ISSN:0975-9646

Bharat Naiknaware et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (6) , 2015, 5265-5267

Prediction of Popular Content from Social Media Mining

Bharat Naiknaware[1] Seema Kawathekar[2] Sachin Deshmukh[3]

[1,2,3]*Dept. of CS & IT,*

Dr. B. A. M. University Aurangabad

Abstract-In recent trends social media websites, such as Facebook, Twitter, LinkedIn, YouTube and Google+ having certain content will attract much more visitors than others. Predicting the popularity of web content has become an active area of research. Predicting which content will become popular is of interest to website owners and market analysts. Popularity of content in social media is unequally distributed, with some items receiving a more attention from users. In Business analysis which newly submitted items will become popular is critically important for both companies that host social media sites and their users. Understanding what makes one item more popular than another, observing its popularity dynamics and being able to predict its popularity has thus attracted a lot of interest in the past few years. Predicting the popularity of web content is useful in many areas such as network dimensioning, online marketing or real-world outcome prediction. In this review, review of the current findings on web content popularity prediction are made. Here different popularity prediction models, present the features that have shown good predictive capabilities and reveal factors known to influence web content popularity.

Keywords: Web content, social media, Popularity, Prediction, Latent Dirichlet Allocation, Predictive Choice Model

INTRODUCTION:

In the social media world the web content has become the main attraction. It gives useful information and entertainment to Internet users or a business opportunity for marketing companies and content providers. Web content is a valuable asset on the Internet. At the same time the growth in social media innovations the ease of content creation and low publishing costs has created a world saturated with information. For example, every minute users around the world send more than 300,000 tweets, share more than 680,000 pieces of content on Facebook and upload 100 hours of video on YouTube[1][2][3][4][5]. Yet the online ecosystem the attention is concentrated on only a few items. Social media identifies the web content that will become popular and becomes of utmost importance.

Online users, flooded by information, can reduce the clutter and focus their attention the most valuable resource in the online world on the most relevant information for them. The explosive growth of user generated messages, Twitter has become a social site where millions of users can exchange their opinion. Sentiment Analysis on Twitter data has provided an economical and effective way to expose public opinion timely which is critical for decision making in various domains. For instance a company can study the public sentiment in tweets to obtain users feedback towards its products while a politician can adjust his/her position with respect to the sentiment change of the public. News articles are a type of content that can be easily produced, have a small size, short lifespan, and low cost, properties that makes them interesting for fast information diffusion through social media platforms or social networks. As a consequence, a significant amount of research has been focused in understanding the interest around news, including general observation on how content is generated, describing the decay of interest over time, community detection, and prediction of popularity. It is the former one, however, that gained most of the research focus both because this problem is very challenging and for its immediate practical implications, where predicting the popularity of online content is valuable for different actors: news sites and news aggregators can better highlight their popular content, online advertisers could propose more profitable monetization strategies, and online readers can filter more easily huge amount of information.

RELATED WORK:

In social media to predict the popular content from various social websites like facebook, twitter, LinkedIn, YouTube and Google+ in this all websites the decision making is core research and it uses the related work of different social medium in this Interpreting the Public Sentiment Variations on Twitter [1], to analyze public sentiment variations and mine possible reasons behind these variations, author propose two methodologies

1) Latent Dirichlet Allocation (LDA) based on models like Foreground and Background LDA (FB-LDA) and
2) Reason Candidate and Background LDA (RCB-LDA).

In this section author illustrate the intuitions and describe the details of the two proposed models**. Author investigated** the problem of analyzing public sentiment variations and finding the possible reasons two Latent Dirichlet Allocation (LDA) based models, Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA). The FB-LDA model can filter out background topics and then extract foreground topics to reveal possible reasons. To give a more intuitive representation, the RCB-LDA model can rank a set of reason candidates expressed in natural language to provide sentence-level reasons. These methodology uses real time Twitter data. Results shown that models can mine possible reasons behind sentiment variations. Moreover, the proposed models are general they can be used to discover special topics or aspects in one text collection in comparison with another background text collection.

Basic Statistics for the 50 Sentiment Variations shown in following table. It uses automatically collected tweets with labels to train the classifier. Then based on the algorithms outputs it will assign the sentiment label (positive, neutral or negative).

Response	Obama	Apple
# Negative Case	14	11
# Positive Case	12	13

We configure FB-LDA to output 20 foreground topics and 20 background topics and set LDA to produce 20 topics. The average word entropies for topics learned from FB-LDA and LDA are 3.775 and 4.389, respectively. It shows that the topics produced by FB-LDA exhibit lower word entropy than those learned by LDA, indicating that FB-LDA can generally obtain topics with less ambiguity and more interpretability.

REASON RANKING OF RCB-LDA

Our model is compared with two baselines: (1) TFIDF: In this method, each tweet or candidate is represented as a vector with each component weighted by term frequency/inverse document frequency (TF-IDF). The association is judged by the cosine similarity between the two vectors representing tweets and candidates. (2) LDA: For this method, we use the standard LDA model to learn topics from foreground tweets and candidates together (i.e., treating the foreground tweets and reason candidates as an entire text set).as shown in following table Ranking Results of Reason Candidates by RCB-LDA.

Cnt	Reason
275	Breaking Shooting at Arlington Apple Store!News Video via Mashable.WTF
191	Apple Patching Serious SMS Vulnerability on IPhone.Apple is working to fix an iPhone
179	Apple warns on iPhone 3GS overheating risk
101	Apple may drop NVIDIA chips in Macs following contract fight
87	Child Porn Is Apples Latest iPhone Headache
84	App store rejections: Apple rejects iKaraoke app then files patent for Karaoke player

RCB-LDA w/o BG	
Cnt	*Aspects*
214	Special Offer
193	Weight
192	Size
166	Button
159	Price
153	Navigation
150	Screen
146	Sensibility
141	Storage

RCB-LDA	
Cnt	*Aspects*
200	Weight
190	Size
183	Button
133	Sensibility
129	Special Offer
122	Storage
93	Screen
90	Navigation
89	Price

The comparison of RCB-LDA results under two settings: (1) RCB-LDA with both foreground and background data (2) RCB-LDA using only foreground data. In both tables the first column shows the count of reviews assigned to each candidate by our model. Since reviews are usually much longer than tweets, each review can cover multiple aspects.

Experiment results on kindle reviews. The left table shows the result of running RCB-LDA using only Kindle 4 reviews (foreground data) and the right table shows the result of running RCB-LDA using both Kindle 4 reviews (foreground data) and Kindle 3 reviews (background data).

In Learning Predictive Choice Models for Decision Optimization [2] the predictive choice model (PCM) author performed two experiments one on synthetic data we performed an in-depth exploration of the properties of the proposed predictive optimization model on synthetic data. In this study, two datasets are generated, the first with three mixture components and the second with 18 mixture components. Our results on the three-component dataset led to some interesting hypotheses about the relative performance of PCM and standard methods, and one on real admissions data .The admissions data is used from this pool of applicants to train and test our predictive optimization model. First create a training set from the admissions data for 2005, 2006, and 2007. Then created a test set from the admissions data for 2008 and then select the four available attributes from the candidates and application records that are continuous or ordinal. Author got a result like new probabilistic predictive choice model for decision optimization problems where the goal is to either maximize revenue or minimize cost.

THREE COMPONENTS

We started our experiments by generating two-dimensional synthetic data from a mixture of three Gaussian distributions. We generated 500 points according to each Gaussian component; the 1500 samples are shown in following table Summary of Results on Synthetic Data in Experiment

Method	RMSE	
	3 comp	18 comp
PCM(Hard Assignment)	0.1020	0.0885
PCM(Soft Assignment)	0.0911	0.0805
LR	0.2264	0.1996
LMT	0.2143	0.1760

Ranking news articles based on popularity Prediction [3] News articles are a captivating type of online content that capture a significant amount of Internet users interest. They are particularly consumed by mobile users and extremely diffused through online social platforms. As a result, there is an increased interest in promptly identifying the articles that will receive a significant amount of user attention. This task falls under the broad scope of content popularity prediction and has direct implications in various contexts such as caching strategies or online advertisement policies. Then compare the ranking capabilities of three prediction models 1) A simple linear regression model that will be proposed in preliminary analysis of the predictive characteristics on 20 minutes articles 2) The linear

regression on a logarithmic scale model proposed by Szabo and Huberman previously evaluated on Digg news, YouTube videos, and news articles from seven Dutch online news websites. 3) The constant scaling model described by Szabo and Huberman evaluated on Digg news and YouTube videos. This particular prediction task the most appropriate prediction method, out of the three that we have analyzed and for this dataset is a simple linear regression. From a general point of view it is observed that prediction methods do have an impact on the ranking accuracy, but their performance is rather limited giving that a simple heuristic, that includes the creation time of articles and the partial number of comments, reveals an accuracy that is comparable to the accuracy obtained when using prediction methods.

Basic Statistics for the 50 Sentiment Variations shown in following table. It uses automatically collected tweets with labels to train the classifier. Then based on the algorithms outputs it will assign the sentiment label (positive, neutral or negative).

Response	Obama	Apple
# Negative Case	14	11
# Positive Case	12	13

We configure FB-LDA to output 20 foreground topics and 20 background topics and set LDA to produce 20 topics. The average word entropies for topics learned from FB-LDA and LDA are 3.775 and 4.389, respectively. It shows that the topics produced by FB-LDA exhibit lower word entropy than those learned by LDA, indicating that FB-LDA can generally obtain topics with less ambiguity and more interpretability.

CONCLUSIONS:

The problem is investigated for analyzing public sentiment variations and finding the possible reasons causing these variations. To solve the problem, here it proposed two Latent Dirichlet Allocation (LDA) based models, Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA). The FB-LDA model can filter out background topics and then extract foreground topics to reveal possible reasons. To give a more intuitive representation, the RCB-LDA model can rank a set of reason candidates expressed in natural language to provide sentence-level reasons. New probabilistic predictive choice model for decision optimization problems where the goal is to on synthetic data show that the model is quite accurate in identifying the true number of mixture components from which the data are generated, estimating the parameters of each component's choice model, and predicting each individual's response probabilities. The predictive choice model substantially outperforms LR and LMT models and on real admissions data the predictive choice model dramatically outperforms LR and LMT models in this evaluation.

REFERENCES:

1] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, Xiaofei He," Interpreting the Public Sentiment Variations on Twitter" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 5, May 2014.pp 1158- 1170

2) Waheed Noor, Matthew N. Dailey,"Learning Predictive Choice Models for Decision Optimization" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 8, August 2014. pp. 1932-1945

3) Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida," Ranking news articles based on popularity Prediction" 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

4] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," J. Amer. Soc. Inform. Sci. Technol., vol. 62, no. 2, pp. 406–418, 2011.

5] C. Hueglin and F. Vannotti, "Data mining techniques to improve forecast accuracy in airline business," in Proc. 7th ACM SIGKDD Int. Conf. KDD, New York, NY, USA, 2001, pp. 438–442.

6] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," Communications of the ACM, vol. 53, no. 8, p. 80, 2008.

7] M. Tsagkias, W. Weerkamp, and M. De Rijke, "News comments: Exploring, modeling, and online prediction," in Proceedings of the 32nd European conference on Advances in Information Retrieval, ser.ECIR2010. Springer, 2010.

8] J. Lee, S. Moon, and K. Salamatian, "An approach to model and predict the popularity of online contents with explanatory factors," in Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01. IEEE Computer Society, 2010.

9] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in Proceedings of the 19th international conference on World Wide Web, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 621–630.

10] A. Tatar, P. Antoniadis, A. Limbourg, M. D. de Amorim, J. Leguay, and S. Fdida, "Predicting the popularity of online articles based on user comments," in WIMS'11. ACM, 2011, pp. 67–75.

11] A. Kaltenbrunner, V. Gomez, and V. Lopez, "Description and prediction of slashdot activity," in Proceedings of the 2007 Latin American Web Conference. Washington, DC, USA: IEEE Computer Society, 2007, pp. 57–66.

12] R. Bandari, S. Asur, and B. Huberman, "The pulse of news in social media: Forecasting popularity," Arxiv preprint arXiv: 1202.0332, 2012.

13] C. Hsu, E. Khabiri, and J. Caverlee, "Ranking comments on the social web," in Computational Science and Engineering, 2009. CSE'09. International Conference on, vol. 4. IEEE, 2009, pp. 90–97.

14] P. Yin, P. Luo, M. Wang, and W.-C. Lee, "A straw shows which way the wind blows: ranking potentially popular items from early votes," in Proceedings of the fifth ACM international conference on Web search and data mining, ser. WSDM '12. ACM, 2012, pp. 623–632.