How Online Content is Received by Users in Social Media: A Case Study on Facebook.com Posts

Yu Cheng, Yusheng Xie, Kunpeng Zhang, Ankit Agrawal, Alok Choudhary
EECS Department, Northwestern University
2145 Sheridan Road
Evanston,IL,USA
{ych133,yxi389,kpz980,ankitag,choudhar}@eecs.northwestern.edu

ABSTRACT

In social media platform such as Facebook.com and Twitter, there are many settings in which users can publicly post content. A number of these sites offer mechanisms for other users to make responses to these content: a canonical example is from Facebook.com, where posts come with annotations like "1,492 people like this" or "view all 307 comments". Usually these user-generated content have the effectiveness of gaining influence for their publishers in social media by getting positive comments or "like". The influence gained from users varies widely across different posts, and reasoning about the effectiveness of a post for gaining influence is an important task in social media analysis. In this paper we develop a framework for analyzing and modeling how the online content get influence in the social media platform, using a large-scale collection of Facebook.com posts as the dataset. We find that the effectiveness of a post to gain influence depends not just on its content but also on when it is published and in a subtle way on how the post relates to their users' interests. As part of our approach, we also propose a simple and natural method to model and predict the effectiveness of a post in gaining social influence. The experimental results on the real-world dataset is consistent with our findings. Our work provides a new method to analyze the social media from data mining view, which contrasts with a number of theories from marketing and sociology.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Application—*data mining*

General Terms

Measurement, Theory, Algorithms

Keywords

Online content, Social media, Social influence, Text mining

SOMA'12, August 12, 2012 Beijing, China.

1. INTRODUCTION

Recent years have seen a large increase in the usage of content creation platforms in social media, such as Facebook, Twitter, online forums etc. - aimed at the general public. Especially a number of them provide functionality by which not only users can publish their contents(text, link, image or video), but other users can also respond to these contents. For example, in Twitter, users post messages as tweets and the other users can "reply" or "retweet" the tweets which they are interested in or like. A popular tweet may have thousands of replies, or are "retweeted" hundreds of thousands times. A user in Twitter may become more and more influential if his/her tweets get more responses from other users. In Facebook.com, a person can create and manage pages (also called "'walls") to represent their organizations or companies. Facebook provides some functions to allow the page managers to create content publicly in their pages to which their audiences can respond to. There are two main "responses" in Facebook.com which users can make: (1)comments: write opinions to the posts; (2)"like": click the button "like" if users think the post is good or are interested in it. Figure 1 shows an example of the responses users made to a post in $BestBuy$ wall. The post talked about an advertisement of a popular Xbox 360 game and got 198 comments (circled blue), and 177 "like" (circled red). Similar to the situation in Twitter, one can expand his/her influence by posting as much "influential" content as he/she can. This is extremely important for companies and business organizations in social media. With the rapid growth of WWW, social media becomes a very important channel for them to communicate with users, and regular posting is the simplest and most effective way for engaging and growing their audiences, in other words, expanding their influences in social media. Understanding how these online content are received and diffused in social media is a fundamental problem in e-commerce, social network analysis as well as in the marketing domain. This issue is also increasingly important in the user interaction dynamics of large participatory web sites.

In this paper we develop a framework for understanding and modeling how online posts is received by users in social media and focus on the case of Facebook.com. The problem is related to several data mining research domain such as text mining, behavioral targeting (BT)[23], sentiment analysis, and subjective content [19]. There are many works about evaluating the "quality" of online user-generated content, mainly for customer reviews [14, 24, 15, 13, 22] and the messages in the community question/answering (QA)

Figure 1: An illustration of users' responses to a particular post from *BestBuy* public wall, Facebook.com

platforms [2, 4]. Intuitively, a "high quality" content is not definitely a "influential" content. The problem of "evaluating the influence of Y" can be formulated as "what did person Z think of post Y ?". But the formulation of the problem of "evaluating the quality of Y" is very different from "what did person Z think of post Y ?". Rather than asking "what did person Z think of post Y", we are asking, "what did person Z think of post Y in wall X ?". There are now three entities in the process rather than two. For example, we want to know, "How people feel about the potato fries post in McDonald's wall", which is absolutely different from "How people feel about the potatoes fries post in Burger King wall?", or "How people feel about the potatoes fries post?". Another issue is the dynamics of the posts' content. Unlike most web content, such as customer reviews, blogs etc, page posts are streaming data and highly dynamic. Once new content is posted, some of the older ones will not show up in the audiences' updates and would get little attention(model details in Section 2). Although there is quite a bit of works for assessing the quality of online text, there has been little investigation of how the three entitie above affects the dynamics of online interaction. This is the main topic we consider here.

We first make the hypothesis based on the observations from the a large-scale collection of Facebook.com posts as the dataset. Then we provide a detailed analysis of those factors that may affect the effectiveness of the posts, including the time when it is posted, the closeness to the users' interest, the structure of the posts, the writing styles, etc. Finally we develop a simple model to predict the effectiveness of the posts for gaining influence. The model uses a regression-based method and is based on some major factors we analyzed above. Extensive experiments were conducted on the datasets which contain thousands of Facebook posts data, demonstrating the consistence of our proposed model. To summarize, we make the following contributions in this paper.

- We collect a large scale posts data from Facebook.com and make some hypothesis and conclusions based on the observations.

- We carefully analyze the possible factors that might affect the effectiveness of the posts to gain influences and identify four important ones: the time when it is posted, the closeness to the users' interest, the structure of the posts and the writing styles.

- We develop a regression-based model that is able to captures some of the important factors for the influence prediction.

The rest of the paper is organized as follows. In Section 2, we introduce the data and make some hypotheses as well as some conclusions from the observations. In Section 3, we proposed a model using regression to predict the effective based on the factors analyzed in Section 2. Experimental results and are presented in Section 4. Section 5 presents related work, and Section 6 concludes this paper and discusses directions for future work.

2. DATA AND OBSERVATIONS

2.1 Data From Facebook.com

In order to make effective analysis for the posts, we collected a dataset of over 1 million posts (corresponding to pages/walls of 1258 companies or organizations in Facebook.com). We made extensive use of the Facebook Graph API (http://developers.facebook.com/docs/reference/api/) to collect this data, and will discuss more details about the process in this section, with particular attention to avoiding sample bias.

Each company or organization in Facebook.com is grouped into a specific category. For example, "Burger King" and "McDonald's" are grouped into "food/beverages". "Amazon Kindle", "HTC Magic" and "Nikon" are categoried as "electronics". There are 166 categories in total in Facebook.com. Usually there is a unique ID corresponding to a company or organization page. Using graph API, we can easily collect all the posts' content as well as their metadata within that company or organization page by sending the query with its ID. The metadata is mainly composed of the time when the post was published, the accumulated number of "like" and comments. Furthermore, we downloaded all the comments to these posts and metadata, which include the comments' content, the time when the comments were made, and the user who made the comment.

In total, the data we collect from Facebook.com contains 1,359,600 posts corresponding to 1258 walls and divided into 10 pre-defined categories based on their walls. And there are totally 78,938,695 users comments corresponding to the 1,359,600 posts. The size of our dataset compares favorably to that of collections used in other studies looking at social media data: Sun et al. [21] used 262,985 Facebook Pages to investigate the diffusion through a large social media network; Bernard J. Jansen and Mimi Zhang [10] used 149,472 micro-blog postings to discover the overall trends of topic in micro blogging; Mor Naaman et al. [17] used 3379 messages from 13 users to examine the characteristics of social activity and patterns of communication on Twitter.

2.2 Hypothesis and Observations

The comments and "like" of posts: If we consider the question "what is the influence of the posts?" to be equivalent to the question "what is the quality of the posts", we can turn to the problem of determining the quality score for the posts using some textural features. However as dis-

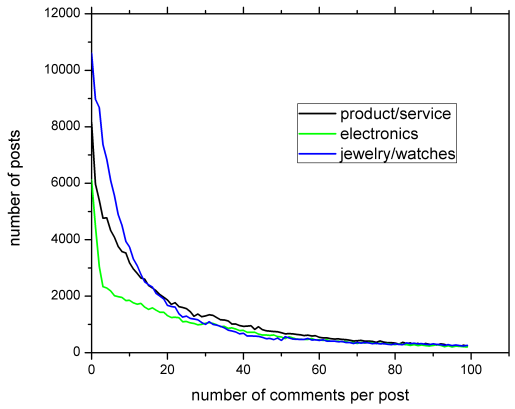

Figure 2: The distribution of received comments of posts on 3 categories

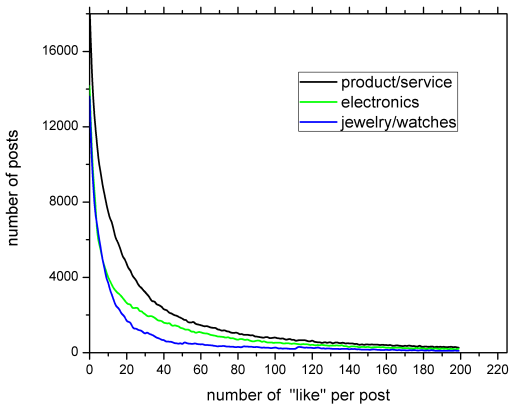

Figure 3: The distribution of received "like" of posts on 3 categories

cuss above, this problem of our problem is defined as "what did person Z think of post Y in wall X ?" rather than the definition "what did person Z think of post Y?". In other words, we should simultaneously calculate content features from "Y" as well as the external information from "X". The other problem is how to find the "influence" evaluation from users. We have seen that some social media sites like Facebook.com provide a way to assess the influence evaluations from the users based on the "like" and "comments" and it is a good way to measure the influence using the two. It is important to note that "like" is considered as positive evaluation but the opinion of comments can be negative, for which we should look into its content. The relation between "like" and "comments" is not well understood. However we can use the two to measure the influence of posts separately.

Distribution of "like" and comments: It is very interesting and important to know the distribution of "like" and "comments" of the posts. Some interesting questions here could be the following. Do all posts have the equal ability to gain influence? How many of them can not get attention from people? What are their difference among their influences? We analyze the distribution of comments and "like" respectively. Figures 2 and 3 show the number of comments and "like" separately versus the number of posts in three categories: product/service, electronics, and jewelry/watches. The trend of the distributions show that a large number of posts receive few or no comments/"like" and only a few post can gain very high influence. However, there is a "long tail" of the distribution and this matches the very famous "power law", which demonstrates that there are huge differences among the posts for gaining influences. The "power law" distribution is very interesting since it is highly consistent with some phenomenons in marketing and sociology [18, 11].

How the posts receive comments: A post is a very highly dynamic content on social media. A Facebook wall may publish hundreds of posts per day. Due to its high dynamics, it will show up only within a very short time in users' updates and the number of comments/"like" declines as time passes by. Figures 4 and 5 show the number of comments versus the time the posts are published (the number

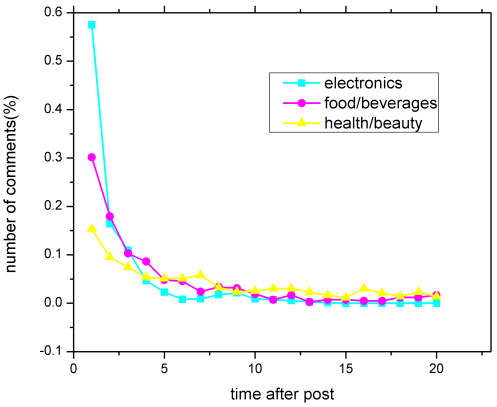

Figure 4: The number of comments vs. time after the post is published in category level

of posts is normalized between 0 $\sim$ 1) at the category and wall level respectively. We observed that most comments are received within a very short time since the posts are published, which confirms our hypothesis that users mainly respond to the posts which are just published. Having observed the trend of number of posts, we hypothesize that the number on comments/"like" received by a post is subject to exponential decay with respect to time. This finding is quite interesting.

2.3 What Affects the Influence?

The quality of content may affect the influence of posts. There has been a wide range of work on evaluating the "true" quality of text. A previous study [9] has shown that the linguistic style can be a very good indicator of the quality of text. Most of them evaluate the text based on their writing style [15], semantic and lexical features [24] and meta-data information [13]. In this subsection we analyze other factors from the users' side that may affect the influence, which will provide the basis for the proposal of the model in the next

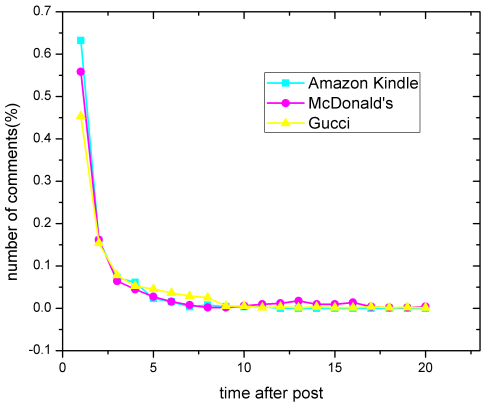

Figure 5: The number of comments vs. time after the post is published in wall level

Figure 6: Users' interest topics for category "retail"

section.

Users' Interest: Facebook walls often involve different fans with personal experience, interests and concerns. The influence of post may depend on the users' interests. We discovers the most talked about keywords and phrases from users' comments for any given social channel or category in any time period. Figure 6 and 7 visualize this and show the most interested topics of categories "retail" and "car". In addition, our analysis deals with streaming data and returns results in real time, to achieve which, we use methods from dynamic topic models [5] and item counting in streaming data [8]. For a category C, we create a words collection $T_C = \{w_1, w_2, ..., w_i, ..., w_t\}$ to express the dynamic interests of users in a period time.

Time stamp: In addition to users' interests, the influence gained by posts is also associated with the particular time stamp, which indicates when the post is published. For instance, research [3] shows that users' activities on social media are affected by time. As a concrete example, Figures 8 and 9 show the average number of comments received per post (normalized between $0 \sim 1$) versus the time for three categories within a day (from 0 to 23) and a week respectively. While off-peak hours from 2pm to 5pm, are the posts receive highest interaction rates and get more comments,

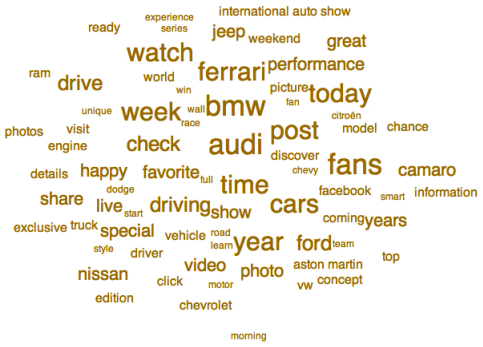

Figure 7: Users' interest topics for category "car"

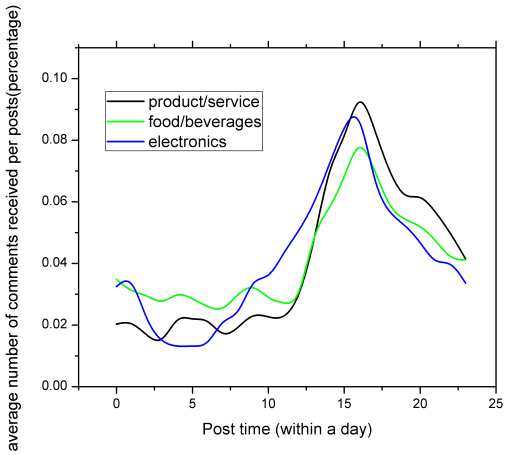

Figure 8: The average comments received in a day

Thursdays, on the other hand, shoulder the highest number of postings and interaction rate during the week. This is quite surprising as people usually think that weekend is the peak of receiving comments.

3. INFLUENCES PREDICTION

In Section 2, we analyzed many factors which may affect the influence of the posts. For a new post, we want to predict its potential influence and maximize the influence before it is posted. In this section, we propose a learning method to model these factors and predict the influence score. Since it is quite complicated to model the effectiveness of timeline, the proposed model currently only considers the factors from the post contents and its relation to the users. We first formally define the learning model and then select several features for predicting the effectiveness.

3.1 Problem Formulation

There can be many ways to measure the "influence" using "like" and comments, such as the number of "like", the number of positive comments, and comments score. The comments score function e for a post p is given as:

$$e(p) = \frac{S_+(p) - S_-(p)}{S_+(p) + S_o(p) + S_-(p)} \quad (1)$$

where $S_+(p)$ is the number of positive comments, $S_-(p)$ is

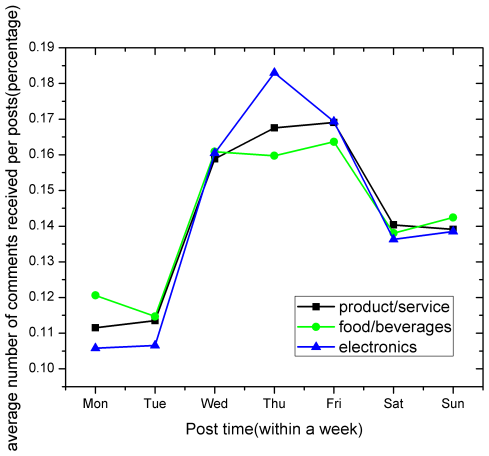

Figure 9: The average comments received in a week

the number of negative comments, and $S_o(p)$ is the number of objective comments. We only consider the influence measured by comments score as an example in this paper. In other words, the "influence" of a post is represented by the its comments score. The problem of estimating influence can be defined as follows: given a training set $P = \{p_i, e(p_i)\}$, $e(p_i) \in [0, 1]$, a number of features $f_1(p_i), ..., f_j(p_i)$, $...f_p(p_i), ...$can be composed. Our task is to construct a model F

$$e(p) = F(f_1, f_2, ..., f_j, ...f_p) \quad (2)$$

that can minimize the error in prediction of effectiveness $e(p)$. This can be viewed as a regression problem. When applied to a new instance p, the model F predicts the corresponding $e(p)$ and outputs the comments score of the prediction.

One solution for learning influence according to Eq. (2) is using SVM Regression [12, 20]. In this paper, we applied an SVM package on the features extracted from posts to learn the model function F.

3.2 Selected Features

One of our aims in this section is to investigate how these different features capture the effectiveness of a post. Generally, a effective post is a "reasonable" mixture of structural, subjective interesting, and readable content. The feature space selected in the learning framework ought to capture these points. Give a page post text T, we compute the following features and feed them into the rank-based algorithm.

Structural Features: Structural features are observations text structure and formatting. An effective post is supposed to be well formatted and informative. Properties such as the post length, whether it contains multimedia data (link,image,video) are hypothesized to relate to the structural of the content. We experimented with the following features:

1.The number of paragraphs in the post.

2.The average length of paragraphs in the post.

3.The number of links in the post.

Table 1: Statistics of Three Datasets for Experiments

Dataset	Mean	Std. Dev	# Posts	# Comments
FB-1	0.724	0.203	19847	$50 \sim 100$
FB-2	0.748	0.265	9566	$100 \sim 150$
FB-3	0.768	0.278	3721	$150 \sim 200$

4.The number of images in the post.

5.The number of video in the post.

User Interests Features: This is an interesting set of features, as we described in Section 2. We also use a list of related company names, brand names to generate the learning features. Specifically, we calculate counts of words in the following clue lists respectively:

1.The list of adjectives/nouns/verbs learnt in Section 2. More precisely, the list of words learned from a large corpus posts and comments with a specific topic.

2.The list of words learnt from a large corpus comments with a specific page.

3.The list of all product names and brand names within a specific page.

Readability Features: We make use of several features at the readability level. The readability is considered to be related to the quality of the text [14]. These features include:

1.The average length of sentences in the post.

2.The number of interjections and emoticons in the post.

3.The number of wh-words that signify either questions or other interesting linguistic constructs such as relative clauses.

4.The number of sentiment words in the post.

In total, we have 12 features for each post.

4. EVALUATION

We conducted extensive experiments on real-world datasets to evaluate the effectiveness of the proposed prediction model. First we describe the datasets in detail.

4.1 Experimental Setting

The experimental data is obtained by extracting three subsets from the data described in Section 2.1, with different number of comments: $50 \sim 100$, $100 \sim 150$ and $150 \sim 200$. Specifically, to maintain the robustness of the prediction model, we only consider posts with at least 50 comments. Table 1 summaries the distribution: mean and standard deviation (Std.Dev) of the influence score as well as number of posts etc. in the three datasets respectively.

In each dataset, we use 30% as training data and the remaining 70% as test data. Two standard metrics are used to evaluate the regression analysis:

Table 2: Correlation between influence score and post length

Dataset	r^2
FB-1	0.0135
FB-2	0.0307
FB-3	0.0452

Table 3: Regression performance on three datasets

Dataset	r^2	σ^2
FB-1	0.3127	0.0947
FB-2	0.3074	0.0907
FB-3	**0.3514**	**0.0582**

- Squared correlation coefficient

$$r^2 = \frac{((\sum_{i=1}^{n}(\mu_i - \bar{\mu})(\hat{\mu_i} - \bar{\hat{\mu}}))^2}{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2 \sum_{i=1}^{n}(\hat{\mu_i} - \bar{\hat{\mu}})^2} \quad (3)$$

- Mean squared error

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \hat{\mu_i})^2 \quad (4)$$

In the two equations above, μ_i and $\hat{\mu_i}$ are the real and predicted scores respectively; $\bar{\mu}$ and $\bar{\hat{\mu}}$ represent the mean of the corresponding sample respectively.

4.2 Experimental Results

Before the regression results are presented, we would like to see how the influence score of a post positively correlates with its length. As shown in Table 2, the correlation between the above two variables is weak. The regression performance on the three review collections (FB-1,FB-2,FB-3) are summarized in Table 3. All results presented in this section are based on 10-fold cross validation.

We have the following observations, based on the experimental results. (1)Across all three collections, the results are relatively similar qualitatively. The strongest model for each collection always achieves $r^2 > 0.30$ and $\sigma^2 < 0.10$, which are better than the result only using the length of post and are quite encouraging. (2)Generally speaking, The performance of regression increases with increasing the number of comments. This shows that the datasets with larger number of comments will be more suitable for our model. We believe that this is because that the standard deviation in FB-3 is smaller than FB-2 and FB-1, which has a more stable comments score.

5. RELATED WORK

In this section, we first review the problem of assessing the quality of user-generated content, and then review related works on user Behavioral Targeting (BT).

5.1 Problem of Assessing the Quality of Online Content

Recently the problem of assessing the quality of online content has attracted increasing attention. Most previous works [14, 24, 15, 13, 22] have typically focused on automatically determining the quality (helpfulness or utility) of customer reviews by using textual features. The problem of determining review quality is formulated as a classification or regression problem with users' votes serving as the ground-truth. Zhang et al. [24] found that syntactic features for the text are very useful. Kim and Pantel [13] proposed that the meta-data information including the review length and the number of sarts in product rating is very helpful. In [15], the authors incorporated reviewers' expertise and review timeliness in a non-linear regression model. Although user votes can be helpful as ground-truth data, some of this research has indicated that the helpfulness votes of reviews are not necessarily strongly correlated with certain measures of review quality. For example, Liu et al [14] identified a discrepancy between votes coming from Amazon.com and votes coming from an independent study. More specifically, they concluded that reviews accumulate votes depending on the number of votes they already have.

The problem of evaluating the quality of user-generated data is also critical in domains other than customer reviews. For example, the works [2, 4] focused on assessing the quality of postings in the community question/answering (QA) platforms. Agichtein et al. [2] combine textual features with user and meta-data features for assessing the quality of questions and answers. In [4], the authors propose a semi-supervised reinforcement framework that jointly models the quality of the author and the review. In [16], Lu et al. exploit contextual information about the authors' identities and social networks for improving review quality prediction. However, their work does not involve the social media, but rather uses the the external information from social media.

5.2 User Behavioral Targeting

The other works related to ours is the user Behavioral Targeting (BT). Behavioral targeting is yet another application of modern statistical machine learning methods to online advertising [23]. However, BT does not primarily rely on contextual information, but from past user behavior, especially the implicit feedback (i.e., click-through, page views) to match the best Ads to users [6]. Recently, there has been a large number of commercial systems proposed for targeted advertising. For instance, Yahoo! smart Ads [1] collects around 169M registered users for behavioral targeting, which also integrates the demographic and geographic targeting. The most crucial part of BT is to derives a relevence score for users' past activity within a category of interest. One of the most common measures is click-through rate(CTR). A well-grounded statistical model can predict click-through rate distribution of an ads from Ads view, page views etc. Canny et al. [7] described a linear Poisson regression model for behavioral count data and adopted the linear mean parameterization. In our paper, we propose to find the distribution of comments and "like" count rate versus time stamp and use exponential model.

6. CONCLUSION

The task of identifying high influential and quality user-generated content in social media sites is becoming increasingly important. We have seen that some social media sites like Facebook.com provide a way to assess the influence evaluations from the users based on the "like" and "comments". We analyze the data from Facebook.com and provide a regression based method to predict the influences of posts, which incorporats several main factors we analyzed. A post's

influence depends not just on its content, but also the time when it is posted and its relation to the users' interests. Extensive experiments on the real-world data set have confirmed the effectiveness of the proposed model.

There are a number of interesting directions for further research. First, we can find a model incorporating users' relationship and communities information. Moreover, it is important not only to model how users receive the content but also how users diffuse this information in social network. Finally, it would also be very interesting to consider social feedback mechanisms that might be capable of modifying the effects we observe here, and to consider the possible outcomes of such a design problem for systems enabling the expression and dissemination of opinions.

7. ACKNOWLEDGMENT

This work is supported in part by NSF award numbers CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CCF-1029166, and OCI-1144061, and in part by DOE grants DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DE-SC0005340, and DE-SC0007456. We also would like to thank Facebook.com, Voxsupinc.com for providing access to their data.

8. REFERENCES

[1] Yahoo! Smart Ads. `http://advertising.yahoo.com/marketing/smartads/`.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 183–194, New York, NY, USA, 2008. ACM.

[3] M. E. Belicove. Facebook Posting Techniques that Really Work. `http://www.entrepreneur.com/blog/220166`.

[4] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 51–60, New York, NY, USA, 2009. ACM.

[5] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[6] Y. Chen, D. Pavlov, and J. F. Canny. Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 209–218, New York, NY, USA, 2009. ACM.

[7] Y. Chen, D. Pavlov, and J. F. Canny. Behavioral targeting: The art of scaling up simple algorithms. *ACM Trans. Knowl. Discov. Data*, 4(4):17:1–17:31, Oct. 2010.

[8] G. Cormode and M. Hadjieleftheriou. Finding frequent items in data streams. *Proc. VLDB Endow.*, 1(2):1530–1541, Aug. 2008.

[9] A. Ghose and P. G. Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proceedings of the ninth international conference on Electronic commerce*, ICEC '07, pages 303–310, New York, NY, USA, 2007. ACM.

[10] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, Nov. 2009.

[11] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[12] T. Joachims. Advances in kernel methods. chapter Making large-scale support vector machine learning practical, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[13] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 423–430, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[14] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342, 2007. Poster paper.

[15] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 443–452, Washington, DC, USA, 2008. IEEE Computer Society.

[16] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 691–700, New York, NY, USA, 2010. ACM.

[17] M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 189–192, New York, NY, USA, 2010. ACM.

[18] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, December 2005.

[19] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, Jan. 2008.

[20] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, Aug. 2004.

[21] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! modeling contagion through facebook news feed. In *Proceedings of the Third International Conference on Weblogs and Social Media*, San Jose, CA, May 2009. AAAI Press, AAAI Press.

[22] O. Tsur and A. Rappoport. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In E. Adar, M. Hurst, T. Finin, N. S.

Glance, N. Nicolov, and B. L. Tseng, editors, *ICWSM*. The AAAI Press, 2009.

[23] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 261–270, New York, NY, USA, 2009. ACM.

[24] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 51–57, New York, NY, USA, 2006. ACM.