CMPS 242 - Machine Learning (Final Report)
Analysis of Rating Dimensions Using Review Text

Bharath Nagesh *bnagesh@ucsc.edu*
Madhu Shivashankaraiah *mshivash@ucsc.edu*
Numra Bathool *nbathool@ucsc.edu*
Tanay Parekhji *tparekhji@ucsc.edu*

Git repository link : `https://github.com/madhusedu/topic_modeling.git`

Abstract

Modern day systems take in both ratings and review texts from the user but generally consider only the ratings for performing different tasks, such as predictions and assessing the overall quality of a business. This according to us is not the optimal method. Ratings are highly subjective and what might be considered as favorable by one could be unfavorable for another. After a detailed observation of the review texts on IMDb, Zomato and Yelp we noticed that users implicitly provide information on their likes and dislikes. We plan to implement a system which takes in not only the rating but also the review text in order to form an opinion. We consider all the reviews of the users and then use data mining to summarize them. We also classify some reviews belonging to the same topic. form an opinion on the preferences of the users, then finally use those opinions in order to improve the overall rating quality for the business and perform better user predictions.

1. Problem Statement

Along with the expansion of the internet, a number of services such as e-Commerce and Review websites have emerged. A lot of people prefer using these services because of many factors, some of them being :

- **Reliability :** These services provide excellent delivery of product/information and then back it up with a good level of customer care. Also, the ratings provided on these sites aren't just a number that is just floated to hoodwink the customer into buying the product/using the service making them very reliable

- **Ease of use :** All these services have a clean web/app interface which makes it very convenient for the people to use them.

The effect of the above factors can be seen in the growth of the industries, as indicated in Figure. 1 and Figure 2.

During holidays such as Thanksgiving the use of these services increase dramatically. The usage statistics is indicated in Figure 3.

The general method used to increase the reliability of a product/business/service is by collecting ratings reviews from the people who have used them. This facilitates future customers to form a better opinion of the product when deciding to whether or not use the product.

After a thorough observation of ratings and reviews provided on websites such as IMDb, Yelp, Amazon the following was observed. The overall rating for any product is formed as a weighted average of the ratings provided by all the users. This method completely disregards the review text provided by many of the users. The users generally provide important information which can be used to improve the overall rating of the product. For an example of this situation, we can consider reviews provided for a restaurant. An user may provide a rating of 3 stars for the restaurant and then provide a review which makes the restaurant actually worth 3.5 or maybe even 4 stars.

This problem that is presented above provided us with the opportunity to develop a system which uses Latent Dirichlet Allocation and Topic Modeling techniques in order to understand the reviews provided by the user. After doing so we attempt to develop an application which fine tunes the overall quality of ratings for any given product/business.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

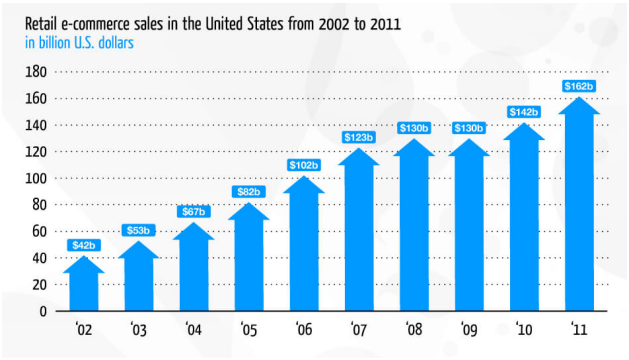

Figure 1. Retail sales in the USA between 2002 and 2011.

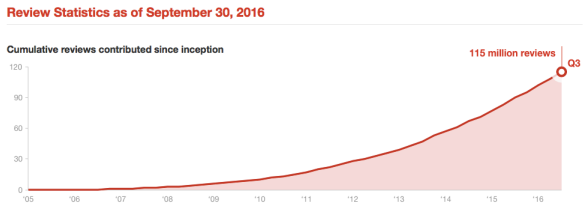

Figure 2. Yelp Review Statistics

2. Feature Engineering

Enter information on the model features and explain variations ...

3. Algorithm Formulation

Explain the algorithm...

4. Evaluation

Enter the evaluation results and parameters used for the evaluation

5. Results

Enter the results obtained after performing the evaluations...

6. Deliverables

Enter all that could be done in case more time was available...

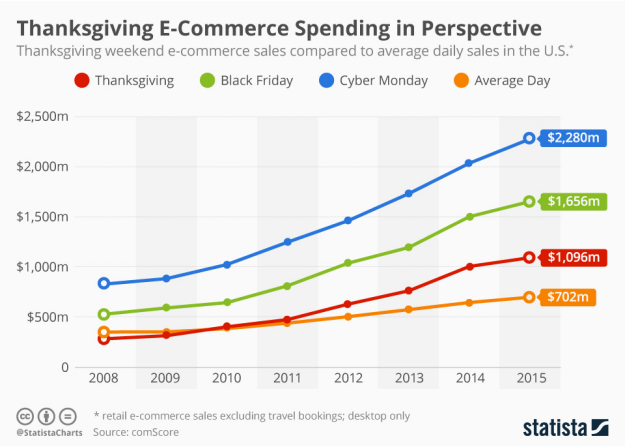

Figure 3. eCommerce sales during the Thanksgiving week.

References

[1] A. A.Yeung. *Matrix Factorization: Implementation in Python*. 2010.

[2] T. L. Griffith, M. Steyvers. *Finding Scientific Topics* PNAS, vol. 10, 6 April - 2004.

[3] Ke Zhai, Jordan Boyd-Graber. *Online Latent Dirichlet Allocation with Infinite Vocabulary* In proceedings, *0th International Conference on Machine Learning*, Atlanta,GA, 2013 JMLR : WCP volume 28. Copyright 2013 by the author

[4] D .M. Blei, A. Y. Ng, M. I. Jordan. *Latent Dirichlet Allocation* Journal of Machine Learning Research. January, 2003

[5] Blei, D. M., Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B. *Hierarchical topic models and the nested Chinese restaurant process* In Advances in Neural Information Processing Systems 16. Cambridge, MA, USA: MIT Press. 2004

[6] Erosheva, E. A. *Grade of membership and latent structure models with applications to disability survey*

data. Unpublished doctoral dissertation, Department of Statistics, Carnegie Mellon University.

[7] Hofmann, T. *Probabilistic Latent Semantic Analysis.* In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.

[8] Zhuang.L, Jing.F, Zhu.X. *Movie Review Mining and Summarization.* CIKM?06, Arlington, VA, USA, 2006.

[9] Hu.M, Liu.B *Mining and Summarizing Customer Reviews.* KDD?04, Seattle, WA, USA. 2004

[10] McAuley.J, Leskovec.J *Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text* In Proceedings of the 7th ACM conference on Recommender systems, Hong Kong, China. 2013