in daily life, from food products to clothes to services, which are provided various business establishments. When it comes to local business, most of the people check reviews to find out about the quality of service provided to the customer from the reviews. Services provided by companies like Yelp, Foursquare, Urbanspoon, Zomato etc. are crowd-sourced reviews about local businesses. These reviews enable fellow users to make a final decision depending on the task, which can be choosing a restaurant to dine, finding a local business. The other side i.e. business side can benefit from such service too as it helps them improve their service or product.

What if a business has not been established and we need to find out people's view about a certain business in a given location and to take a decision about the feasibility of opening up a local business? One of the important questions of the feasibility of the business idea is assessing the competitive advantage and this can be determined by finding people's opinion about the establishment. People's reviews can help a business provide services that can help them differentiate themselves from the competition. With the help of review data one can read up on the negative reviews and see the current demands of people that are lacking in the establishment, which can be used to further improve the services of the business. Companies mentioned above provide such services to the business and Yelp is one of the big players in the market who provided data analysis to various businesses to help them make the appropriate business decisions in future. For academic purposes, Yelp has provided students with large data sets to analyze and we have decided to make use of it in our project [1]. The analysis would involve; clustering of the business rating given to the business, reviews and tips for that business using rating given by the reviewer and parsing through the review text to identify if the review is positive or negative or neutral [1] that considers the user's view towards the business as an important factor in rating it.

The project's main challenge is to combine the analysis of reviews and tips by the reviewers to provide a specific conclusion.

The data set which we have provides a large variety of data with possibilities of obtaining numerous kinds of information from the dataset. We had to narrow down our decision to selecting one such data mining process where we also finalized to do analysis for local business establishment process. The fact that we could use the consumer's data to benefit the businesses was more lucrative to us.

2. DESIGN CONSIDERATIONS

2.1 JSON

For our dataset, we have chosen the data from Yelp dataset challenge website. The data obtained by us is presented in a JSON format which is semi-structured. Our first challenge was to convert this semi-structured data into a structured RDBMS form, one that would be more appropriate for us to mine or analyze using tools such as R and Rattle. We first looked into converting the data set into CSV format so that it could be easily read by R / Rattle tool but then we decided to learn some of the connecting with Java/JDBC and our MySQL server to parse through the JSON data and store it in our MySQL.

2.2 MySQL

MySQL is the open software for handling of relational databases. We had to deal with data which had a huge number of entries in the form of unstructured JSON format, about a million reviews from 3,00,000 users. Our project required optimized search and RDBMS becomes a natural choice as it can handle large databases. MySQL being open source became our preferred tool to handle such data. As it uses standard query language it is also widely supported by number of data mining tools including R. Using MySQL we created a database called "yelp" and created a user called "sqluser". We granted this user all privileges so that we could access the database via our JDBC.

2.3 Java/JDBC

Java is an excellent platform for us to write our code. For one it's highly customizable, we can take exactly the data we

need and leave out sections that wouldn't be useful for us. We imported the necessary libraries which were the JSON and MySQL connector libraries. This helped us connect to our created database mentioned in the earlier section. We created two files - a driver and a parser. Our driver files connect to the respective DB and our parser file parses through the JSON data and stores it in its proper columns in our RDBMS. The important part of the whole loading process was to design the RDBMS schema and it took us quite a while. We had a number of JSON files so we all got together a number of times to discuss and polish our RDBMS schema. Once decided, we implemented the same structure on our Java code to populate the SQL tables accordingly.

2.4 R and Rattle

R being a powerful statistical tool and also used for data mining. The rattle is GUI which sits on top of R and provides better usability. R can be easily connected a MySQL server with RMySql package. This package provides full functionality to use the SQL and obtain the results from the database. The next step is sentiment analysis of the reviews. For this, we shall be using the method proposed by Jeffrey Breen [2] which uses Bing Liu and Minqing Hu's [3] opinion lexicon words to find the polarity of the text. Based on the results obtained from these methods applied on the reviews we can use R for data visualizations.

3. ARCHITECTURE

Here we are loading the Yelp JSON files into MySQL DB.The JSON files contain a non-flat structure, so we need to re-format the data and convert them to flat structures that we'll insert into the DB tables.We are also creating validation tables from the data available so as to maintain the integrity of the data in the DB.

The project has been divided into following steps:

Data Preparation: This is the phase which we have completed the interpreting of the data, converting it from JSON to RDBMS MySql format, these are data retrieval and data loading stages. Then we validated and clean the data by removing the values which are irrelevant. The intermediate stage in this was the development of ER models and designing of table structures before loading and populating the data.

Data Mining, Analysis and Visualization: This is the phase where we will be working on packages which we have installed for sentiment analysis and retrieving the sample results after the analysis. The data mined from MySQL database will be used to perform sentiment analysis on the results will be displayed on the R console.

Interactive Interface: This phase will be implemented looking at its feasibility and will have an interface to display the results based on the inputs. This will take the data which is output from the R and display it accordingly.

4. IMPLEMENTATION

The data that has been provided to us by Yelp has information about businesses in four continents and 10 cities. The number entries for users data is close to 3,00,000. The database provided contains 5 datasets namely business, checkins, reviews, tip, and users. The data which we need would be location and category of the business. Then we will need to analyze the reviews of the similar category in the area and we will mine the reviews from these businesses. These reviews will be used to find an overall sentiment of the business in the area. Of course, we will have to modify this to include greater details with features of providing top 5 business and their reviews. Also, we need to identify the negative comments and provide a sample data of those negative comments to the business for analysis which helps them to find more about the problems associated with the business in that area. Next thing which we need is to find some general tips given by users about businesses to further help them relate to users needs.

The technologies which we found helpful are R for data mining; this is connected to a MySQL database. To achieve the conversion of the available dataset, which is in JSON format, to an RDBMS we used Java to process the data. R and Rattle shall be used as the tools for data visualization and sentiment analysis. For visualization, ggplot2 and for sentiment analysis sentiment packages shall be used. We chose MySQL because it provides us with scalability and integrity of data along with good security. Even if this data doesn't contain any information about the users, but we need to consider the security of the dataset to maintain the confidence in this model. The tools and packages chosen here go in perfect synchronization. We shall also be adding new packages as in the next phase as the project advances.

We have set up our MySQL servers as well on our mac systems using MySQL homebrew. It took us a while to get past some permission and other errors, but we've got past all that now and it's working fine for all of us. We've also finished writing our java codes to push data into 3 databases after parsing through the 3 JSON files.

Using a JSONObject in Java we can use its pre-defined API like "get(String)" to pull out necessary information from our JSON file. Once we get the string we can push the same values into a "PreparedStatment" object using "setString()","setLong()" etc and finally update the values into the MySQL DB using the "executeCommand()". We have set-up 3 databases so far, the first is the "business" database, the second is "review" database and the third is "users" database. Our "business" database comprises of five tables. They are business, business hours, business category, categories, and states. For two of the tables, we use business_id as our primary key and category name and state abbreviation as primary keys for two other tables. Our business category table will have a manytomany relation and hence would not have a primary key associated with it. Our "review" database will have just one table for now and our "users" database will have two tables. Both the tables will have user_id as the primary key. Our relational scheme is as shown in figure 1.

5. LESSONS LEARNED

We were going to use CSV to read the dataset but realized that it's won't provide the speed required for mining of the dataset. We also learned that instead of trying to put in too many features and doing different data analysis we should concentrate on one which we want to focus the most. We also learned that the some requirements of the project may change and we will have to take that into consideration like we did in the case of the using JDBC instead of CSV files. We also realized as research for the development of this project proceeds we will be adding new tools and packages to achieve our work.

6. CURRENT STATUS & FUTURE WORK

6.1 Tables, Figures, and Citations/References

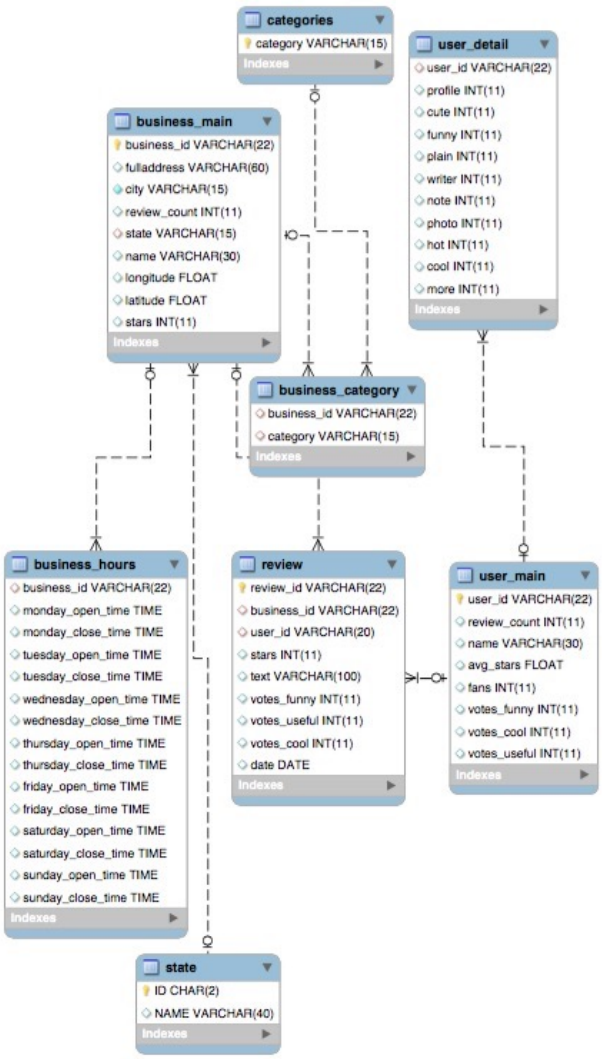

Figure 1: Relational Schema

6.2 Future Work

Next we'll need to look into the following points. We would have to load all the datasets obtained form yelp. Once we've loaded all the datasets we can proceed with analyzing the data. We're interested in performing sentiment analysis on the text reviews.

7. REFERENCES

[1] Yelp Dataset Challenge. `http://www.yelp.com/dataset_challenge/`, 2015.

[2] J. Breen. Twitter text mining R slides. `https://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/`, 2011.

[3] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM.