POPULARITY PREDICTION FOR SINGLE TWEET BASED ON HETEROGENEOUS BASS MODEL

[1]Vidhya Shenigaram,[2]Katakam Krishna Chaitanya,[3]Pallavi Bhramarautu,[4]Mosheck Menta

[1,2]Assistant Professor,[3,4]Associate Professor

Department of CSE

Kshatriya College of Engineering

Abstract

Predicting the popularity of a single tweet is useful for both users and enterprises. However, adopting existing topic or event prediction models cannot obtain satisfactory results. The reason is that one topic or event that consists of multiple tweets, has more features and characteristics than a single tweet. In this paper, we propose two variations of Heterogeneous Bass models (**HBass**), originally developed in the field of marketing science, namely Spatial-Temporal Heterogeneous Bass Model (**ST-HBass**) and Feature-Driven Heterogeneous Bass Model (**FD-HBass**), to predict the popularity of a single tweet at the early stage and the stable stage. We further design an Interaction Enhancement to improve the performance, which considers the competition and cooperation from different tweets with the common topic. In addition, it is often difficult to depict popularity quantitatively. We design an experiment to get the weight of favorite, retweet and reply, and apply the linear regression to calculate the popularity. Furthermore, we design a clustering method to bound the popular threshold. Once the weight and popular threshold are determined, the status whether a tweet will be popular or not can be justified. Our model is validated by conducting experiments on real-world Twitter data, and the results show the efficiency and accuracy of our model, with less absolute percent error and the best Precision and F-score. In all, we introduce Bass model into social network single-tweet prediction to show it can achieve excellent performance.

1. Introduction

TWITTER, centered by users and communications, is one of the best-known social networks in the world. In recent years, research on prediction in social networks has received increased attention from both academia and industry. Many items in social networks are worth predicting, such as user's personality [1], popular stories [2], and interesting events [3]. Even a film's box office [4] and stock trend [5] with little relevance to social networks can be predicted through the contents posted by users. Our goal is to quantitatively predict the popularity of a single tweet at any given time during its life cycle. Meanwhile, we also want to make a qualitative prediction by classifying a tweet as popular or unpopular. This work is significant for both ordinary users and companies. For users, it provides a tool to help them filter through a large amount of new content in order to identify interesting items in a timely manner. Another important application of this work is to help companies seize the opportunity to lead and generate a trend or hot topic. Lastly, abnormal popular tweets can set alarm for disaster, crime, or catastrophe. For Example , Face book is helping to catch criminals. Sometimes, the suspect inevitably brags about his deviant behavior on the social networks, which captures the user's attention. The police can get tips from this abnormal popular content. In addition, there are also some negative cases, such as "Toyota brake". It could have avoided car accidents if the prediction of explosion was adopted, which would help the company to recall the products earlier.

Majority of existing work focus on predicting the popularity of topics or events, which consist of sets of single tweets. However, there are a few works that focus on predicting the popularity of a single tweet. Predicting a single tweet is a more challenging task because it only uses its own textual information and user's information with a timeline. Moreover, the lifespan of a single tweet is often shorter than a topic or event. Most of the tweets are out of sight before long, which lacks enough time to compare the prediction with real trend.

Several works related to single tweet [6], [7], [8], [9], [10], [11] mostly concentrate on qualitative prediction. They predicted whether a tweet will be retweeted, which is a 2-class classification problem. In addition, the prediction of the tweet's trend can be categorized as a regression problem. However, the accuracy of regression models are often unstable, because regression models cannot capture the randomness that exists in the trend after a tweet has been posted. On the other hand, several works related to feature based methods [2], [12], [13], [14], [15], [16], [17], [18], [19] relied heavily on the most effective features for estimating popularity. Meanwhile, time-series methods have a better performance than feature based methods overall. Time-series based methods relied on statistical models or point process based methods, and those models are widely used in recent years. However, time-series based methods also have many drawbacks. For instance, Q. Zhao et al. [20] and S. Mishra et al. [21] built a model based on the random point process to predict the trend of a tweet. However, it requires hard-collected features, such as the post time of all retweets related to the original tweet and the number of followers of each retweet user. Most of the above works use traditional machine learning methods or random point process methods. Nevertheless, those methods [20], [26] overlooked human activities that may contribute to popularity, such as favorite or reply. Furthermore, J. Xin et al. [27] introduced the Bass model into flickr prediction. However, they did not incorporate th characteristics of the social networks.

In this paper, we design the Heterogeneous Bass model (**HBass**) which contains two varieties, namely Spatial-Temporal Heterogeneous Bass Model (**ST-HBass**) and Feature-Driven Heterogeneous Bass Model (**FD-HBass**), to predict the popularity of a single tweet. The Bass model [28] is one of the most widely applied models in management science. It is originally used to model the sales of a newly put-on market product to a group of people, which can predict the popularity of a newly posted single tweet in a social community. [56] created 12 features for each topic and forecasted topic popularity using standard bass diffusion model in Twitter. However, it is not immediately clear the Bass model is applied to single tweet prediction. First of all, it is hard to add the features of social network with only two parameters. In addition, the Bass model assumes spatial and temporal homogeneity, leading to no distinction of individuals.

To overcome the limitations of the Bass model, we incorporate Twitter features into the original model to form the HBass, and relax it to individual-level heterogeneity. In particular, we propose two Bass models based approaches, ST-HBass and FD-HBass. ST-HBass focuses more on the spatial and temporal heterogeneity, while FD-HBass focuses more on the effect of different features. As a result, FDHBass has better performance than ST-HBass in quantitative prediction and qualitative prediction. In reality, a series o tweets may concentrate on one common topic. Those tweets often influence each other's popularity. Considering the interaction between those tweets, we propose an Interaction Enhancement to measure the competition and cooperation from different tweets with the common topic. Furthermore, the performance of HBass is improved, which forms two variations, the STI-HBass and FDI-HBass. As a result, FDIHBass has the best performance.

Moreover, compared with many previous works, we redefine the quantitative definition of popularity and the threshold to classify popular and unpopular tweets. We adopt a linear regression model to calculate popularity by combining favorite count, retweet count, and reply count as features. Instead of choosing the threshold by experience, we use DBSCAN and k-medoids clustering algorithms to get the median of two cluster centers, and multiply the indicators' weights respectively to get the threshold. Hence, we get the popularity and threshold from real datasets.

We focus on predicting the trend of a single tweet in all its life cycle including early stage (1h-24h after posted and stable stage (25h-240h after posted). We further justify whether a single tweet will be popular. Our model does not need a large training set like traditional machine learning methods. In addition, we do not require the topology of the social networks such as following and follower relationships. Generally, our model is not only suitable for predicting a single tweet's popularity, but also suitable for other social networks, which have asymmetric following follower relationship, such as Weibo and Digg.

With real-world Twitter dataset, our experimental results validate the efficiency and accuracy of our model to predict the trend, with less absolute percent error and closer the reality overall. In addition, our models have best Precision and F-score, which shows that our models have a better classification detection.

We summarize our contributions as follows:

_ We incorporate Twitter features into the Bass model in social network single-tweet prediction to form the HBass model. In addition, HBass has two variations, namely **ST-HBass** model, which focuses on spatial and temporal heterogeneity, and **FD-HBass** model, which focuses on the effect of different features. To be specific, we aim to predict the trend of a single tweet, and whether the tweet will be popular in the end.

_ We propose the Interaction Enhancement to consider the real situation that the different tweets with common topic have the interaction of competition and cooperation between each other.

_ We redefine the quantitative definition of popularity that combines the relationship among favorite, retweet, and reply, and threshold to classify popular and unpopular tweets based on clustering method, instead of choosing the threshold by experience.

_ We use real-world Twitter data to examine the efficiency of HBass. The simulation results show that the efficiency and accuracy of the quantitative prediction with less absolute percent error and the qualitative prediction with a better classification detection.

2. Literature Survey

The most significant step in any kind of study is the literature survey. We want that to review the papers that are formerly used in our domain that we are working on before we start developing and we can predict or generate the disadvantage on the basis of study and using the references from previous papers as a guide. In this segment, we momentarily audit the connected work on popularity prediction and their different techniques and also compare methods or techniques advantages and limitations or future work.

Paper 1: XiaomingChen and et al had proposed [1] to used Multivariate Linear (ML) and ModelSzabo-Huberman (S-H) Model which gives better output. The advantage of this model are used for considering possible web content popularity based on recorded data provided by effective popularity measures. It is feasible to utilize diverse ubiquity expectation models for each example. That can prompt decreased forecast mistakes due to investigating the contrasts between designs in a more unequivocal structure.

Paper 2: Sirisup Laohakiat and et al had proposed [10] the twitter fame profile mining and expectation effectively. The proposed structure not exclusively can perform twitter ubiquity profile mining and expectation proficiently, yet additionally can be applied to mine the arrangements of any time arrangement. Later on, to improve the structure to consider both the substance of the posts just as the fleeting profile to make more precise expectation.

Paper 3: Przemys law Rokita and et al had proposed [11] to used Support Vector Regression to foresee the fame of online video content estimated as the quantity of perspectives. Later on work is to broaden the arrangement of highlights utilized for expectation by adding more semantic signs, like video point or the assumption of the social communications, to more readily comprehend.

Paper 4: Cheng-Te Li and et al had proposed [4] to used personalized self-exciting point process Model (PSEISMIC)for oneself energizing point cycles to build up a factual model, PSEISMIC, which prompts exact expectations of the personalized self-exciting point process Model doesnt deal with the various tweets for ongoing.

Paper 5: Almeida and Marcos A. Goncalves and et al [5] generally based on the Multivariate Linear (ML) Model Szabo-Huberman (S-H) Model and these model are generally used for foreseeing the future prevalence of Web content dependent on verifiable data given by early fame measures. It is feasible to utilize diverse notoriety expectation models for each example. That can prompt diminihed forecast mistakes due to investigating the contrasts between designs in a more unequivocal structure.

Paper 6: Yeyun Gong and Xuanjing Huang and and et al is based [2] on the Deep Neural Network and the attention based deep neural network is used to fuse context oriented and social data for the undertaking. They utilized inserting to mirror the client, the interests of the clients advantage, the creator and tweet, separately. In this strategy, the undertaking of retweet forecast is just in binary classification.

3. System Analysis

Existing System:

The researches for content prediction include events, topics [3], topics [9], and single post [22], [6], [3]. Most of the content prediction paid attention to predict events or topics that a group of people created, not that created by an individual. These predictions tend to predict whether a topic or event will be popular or how popular in the future. All these topics or events models need extra tools to generate topics or events at the first step, and then use self designed model with machine learning methods to achieve their goals. In this table, we choose 6 critical metrics to describe the related works, and their meanings are listed as follows.

_ **Content.** Content is used to express the detail content type, such as text, video, image and media.

_ **Dataset.** Dataset shows the data sources in different social network of those works.

_ **Prediction Type.** Prediction type can be divided into Boolean and Numerical, where Boolean represents that works aim at qualitative prediction and Numerical represent that they are quantitative prediction.

_ **Reference Objective**. Reference objective represents the common method type for prediction. Generally, there are briefly two different methods to predict those situations, feature based methods and time-series based methods. As for feature based methods, many researches studied that the different features have a different effect on the popularity. Therefore, they always adopted the methods rely on data to find out the most effective feature for popularity. The feature-based methods have a moderate performance, which is stable during the peaking time. Several researches based on time-series approach have been proposed recently to predict the popularity. They tended to design the timeseries methods based on statistical models. The timeseries methods always improve greatly over time and produce quite satisfactory results.

_ **Methodology**. Methodology shows the basic methodology of those works.

_ **Feature**. If the method belongs to feature based methods or some special time-series methods, we will list the main features those works used.

Disadvantages

1) The system less effective since it is not implemented FD-HBass for large number of datasets.

2) The system doesn't implement Data Preprocessing and not compared with number of classifiers.

4. Proposed System:

The proposed system incorporates Twitter features into the Bass model in social network single-tweet prediction to form the HBass model. In addition, HBass has two variations, namely **ST-HBass** model, which focuses on spatial and temporal heterogeneity, and **FD-HBass** model, which focuses on the effect of different features. To be specific, we aim to predict the trend of a single tweet, and whether the tweet will be popular in the end. The system proposes the Interaction Enhancement to consider the real situation that the different tweets with common topic have the interaction of competition and cooperation between each other. The system redefine the quantitative definition of popularity that combines the relationship among favorite, retweet, and reply, and threshold to classify popular and unpopular tweets based on clustering method, instead of choosing the threshold by experience. The system uses real-world Twitter data to examine the efficiency of HBass. The simulation results show that the efficiency and accuracy of the quantitative prediction with less absolute percent error and the qualitative prediction with a better classification detection.

Advantages

1) The system is more effective due to presence of DBSCAN and k-medoids clustering algorithms.

2) The system designs the Heterogeneous Bass model (HBass) which contains two varieties, namely Spatial- Temporal Heterogeneous Bass Model (ST-HBass) and Feature-Driven Heterogeneous Bass Model (FD-HBass), to predict the popularity of a single tweet.

5. Conclusion

In this paper, we design the Heterogeneous Bass model (**HBass**) which contains two varieties, namely Spatial-Temporal Heterogeneous Bass Model (**ST HBass**) and Feature-Driven Heterogeneous Bass Model (**FD-HBass**), to predict the popularity of a single tweet. We also propose an Interaction Enhancement to measure the competition and cooperation relationship of different tweets with the common topic. We further design an clustering method to bound the popularity threshold based on real world dataset.

Our experiments use real-world Twitter data to validate the efficiency and accuracy of our model in quantitative prediction, with less absolute percent error. For qualitative prediction, our model gets the best Precision and F-score, which shows that our models have a better classification detection. To summarized, we introduce the Bass model into social network single tweet prediction, and we also show that it achieves great performance.

REFERENCES

1. J. Golbeck, C. Robles, and K. Turner, “Predicting personality with social media,” in ACM CHI Conference on Human Factors in Computing Systems (CHI), 2011, pp. 253–0262.

2. K. Lerman and T. Hogg, “Using a model of social dynamics to predict popularity of news,” in ACM International Conference on World Wide Web (WWW), 2010, pp. 621–630.

3. X. Zhang, X. Chen, Y. Chen, S. Wang, Z. Li, and J. Xia, “Event detection and popularity prediction in microblogging,” in Elsevier Neurocomputing, vol. 149, pp. 1469–1480, 2015.

4. S. Asur and B. Huberman, “Predicting the future with social media,” in IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010, pp. 492–499.

5. Y. Huang, S. Zhou, K. Huang, and J. Guan, “Boosting financial trend prediction with twitter mood based on selective hidden markov models,” in Springer International Conference on Database Systems for Advanced Applications (DASFAA), 2015, pp.435–451.

6. Q. Zhang, Y. Gong, Y. Guo, and X. Huang, “Retweet behavior prediction using hierarchical dirichlet process.” in AAAI Conference on Artificial Intelligence (AAAI), 2015, pp. 403–409.

7. Q. Zhang, Y. Gong, J. Wu, H. Huang, and X. Huang, “Retweet prediction with attention-based deep neural network,” in ACM International on Conference on Information and Knowledge Management (CIKM), 2016, pp. 75–84.

8. J. Bian, Y. Yang, and T. S. Chua, “Predicting trending messages and diffusion participants in microblogging network,” in ACM International Conference on Research on Development in Information Retrieval (SIGIR), 2014, pp. 537–546.

9. P. Cui, S. Jin, L. Yu, F. Wang, W. Zhu, and S. Yang, “Cascading outbreak prediction in networks:a data-driven approach,” in ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2013, pp. 901–909.

10. B. Suh, L. Hong, P. Pirolli, and E. H. Chi, “Want to be retweeted? large scale analytics on factors impacting retweet in twitter network,” in IEEE International Conference on Social Computing (Socialcom), 2010, pp. 177–184.

11. M. Jenders, G. Kasneci, and F. Naumann, “Analyzing and predicting viral tweets,” in ACM International Conference on World Wide Web (WWW), 2013, pp. 657–664.

12. R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," in Association for the Advancement of Artificial Intelligence International AAAI Conference on Web and Social Media (ICWSM), 2012, pp. 26–33.

13. A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev, "Prediction of retweet cascade size over time," in ACM International Conference on Information and Knowledge Management (CIKM), 2012, pp. 2335–2338.

14. J. Li, J. Qin, T. Wang, Y. Cai, and H. Min, "A collaborative filtering model for personalized retweeting prediction," in Springer International Conference on Database Systems for Advanced Applications (DASFAA), 2015, pp. 122–134.

15. A. N. H. Man and M. L. Khodra, "Predicting information cascade on twitter using support vector regression," in International Conference on Data and Software Engineering (ICODSE), 2014, pp. 1–6.

16. M. Yang, K. Chen, Z. Miao, and X. Yang, "Cost-effective user monitoring for popularity prediction of online user-generated content," in IEEE International Conference on Data Mining (ICDM),2014, pp. 944–951.

17. H. Zhao, G. Liu, C. Shi, and B. Wu, "A retweet number prediction model based on followers' retweet intention and influence," in IEEE International Conference on Data Mining (ICDE), 2014, pp. 952– 959.

18. R. Lemahieu, S. Van Canneyt, C. De Boom, and B. Dhoedt, "Optimizing the popularity of twitter messages through user categories," in IEEE International Conference on Data Mining (ICDM), 2015, pp. 1396–1401.

19. P. Bao, H.-W. Shen, J. Huang, and X.-Q. Cheng, "Popularity prediction in microblogging network: a case study on sina weibo," in ACM International Conference on World Wide Web (WWW), 2013, pp. 177–178.

20. Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "Seismic: A self-exciting point process model for predicting tweet popularity," in ACM Knowledge Discovery and Data Mining (SIGKDD), 2015, pp. 1513–1522.

21. S. Mishra, M.-A. Rizoiu, and L. Xie, "Feature driven and point process approaches for popularity prediction," in ACMInternational on Conference on Information and Knowledge Management (CIKM), 2016, pp. 1069–1078.

22. T. Zaman, E. B. Fox, E. T. Bradlow et al., "A bayesian approach for predicting the popularity of tweets," in The Annals of Applied Statistics (AOAS), vol. 8, no. 3, pp. 1583–1611, 2014.

23. B. Peng, "Modeling and predicting popularity dynamics via an influence-based self-excited hawkes process," in ACM International on Conference on Information and Knowledge Management (CIKM), 2016, pp. 1897–1900.

24. Y. Li, Z. Feng, H. Wang, S. Kong, and L. Feng, "ReTweetp : Modeling and predicting tweets spread using an extended susceptibleinfected-susceptible epidemic model," in Springer International Conference on Database Systems for Advanced Applications (DASFAA), 2013, pp. 454–457.

25. L. Yu, P. Cui, F.Wang, C. Song, and S. Yang, "From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics," in IEEE International Conference on Data Mining (ICDM), 2015, pp. 559–568.

AUTHORS PROFILE

Mrs.P.Anusha, working as Associate Professor Of Computer Science And Engineering Department in Qis College Of Engineering and Technology(Autonomous), Ongole, Andhra Pradesh, India

K.Brahma Teja pursuing B.Tech in the department of Computer Science &Engineering from Qis college of Engineering and Technology (Autonomous&NAAC'A'Grade),PonduruRoad,Vengamukkalapalem,Ongole,PrakasamDist.AffiliatedtoJawaharlalNehruTechnologicalUniversity,Kakinadain2018-2022respectively.

B.Eswara Raju pursuing B.Tech in the department of Computer Science &Engineering fromQiscollege of Engineering and Technology(Autonomous& NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, PrakasamDist. Affiliated to Jawaharlal Nehru Technological University, Kakinada in2018-22respectively.

D.Dheeraj Ashish Preetham pursuingB.Tech in the department of ComputerScience&EngineeringfromQiscollegeofEngineeringandTechnology(Autonomous&NAAC 'A'Grade),PonduruRoad,Vengamukkalapalem,Ongole,PrakasamDist.AffiliatedtoJawaharlalNehruTechnologicalUniversity, Kakinadain2018-22 respectively.

D.ChennaRayudu pursuing B.Tech in the department of Computer Science & Engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, PrakasamDist. Affiliated to Jawaharlal Nehru Technological University, Kakinadain 2018-22 respectively.