2010 IEEE International Conference on Data Mining

What Do People Want in Microblogs? Measuring Interestingness of Hashtags in *Twitter*

Jianshu Weng*, Ee-Peng Lim*, Qi He†, and Cane Wing-Ki Leung*
*School of Information Systems, Singapore Management University
Email: jianshu@acm.org, {eplim,caneleung}@smu.edu.sg
†College of Information Sciences and Technology, Pennsylvania State University
Email: qhe@ist.psu.edu

***Abstract*—When microblogging becomes a very popular social media, finding interesting posts from high volume stream of user posts is a challenging research problem. To organize large number of posts, users can assign tags to posts so that these posts can be navigated and searched by tag. In this paper, we focus on modeling the interestingness of hashtags in Twitter, the largest and most active microblogging site. We propose to first construct communities based on both follow links and tagged interactions. We then measure the dispersion and divergence of users and tweets using hashtags among the constructed communities. The interestingness of hashtags are then derived from these community-based dispersion and divergence features. We further introduce a supervised approach to rank hashtags by interestingness. Our experiments on a *Twitter* dataset show that the proposed approach achieves a fairly good performance.**

***Keywords*-Twitter, hashtag, interestingness, ranking**

I. Introduction

Microblogging as a form of social media has attracted huge number of users in recent years. As microblog posts are very short in length and mostly casual in content, the vast majority of them are not interesting to serious users. On the other hand, the realtime and lightweight nature of microblogging has also made it popular among users. In the case of *Twitter*, users actively post tweets to document their daily lives. While many tweets may never be read by other users, there are other tweets that report interesting events that happen both in real world and online world. Such activities are known to be citizen journalism. For example, the Iranian election protests in 2009 were largely reported by *Twitter* users.

Twitter allows users to assign hashtags to tweets so that they can be navigated and searched more easily. Hashtags (or simply tags without loss of generality) are free style keywords or key phrases that give concise semantics to the tweets. In the case of Iranian election protests, the tag "#iranelection" has been used. Unfortunately, tags also come in large number and many new tags get created each day. Not all are meaningful and interesting for serious consumption. Tags automatically generated by programs further add complexity to the problem. It is thus crucial to automatically determine the *interestingness* of tags to help searching and navigating *Twitter* contents.

The ability to distill a set of interesting tags can lead to several useful applications. Traditional news media companies can monitor tweets assigned with interesting tags to gain insights into events reported by *Twitter* users. Reporting such events enlarges the traditional media coverage and offers different perspectives to the same news items.

In this paper, we therefore aim to measure tag interestingness by considering the social aspect. We propose the following desired properties for tag to be considered interesting:

- *Targeted user attention*: The tag should gain attention of one or more user community. Assuming that each user community has some common interest, a tag that catches the attention of users' in some communities is thus more interesting than tag already used by the entire user population.
- *Unexpected user attention*: The existence of different types of users and user communities in *Twitter* gives rise to some expected distribution of attention levels. The heavy tag users tend to pay more attention to tagged tweets while tag-averse users do not show their attention much. Hence, a tag that commands unexpected user attention distribution should be interesting.

To compute tag interestingness, we propose the MEDIC framework (MEasuring Interestingness based on Discussion In Communities). Figure 1 gives a complete overview of MEDIC.

We evaluate our MEDIC approach on a Twitter dataset consisting of more than 15K users located in Singapore, their follow links and their tweets. We compare our proposed tag *interestingness* measure with the baseline measures considering tag popularity among users and tweets. We also assemble a small ground truth data from Google Trend for training the scoring function for our interestingness measure. Our results show that community based approach to measure tag interestingness is better than the non-community based one. Through empirical case study, we also confirm the usefulness of our proposed measure.

1550-4786/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDM.2010.34

```
input :
    1) a set $V$ of users,
    2) a set $\mathcal{R}$ of social connections among users in $V$,
    3) a set $\mathcal{T}$ of posts uttered by users in $V$,
    4) a set $S'$ of tag-pairs and the relative ordering about their
       interestingness.
output: A set $S$ of tags discussed in $\mathcal{T}$ and their interestingness.
1  identify the tags discussed in $\mathcal{T}$, denote the set of topics as $S$;
2  foreach tag $s \in S$ do
3  |   find all the uses using $s$;
4  |   establish the interactions among users in $V$;
5  build the discussion graph $\mathcal{G}$, taking into account the social
   connections defined in $\mathcal{R}$ and interactions derived in line 2-4;
6  construct communities from $\mathcal{G}$;
7  foreach tag $s \in S$ do
8  |   based on the community structure, derive values of features
   |   $DPU(s)$, $DGel^{U}(s)$, $NDPM(s)$, and $DGel^{M}(s)$;
9  |   conduct tag-specific projection of the community structure;
10 |   derive value of features $DPD(s)$, $DGel^{D}(s)$, $DPC(s)$, and
   |   $DGel^{C}(s)$ based on the projection;
11 |   represent $s$ as a feature vector;
12 based on the ordering in $S'$ and the feature vectors derived in line
   7-11, learn the interestingness measure scoring function;
13 apply the learned scoring function to measure the tags'
   interestingness;
14 (optional) order the tags in $S$ according to their interestingness;
```

Figure 1. Overview of MEDIC

The rest of this paper is organized as follows. Section II describes how the communities are constructed. Following that, Section III details the community-based features and the supervised approach to learn *interestingness* measure. The effectiveness of MEDIC is validated empirically in Section IV. A brief review of related work is presented in Section V. Finally, Section VI concludes with directions for further research.

II. Community Construction

A. Discussion Graph

The *discussion graph* is an aggregation of the social connections and the tag-specific interactions among users.

The social connections can be represented by a directed graph $G^{\mathcal{SC}} = (V, E^{\mathcal{SC}}, W^{\mathcal{SC}})$, where V represents a set of users; $E^{\mathcal{SC}} = \{(v_a, v_b) | v_a, v_b \in V \text{ and } v_a \text{ follows } v_b\}$; and $W^{\mathcal{SC}} = \{w_{ab}^{\mathcal{SC}}\}$ where

$$w_{ab}^{\mathcal{SC}} = \begin{cases} 1 & e_{ab}^{\mathcal{SC}} \in E^{\mathcal{SC}} \\ 0 & \text{otherwise} \end{cases}$$

Note that, it is possible to have $w_{ab}^{\mathcal{SC}} \neq w_{ba}^{\mathcal{SC}}$. $G^{\mathcal{SC}}$ is a non-tag-specific graph.

The interactions among users regarding a specific tag s is also represented as a directed graph $G^{\mathcal{TC}_s} = (V, E^{\mathcal{TC}_s}, W^{\mathcal{TC}_s})$. It shares the same node set V as $G^{\mathcal{SC}}$. An edge $e_{ab}^{\mathcal{TC}_s} = (v_a, v_b)$ exists in $E^{\mathcal{TC}_s}$ only if user v_b has certain influence over v_a in interaction about tag s, with the influence strength determined by $w_{ab}^{\mathcal{TC}_s}$.

We assume that a user's tweet post assigned with a tag is influenced by earlier posts carrying the same tag contributed by all the other users [1], [2]. Such an effect can be realized by a time-decay function. Let x_k to be the time point user v_b posting a tweet x_k using a tag, and t_l be the future time point v_a posting another tweet x_l using the same tag. $\Delta(x_k, x_l) = t_l - t_k$ is the amount of time between the two posts. The strength of the influence v_b has over v_a in this pair of tweet posts is defined as:

$$w_{ab}^{\mathcal{TC}_s}(x_k, x_l) = \begin{cases} \lambda^{\Delta}, & if\ 0 \leq \Delta < T \\ 0, & \text{otherwise} \end{cases}$$

Here $\lambda \in (0, 1]$ is a decay rate factor. T is a threshold set to ignore the influence that is too weak[1].

Note that user v_b may have multiple posts before v_a posts about s, and v_a may also post multiple tweets about s. To account for this, we derive an aggregated influence from v_b to v_a for tag s as: $w_{ab}^{\mathcal{TC}_s} = \sum_{x_k \in X_b(s), x_l \in X_a(s)} w_{ab}^{\mathcal{TC}_s}(x_k, x_l)$ where $X_a(s)$ and $X_b(s)$ represent the tweet posts from v_a and v_b respectively, using tag s.

The *discussion graph* $\mathcal{G}$ is then defined as an aggregation of $G^{\mathcal{SC}}$ and $G^{\mathcal{TC}_s}$. $\mathcal{G} = (V, E, W)$, which shares the same node set V as $G^{\mathcal{SC}}$. $E = \bigcup E^{\mathcal{TC}_s} \cup E^{\mathcal{SC}}$, and $W = W^{\mathcal{SC}} + \sum_s W^{\mathcal{TC}_s}$. The weight $w_{ab} \in W$ captures the strength of the influence that user v_a receives from v_b in the use of tags. The influence is strongest when v_a follows v_b, and v_a always discusses shortly after v_b discusses about any tag.

We can represent graph $\mathcal{G}$ in matrix form as follows:

$$\mathcal{G} = \begin{pmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & \ddots & \vdots \\ g_{n1} & \cdots & g_{nn} \end{pmatrix} \tag{1}$$

where $g_{ab} = \begin{cases} w_{ab}, & (v_a, v_b) \in E \\ 0, & \text{otherwise} \end{cases}$.

B. Modularity-based Community Construction

MEDIC applies modularity-based graph partitioning to extract communities from the *discussion graph*. Community here refers to a strongly connected group of nodes, with strong links within group members and sparse weak links across groups. *Modularity* is proposed by Newman et al. as a metric to measure the quality of graph partitioning [3], [4]. Based on the matrix representation of $\mathcal{G}$, we can define node v_a's in-degree and out-degree as $d_a^{in} = \sum_b g_{ba}$ and $d_a^{out} = \sum_b g_{ab}$ respectively. The sum of all edge weights in $\mathcal{G}$ is defined as $m = \sum_{ab} g_{ab}$. The *modularity* of the partitioning is then defined as [4]:

$$Q = \frac{1}{m} \sum_{ab} (g_{ab} - \frac{d_b^{in} \cdot d_a^{out}}{m}) \delta_{c_a, c_b} \tag{2}$$

where c_a and c_b denotes the communities that node v_a and v_b belongs to respectively, $\delta_{c_a, c_b} = 1$ if $c_a = c_b$, 0

[1] In this paper, λ and T are fixed for all users and all topics. The impact of these two parameters will be studied in future work.

otherwise. A positive modularity indicates possible presence of community structure.

The goal here is to partition $\mathcal{G}$ into communities $\{c_i\}$ such that Q is maximized. Newman et al. propose an intuitive and efficient spectral graph theory-based approach to achieve this [4]. It first constructs a modularity matrix (B) of the graph $\mathcal{G}$, whose elements are defined as:

$$B_{ab} = g_{ab} - \frac{d_b^{in} \cdot d_a^{out}}{m} \tag{3}$$

Eigen-analysis is then conducted on the symmetric matrix $B + B^T$ to find its largest eigenvalue and corresponding eigenvector ($\overrightarrow{v}$). Finally, $\mathcal{G}$ is split into two communities based on the signs of the elements in $\overrightarrow{v}$. This treatment provides an initial set communities, but may still result in less-than-optimal community assignments. This can be remedied by a "fine-tuning" stage in which users are moved between communities in an effort to increase modularity, until no further improvement can be achieved [4], [3]. The spectral method is recursively applied to each of the two communities to further divide them into sub-communities. This process iterates until no more community can be constructed[2].

Small communities with less than three users are considered trivial and filtered. The total number of communities extracted after filtering is denoted by N_C.

III. *Interestingness* MEASURE

Recall that we would like tag *interestingness* to consider both targeted user attention and unexpected user attention properties, which can be modeled by dispersion and divergence respectively. In this section, we first define a set of tag features considering both dispersion and divergence to aggregate features about users, posts and tag propagations in the different communities. Next, we define the *interestingness* measure function using a learning to rank approach.

A. Tag Features from User Data

This set of tag features aggregates the numbers of users from different communities using a given tag.

Definition *1:* Let $n_i(s)$ denote the number of users using tag s in community c_i. The ***user dispersion*** of tag s, denoted as $DPU(s)$, is defined as:

$$DPU(s) = -\sum_{i=1}^{N_C} p_{c_i}(s) \cdot \log p_{c_i}(s) \tag{4}$$

where N_C denote the number of user communities, $p_{c_i}(s) = \frac{n_i(s)}{n(s)}$, $n(s) = \sum_i n_i(s)$. ∎

As the entropy of user counts using s across different communities, $DPU(s)$ is small when the users using s belong to few communities, and large when the users are scattered evenly in different communities.

The aggregated tag feature using *user divergence* is defined by comparing tag-using user distribution with the general user distribution. The larger the difference between the two user distributions, the more unexpected is tag s in drawing user attention.

Definition *2:* Let the total number of users be $n = |V|$, and the number of users in community c_i be n_i. The ***user divergence*** of tag s is defined as:

$$DGel^U(s) = \sqrt{2 * D_{JS}(\mathcal{D}^U, \mathcal{I}^U(s))}\text{[3]} \tag{5}$$

where general user distribution $\mathcal{D}^U = \{p_{g_i}^U\}$ with $p_{g_i}^U = \frac{n_i}{n}$, tag-using user distribution $\mathcal{I}^U(s) = \{p_{c_i}(s)\}$, and $D_{JS}(P, Q)$ is the Jensen-Shannon Divergence between probability distributions P and Q. ∎

B. Tag Features from Tweet Post Data

We next define a set of tag features derived by aggregating the tweet posts using a given tag s. Again, we apply the concepts of dispersion and divergence. Unlike user dispersion and divergence, we normalize the number of posts in each community by the number of users in the community to avoid the effect of community size.

Definition *3:* Let the number of posts using tag s in community c_i be $m_i(s)$. The normalized number of posts using s in community c_i is $m_i'(s) = \frac{m_i(s)}{n_i(s)}$. The normalized total number of posts using s is $m'(s) = \sum_{i=1}^{N_C} m_i'(s)$. ∎

Definition *4:* The aggregated tag feature using ***post dispersion*** $NDPM(s)$ is defined as:

$$NDPM(s) = -\sum_{i=1}^{N_C} p_{m_i'}(s) \cdot \log p_{m_i'}(s) \tag{6}$$

where $p_{m_i'}(s) = \frac{m_i'(s)}{m'(s)}$. ∎

Definition *5:* For each community c_i, denote the count of posts uttered by users in this community as m_i. $m_i \geq m_i(s)$. The normalized post count is $m_i' = \frac{m_i}{n_i}$, and $m' = \sum_i m_i'$. The normalized post distribution of the general discussion is denoted as $\mathcal{D}^M = \{p_{g_i}^M\}$, where $p_{g_i}^M = \frac{m_i'}{m'}$. Similarly, tag-using post distribution is denoted as $\mathcal{I}^M(s) = \{p_{m_i'}(s)\}$. The ***post divergence*** of tag s is defined as:

$$DGel^M(s) = \sqrt{2 * D_{JS}(\mathcal{D}^M, \mathcal{I}^M(s))} \tag{7}$$

∎

C. Tag Features from Tag Propagation

In this family of tag features, we care about whether a tag has prompted active tag propagation in the communities. We focus on tag propagations where a tag is used by a user and shortly after another user.

To obtain tag propagation instances, we first induced a tag-specific graph for each community. For a community

[2] Due to space constraint, readers are referred to [4], [3] for detail of the spectral method.

[3] It has been proved that $\sqrt{2 * D_{JS}}$ is a metric which fulfills the triangle inequality [5].

c_i, the tag-specific graph of tag s is a unweighted directed graph $\mathcal{G}_i(s) = (V_i, E_i(s))$ induced by V_i which represents the users in c_i ($|V_i| = n_i$). Let the edge weights in $G^{\mathcal{SC}}$ and $G^{\mathcal{TC}_s}$ defined over V_i be $W_i^{\mathcal{SC}}$ and $W_i^{\mathcal{TC}_s}$ respectively. An edge in $E_i(s)$ from user v_a to v_b exists if and only if $w_{i_{ab}}^{\mathcal{SC}} + w_{i_{ab}}^{\mathcal{TC}_s} \neq 0$.

Now, we define the *density* of tag s propagations within community c_i as $den_i(s) = \frac{|E_i(s)|}{n_i*(n_i-1)}$. We also define the *largest connected component size* of tag s within c_i to be the number of users in the largest *connected components* in $\mathcal{G}_i(s)$. We then define the *propagation density dispersion* of tag s as follows.

Definition** 6:* Let the aggregated propagation density across all communities as $Den(s) = \sum_i den_i(s)$. The ***propagation density dispersion of tag s is defined by:

$$DPD(s) = -\sum_{i=1}^{N_C} p_{d_i}(s) \cdot \log p_{d_i}(s) \tag{8}$$

where $p_{d_i} = \frac{den_i(s)}{Den(s)}$. ∎

We can also derive the density of the induced discussion graph of a community c_i in a similar manner. In this case, the density is defined on an induced subgraph $\mathcal{G}_i = (V_i, E_i)$. V_i consists of all users in community c_i, and E_i is a subset of E which only contains the edges connecting two nodes in V_i. Then the density of the induced graph for community c_i is $den_i = \frac{|E_i|}{n_i*(n_i-1)}$.

Definition** 7:* Let the aggregated propagation density across all communities be $Den = \sum_i den_i$, the propagation density distribution of the general discussion be $\mathcal{D}^D = \{p_{g_i}^D\}$ where $p_{g_i}^D = \frac{den_i}{Den}$, and tag-specific propagation density distribution be $\mathcal{I}^D(s) = \{p_{d_i}(s)\}$. The ***propagation density divergence is defined as:

$$DGel^D(s) = \sqrt{2 * D_{JS}(\mathcal{D}^D, \mathcal{I}^D(s))} \tag{9}$$

∎

In the similar way, we can define the ***largest connected component size dispersion*** and ***divergence*** tag features. Denote them as $DPC(s)$ and $DGel^C(s)$ respectively.

D. Interestingness Measure Learning

Users are generally more concerned about relative ordering of tags' *interestingness* rather than the absolute degree of *interestingness*. Therefore, MEDIC formulates the design of *interestingness* measure as a pair-wise learning to rank process, which learns a ranking function from training data with pair-wise relative ordering [6].

Represent each tag as a feature vector; and denote the set of all feature vectors as S. The goal here is basically to learn a scoring function $f : S \to \mathbb{R}$ such that $f(s_i) > f(s_j)$ for tags $s_i \succ s_j$. Here, $\succ$ captures the ordering between a pair of tags, i.e. $s_i \succ s_j$ means s_i is more interesting than s_j. Current design of MEDIC applies Ranking SVM [6] to derive the *interestingness* measure.

Ranking SVM transforms the problem into a pairwise classification problem, and solves the transformed problem using conventional classification SVM algorithms [6]. In the case of linear function f, f is of the form $f(s; \mathrm{w}) = \langle \mathrm{w}, s \rangle$. Here, $\langle \mathrm{w}, s \rangle$ denotes the dot product of w and s. The goal now is to learn the weight vector w, which is achieved by solving the following optimization problem:

$$\begin{array}{ll} minimize: & \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{l} \xi_i \\ subject\ to: & \langle \mathrm{w}, s_{i1} - s_{i2} \rangle \geq 1 - \xi_i,\ i = 1, \cdots, l \\ & \xi_i > 0 \end{array} \tag{10}$$

With application of kernel trick, f can easily to extended to non-linear case [6].

IV. Empirical Evaluation

A. Dataset

The dataset used in the study is collected using the following procedure:

1) Obtain the most followed 1000 Singapore-based *Twitter* users[4] from *twitterholic.com*. Denote this set as V^*.
2) For each user in V^*, add her Singapore-based followers and friends within 2 hops. Denote this aggregated user set as V.
3) For each user in V, obtain all her published tweets.

The majority of the tweets collected are in the period from Sep 2009 to Apr 2010. All the results reported in the rest of this paper are based on this dataset.

There is a total of 3,092,955 tweets collected, published by 19,256 unique users. We exclude the isolated users who do not make any social connection with other users in our dataset and inactive users with less than 10 tweets. There are 15,630 non-isolated active users, who published 2,837,558 tweets.

B. Experimental Design

There is a total of $12,731$ hashtags in the dataset. This study focuses on the tags that appear in not less than 100 tweets as tags appearing in very few tweets are considered not interesting. There are 144 such tags.

We use Google Trend to establish the "ground truth" tag ordering as follows:

1) for each tag, obtain its definition from http://tagdef.com or http://www.whatthetrend.com;
2) craft a query for each tag based on the meaning obtained and submit to http://google.com/trends, with the region set as "Singapore" and the time frame as "Last 12 months" (which is enough to cover the focal time range in this study, i.e. Sep 2009 to Apr 2010);

[4] A user is considered Singapore-based if she specifies "Singapore" as her location in the profile.

3) obtain the tag's total search traffic in the focal time range[5];
4) choose two tags to form a tag-pair, and mark the one with higher search traffic as more interesting;
5) repeat Step 4) so that every tag is compared with all other tags once.

Eventually, 40 tags with definitions and non-zero search traffic are included in the "ground truth".

The performance of MEDIC is compared against two baseline methods. The first baseline RM measures a tag's *interestingness* by how often it has been used. RM is believed to be used in deriving trending topics in *twitter* [7]. The second baseline RU measures by the number of unique users using the tag. The three methods are applied to generate three ranked lists of tags. The correlation between the three ranked lists and the one in the "ground truth" is then measured by Kendall's τ [8].

C. Experimental Results

1) ***Community Construction****:* The *discussion graph* is built with $\lambda = 0.5$ and time unit as hour. The threshold T is set as 12-hour. 655 communities are then constructed. Among them, there are 525 trivial communities (with less than three users), making $N_C = 130$. The largest non-trivial community contains 4626 users, while there are 31 communities with the smallest size of 3.

2) **Interestingness** ***Scoring Function Learning****:* All the feature values are linearly scaled to the range of $[0, 1]$ before learning starts. Currently, MEDIC has two factors to tune for a better learning performance. One is the parameter C in Eq (10), while the other is the kernel function used in Ranking SVM. Repeated random sub-sampling cross validation is applied to obtain a reasonable settings. Optimal setting is chosen according to the average performance across 10 rounds of cross validations. Eventually, $C = 1$ and RBF kernel with $\gamma = 2$ are chosen to learn the *interestingness* scoring function over the complete training data.

The learned scoring function is then applied to generate a ranked list of all the 40 tags identified earlier in the "ground truth". Its agreement with the ranked list in the "ground truth" is $\tau = 0.695$ with corresponding one-tailed $p = 1.139 * 10^{-10}$. The extremely small value of p indicates that the high agreement is statistically significant.

3) ***Comparison with Baselines****:* The learned *interestingness* scoring function's performance is further compared against RM's and RU's. The comparison is listed in Table I.It is observed that τ value achieved by RM and RU is close to zero with a fairly large p value ($p > 0.05$). Statistically speaking, there is no enough evidence to reject the hypothesis that τ value achieved by RM or RU is zero. $\tau = 0$ means that the ranked list generated by RM or RU are totally independent from the one in "ground truth".

Table I
COMPARISON WITH BASELINE METHODS

Methods Results	MEDIC	RM	RU
τ	0.695	0.060	0.035
p (one-tailed)	$1.139 * 10^{-10}$	0.295	0.381

D. The Benefit of Considering Community

To justify the benefit of considering community, we exam the correlation between the "ground truth" and the eight features used in MEDIC as well as four other features derived without community structure considered. The two features used in RU and RM are denoted by SM and SU respectively. The density of the tag-specific induced *discussion graph* is denoted by SD. Feature $SC = n^g(s)/n(s)$. $n^g(s)$ denotes the user count in the largest connected component in tag s' projection of the *discussion graph*, while $n(s)$ is the count of all the users discussing s. We use FNC to denote the set of four features, i.e. $\{SU, SM, SD, SC\}$. For each feature studied, a two-tailed rank correlation test is run individually, with Kendall's τ employed as the statistics.

Table II summarizes the comparison. There are two major observations. First of all, the eight features used in MEDIC show higher correlation (higher $|\tau|$ value) with the "ground truth" than the four features in FNC. This means that features considering the community structure are better indicators of tags' *interestingness* than those not doing so. This justifies the need of considering community structure. Second, although the features in MEDIC achieve higher $|\tau|$ values, the corresponding p values are higher than 0.05. It means that the correlation is not statistically significant, i.e. there is no enough evidence to reject the hypothesis that the (linear) correlation achieved by each individual feature is 0. In other words, each individual feature is not a statistically significant indicator of tags' *interestingness*. This justifies the need of devising a function to aggregate the eight features to make a better indictor.

V. RELATED WORK

The immediate related research area is event detection [9], [10], [11]. An underlying assumption of event detection is that bursts could be observed in certain keywords' appearance frequency when an event happens [9]. This frequency-based assumption is also generally believed to be the design principle of the "trending topics" provided by *Twitter* [7].

With the frequency-based assumption, event detection essentially equates *interestingness* with the bursty-ness in content. It suffices in the context with very few user interactions, such as the traditional paper-based news media. Nevertheless, it simply overlooks the social dimension in the online context which is dominated by user-centered interaction and user-generated contents.

There is an emerging research interest in exploring the social dimension in event detection [12], [13]. In [12],

[5] I see a number next to my search term at the top of the graph. What does this mean?: http://www.google.com/intl/en/trends/about.html#11.

Table II
CORRELATION OF DIFFERENT FEATURES

Features Results	SU	SM	SD	SC	DPU	$NDPM$	DPD	DPC	$DGel^U$	$DGel^M$	$DGel^D$	$DGel^C$
τ	0.035	0.060	-0.122	0.0439	0.264	0.234	-0.129	0.323	-0.141	-0.153	0.181	-0.287
p (two-tailed)	0.762	0.589	0.271	0.701	0.566	0.840	0.242	0.778	0.200	0.637	0.312	0.804

Chen et al. propose to detect event by taking into account the social content contributed by users, such as the tags used in *flickr* [12]. However, the social connections among users are still ignored in [12]. The social connections are exploited together with content in [13]. However, [13] only explores the social dimension from a global perspective, without considering the differences across communities. To the best of our knowledge, MEDIC is the first to explore the social dimension with the difference across communities considered, whose benefit has been shown empirically.

VI. CONCLUSIONS AND FUTURE WORK

This paper focuses on measuring the *interestingness* of hashtags in *Twitter*. This paper proposes MEDIC, which measures hashtags' *interestingness* by studying how they are discussed within and across communities. Experimental studies show that MEDIC achieves a fairly good performance. Nevertheless, as an early attempt to measure *interestingness* from the new perspective, MEDIC still has space for improvement.

First of all, the influence between users is currently modeled with a fixed time decay λ and threshold T, and the *discussion graph* is modeled as a linear aggregation of the social connections and tag-specific interactions. It remains unclear how much a user is influenced by early discussion on the same tag and her social connection as well. Studying such influence would help to construct the community structure more accurately. Second, MEDIC only considers the community-based features. We plan to study how to include content-based features to help improving the performance in future. Third, MEDIC may exploit the unlabeled data to improve the performance of learning to rank. Last but not least, a study about MEDIC's applicability in other forms of social media is also planned.

ACKNOWLEDGMENTS

This work is supported by Singapore's National Research Foundation's research grant, NRF2008IDM-IDM004-036.

REFERENCES

[1] M. Mathioudakis, N. Koudas, and P. Marbach, "Early online identification of attention gathering items in social media," in *WSDM '10*, pp. 301–310.

[2] G. Kossinets, J. Kleinberg, and D. Watts, "The structure of information pathways in a social communication network," in *KDD '08*, pp. 435–443.

[3] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[4] E. A. Leicht and M. E. J. Newman, "Community structure in directed networks," *Physical Review Letters*, vol. 100, p. 118703.

[5] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[6] T. Joachims, "Optimizing search engines using clickthrough data," in *KDD '02*, pp. 133–142.

[7] StackOverflow, "What is search.twitter.com's 'trending topics' algorithm?" http://stackoverflow.com/questions/143781/what-is-search-twitter-coms-trending-topics-algorithm, Sept 2008.

[8] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.

[9] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *SIGIR '98:*, pp. 28–36.

[10] J. Kleinberg, "Bursty and hierarchical structure in streams," in *KDD '02*, pp. 91–101.

[11] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *WWW '10*, pp. 851–860.

[12] L. Chen and A. Roy, "Event detection from flickr data through wavelet-based spatial analysis," in *CIKM '09*, pp. 523–532.

[13] Q. Zhao, P. Mitra, and B. Chen, "Temporal and information flow based event detection from social text streams," in *AAAI '07:*, pp. 1501–1506.