Algorithms for Data Guided Business Intelligence

Community Detection using Attributed Graphs & Sentimental Analysis

By:
Udit Deshmukh(udeshmu)
Sachin Ahuja(sahuja3)
Jitin Kumar(jkumar3)

(1) Yelp dataset examination and business value assessment:

Yelp dataset consists of files containing business, user, user review, check-in, tip, and photographic data. Each file consists of one JSON record per line. These records and their business value assessment is explained below:

1. **yelp_academic_dataset_business.json** :
   This file contains business like id, name, neighborhood, address, stars, no. of reviews etc. This data can be used to find basic recommendations like top few restaurants can be selected using star ratings or we can use location to filter top rated ones by location.

2. **yelp_academic_dataset_review.json** :
   This file contains review data like review id, text, stars, and categories as useful, funny, cool etc. Using review data, we can identify fake reviews by analyzing no. of reviews from a person and their type. We can also analyze how useful the reviews from a user are. Categorizing review text can give basic sentiment analysis which can further be used for recommending a restaurant.

3. **yelp_academic_dataset_user.json** :
   This file contains user id, name, review count, categorical data like useful, funny etc., average stars, his friends, elite, and compliments data. Based on this data, we can extract recommendations based on user's friend network and can also do basic sentiment analysis based on funny, elite, useful categories. This can further be used to find key personalities (based on fans, friends network etc.) which could impact the business a lot.

4. **yelp_academic_dataset_checkin.json** :
   This file contains check-in data like time of check-in and business id where user checked-in. This data can be used to extract knowledge like which restaurants a user frequently visits. We can further analyze patterns like how frequently that person goes to these restaurants and suggest similar options.

5. **photos**:
   This file contains encrypted photo id, business id, caption, and categorical labels if any.

(2) Sentiment Analysis/Recommendation System Use Cases:

Few of the use cases where Yelp dataset can be used for Sentiment Analysis or Recommendation Systems are as follows:

1. **Finding the rating comparing with actual rating:**
   Based on user reviews, Sentiment analysis can be done and we can calculate actual rating of the restaurant. This can be useful in following cases:

a. We can compare the actual rating of a restaurant with its calculated rating
b. Actual rating might not be re-calculated very frequently. Say, we it is calculated once very quarter, then actual rating will not correctly represent the current rating. In this case, newly calculated rating can give better estimate.

2. **Filtering fake reviews:**
   Fake reviews can portray an image about a product that is far from the truth. It is increasingly becoming a major problem because one cannot be sure whether a review truly represents a sentiment or it is merely a paid one. We can filter the fake reviews as follows:
   a. Check the no. of reviews posted by a reviewer and for which restaurant. For example, if user A has always posted good reviews about restaurant R but never posted any other reviews, or posted very bad reviews for many other restaurants, then his reviews might be fake
   b. We can check the reviewer's stars and other categorical data like funny, useful etc. to analyze the significance of his reviews.

   Once we filter fake reviews, we can get better idea about the restaurant from true reviews.

3. **Analyze current trend for top restaurants:**
   We can analyze the check-in data for top 50 restaurants in a city. Then we can do the following analysis:
   a. Find the restaurants where no. of check-ins has increased/decreased to great extent in past weeks/months.
   b. For these restaurants, we can filter reviews and do sentiment-analysis.
   c. Based on this sentiment-analysis, we can figure the reason for the decline in no. of check-ins. Say restaurant A's check-ins have decreased a lot, and sentiment analysis shows that A's service is not as good as earlier, then we know the major issue for reduced no. of check-ins could be poor customer service.
   d. We can notify the problem and suggest the solution to the restaurant A to improve their performance.

4. **Finding influential people and community detection:**
   Based on user's friends data, we can use community detection to find the network and then use influence propagation to find the most influential people in a group. We can then target these influential people for any product/services which will be more likely to create an impact over that community.

(3) Most recent R&D advances in Sentiment Analysis and Recommender Systems:

Sentimental Analysis from Multimodal Content:
Multimodal sentiment analysis uses texts, audio, and video data as source of information for sentiment analysis. It employs both feature and decision-level fusion methods to merge effective information from multiple modalities. Sentiment analysis from facial expressions has been researched upon since 1970s, speech based emotional analysis focuses on pitch, intensity of utterance, bandwidth, and duration. Best accuracy achieved so far is 81% on Berlin Database of Emotional Speech(BDES) using two step classification approach and a unique set of spectral, voice features, selected with Sequential Floating Forward Selection algorithm.
Sentiment analysis from textual data is currently mostly based on rule-based techniques like Bag of Words using large sentiment or statistical approaches that rely on large datasets.
For visual analysis, a video is segmented into several parts and then converted to images based on frame rate. Then facial features are extracted from each image and average is computed to get final feature vector. Similarly, audio and text features of video are extracted and fused together. Using supervised classifier, we identify overall polarity of each video segment. [1]

Measuring Term Weights for sentiment analysis:
Current work in sentiment analysis is mostly data-driven and lexicon based. This paper introduces how to represent and evaluate the weights of sentiment terms. The characteristics of good sentiment terms can be prominence, topic-relevance and sentiment analysis using collection statistics etc. Topic-independently, a good sentiment term is discriminative and prominent such that its appearance poses greater influence on judgement of analysis system. Topic-dependent terms are terms relevant to topic of the text. These can be regarded more useful than extraneous terms. Opinion Retrieval Model is then used to find set of documents that are relevant to given topic and then, sentiment analysis is performed on these topics. This gives a list of topically-related and opinionated documents. The best performing feature of word generation model is VS giving 4.21% improvement over baseline f-measure.[2]

Collaborative Filtering for Implicit Feedback Datasets: Recommender systems are based on the idea of helping users select a product/service based on user's data. In first approach, user's data is matched with the product data and relevant relationship is extracted. For example, if a user likes sci-fi movies, then sci-fi movies are selected from movies category and then recommended to the user. This approach is called Content Filtering. Major problem in this is we don't usually have the user's and product's content data.
Another approach is **Collaborative Filtering** where transaction history is used instead of content attributes. For example, if user U1's transactions show that user purchased sci-fi movies more often than any other genre, then we can assume that user likes that genre and then we can

recommend other movies from similar genre to the user. This is called collaborative filtering because past data is used instead of direct user's preferences.[3]

(4) Target Question we address in this project:

The concept of identifying the most influential people plays a major role in improving the services of any business. We can expect one person to have far more influence over a community than the other. We can use this as follows:

a. We can do community detection to find communities from our data. To find communities, we can use friend's data from user.json. This will give us a network of connected people.
b. Based on this community, we can measure influence propagation tests and observe the propagation.
c. Based on influence propagation, we can find the most influential people in a group. This will give few key people in the group which affect the performance of the restaurant directly/indirectly. For example, say user U1 has 200 followers, and influence propagation shows that starting from U1, influence propagation to other nodes is fastest, we can identify U1 as key person.

(5) Business Value of finding most influential people:

Finding most influential people can be used in many ways as follows:

a. We can give special treatment to more influential people. It will be like MVP card for pet users, just that user will not even have to carry a card and will be able to enjoy the privileged services.
b. These privileged services will most likely leave very positive influence which again, based on influence propagation give advantage by influencing the whole community.

(6) Evaluating quality of different answers:

Our goal is to find strong communities based on the users' friends as well as their attributes. A strong community has the property that the rate of influence propagation within the community is high. From a business perspective, we can target the people who have large number of connections within a strong community and provide higher quality services to them. We then see how this affects the other people in that community. If the other people start giving positive reviews as well or the revenue increases substantially, we can say that the quality of community is good and we were successful in finding such communities.

(7) Research papers relevant to our business problem:

Community Detection Based on Structural and Attribute Similarities: The study of social networks has been a central research topic in the recent years. The biggest challenge of social network analysis is to identify communities in the network. A community is defined as the group of users in the social network who interact with each other more frequently than with those outside the group. If identified accurately, these communities can guide us in finding solutions to business problems such as making recommendation systems, targeting potential customers/influencers to improve business. In the past years many algorithms have been proposed to solve this problem but all of them only took the structural features into account. But the attributes of users such as their age, location, ratings, reviews etc. In this paper, the authors have introduced two algorithms to partition a graph with attributes into communities so that the nodes in the same community are densely connected as well as homogeneous.
In the first algorithm Structure-Attribute Clustering Algorithm (SAC1) they used similarities between users based on their attributes and gain in modularity as the basis for their algorithm. The started by assuming each user to be a community and slowly increases the communities by combining different communities based on similarity and modularity gain. The second algorithm was just a small change to the first one to optimize it. In this algorithm, instead of taking each node into consideration with every other node to compute modularity gains, the constructed a k nearest neighbor graph for each node and took only these k nearest neighbors into consideration.[4]

Sentiment classification using Machine Learning Techniques:
Here, we consider classifying documents not by topic but overall sentiment. A lot of research has been done in the field of Text categorization, most of which focuses on sorting documents based on subject matter. This paper however focuses on overall sentiment analysis which can give succinct summaries to readers. We examine effectiveness of machine learning techniques to sentiment analysis problem, which is different from traditional topic-based classification because sentiment can be very subtle. Three algorithms namely Support Vector Machines(SVM), Naïve Bayes and maximum entropy classification are used for experiments using Bag of Words framework. In terms of results, Naïve Bayes tends to do the worst and SVM tends to do the best, but the difference isn't large. Moreover, achieved accuracy using sentiment classification is quite poor compared to topic-based classification despite of several types of features used. Further, Unigram presence information turns out to be very effective.[5]

Twitter Sentiment Analysis Using Distant Supervision:

In this paper, the authors have discussed the importance of performing sentiment analysis on twitter tweets. Doing sentiment analysis on these tweets can provide us with the information about the overall view of audience on a topic that is trending on twitter. The authors used

different models such as unigram, bigram, part of speech, etc. to extract features out of these tweets. They then performed various machine learning algorithms in these tweets to categorize the tweets either as positive or negative. They then discussed which model performed better and which algorithm gave the highest accuracy and why.[6]

(8) Features used from Yelp dataset:

We used yelp_academic_dataset_review.json and yelp_academic_dataset_user.json from Yelp dataset as follows:

a. From review.json, we used continuous data(useful, funny, etc.), user id, stars and text
b. From user.json, we used name, review count, friends data and other continuous data like funny, useful, elite etc.

(9) Solution Prototype:

Our approach is mainly comprised of these three steps:

(a) **Reducing the number of Users:** The Yelp Dataset originally consisted of around 1 million users. So, to reduce the number of users, we picked only those who have given more than 500 reviews. We filtered out these users and also found their friends. Both the users who have more than 500 reviews and their friends were taken as target users.

(b) Performing Sentiment Analysis: From yelp_academic_dataset_review.json file we filtered reviews of those users found in step-(a). After cleaning the data we used doc to vec method to calculate features for these reviews. We then built naïve bayes models to find the overall sentiment for each user.

(c) **Making Attributed Graph and Applying Community Detection:** We extracted the friends of users found in step-(a) from yelp_academic_dataset_user.json and created an edge list so that we can make a graph out of it. We also extracted the attributes of these users using the same file and also appended the sentiment score found in step-(b). The rationale behind adding the sentiment score of each user as an attribute is that every user has his own way of reviewing the same thing. If the sentiment score of two user is similar, it is highly likely that these users have a similar way of reviewing things. We made a csv file containing these attributes after discretizing them. We then applied the SAC-1 algorithm to find the communities using the structure as well as attribute data of the graph.

We can observe that the final communities obtained were densely connected. From each community, we can find the users that have the most number of friends in that community. We can pick 25 such customers in each community and offer them special services and promotional offers. As these users have the most number of friends in the community they can be seen as the most influential users and these users can influence other users to go to a particular hotel if

they have a positive experience with the same place.

References:

[1] Poria, Soujanya, et al. "Fusing audio, visual and textual clues for sentiment analysis from multimodal content." *Neurocomputing* 174 (2016): 50-59.

[2] Kim, Jungi, Jin-Ji Li, and Jong-Hyeok Lee. "Discovering the discriminative views: measuring term weights for sentiment analysis." *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 2009.

[3] Hu, Yifan, Yehuda Koren, and Chris Volinsky. "Collaborative filtering for implicit feedback datasets." *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. Ieee, 2008.

[4] The Anh Dang and E. Viennet, *"Community Detection based on Structural and Attribute Similarities"*

[5] *Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.*

[6] *Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1.12 (2009).*