Brand Data Gathering From Live Social Media Streams

Yue Gao, Fanglin Wang, Huanbo Luan, Tat-Seng Chua
School of Computing, National University of Singapore, Singapore
{dcsgaoy, dcswangfl, dcsluanh, dcscts}@nus.edu.sg

ABSTRACT

Social media streams, such as Twitter, Facebook, and Sina Weibo, have become essential real-time information resources with a wide range of users and applications. The rapidly increasing amount of live information in social media streams has important societal and marketing values for large corporations and government organizations. There is a strong need for effective techniques for data gathering and content analysis. This problem is particularly challenging due to the short and conversational nature of posts, the huge data volume, and the increasing heterogeneous multimedia content in social media streams. Moreover, as the focus of "conversation" often shifts quickly in social media space, the traditional keywords based approach to gather data with respect to a target brand is grossly inadequate. To address these problems, we propose a multi-faceted brand tracking method that gathers relevant data based on not just evolving keywords, but also social factors (users, relations and locations) as well as visual contents as increasing number of social media posts are in multimedia form. For evaluation, we set up a large scale microblog dataset (Brand-Social-Net) on brand/product information, containing 3 million microblogs with over 1.2 million images for 100 famous brands. Experiments on this dataset have demonstrated that the proposed framework is able to gather a more complete set of relevant brand-related data from live social media streams. We have released this dataset to promote social media research.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

General Terms

Algorithms

ICMR'14, April 01-04, 2014, Glasgow, United Kingdom.

Keywords

Brand tracking, Social media, Extended data gathering, Social context, Visual content.

1. INTRODUCTION

Social media platforms [15, 17], such as Twitter[1], Facebook[2], and Sina Weibo[3], have become essential real-time information resources with a wide range of users and applications. Consumers normally provide positive or negative comments when they post brand related information in microblog platforms, and such comments may spread quickly and widely across the entire social network. Such knowledge and insights have important marketing values for enterprises [8, 12, 20] which need to know about brand exposure and acceptance by users. Even for individual users, such insights are extremely useful to help them make purchase decisions on brands and products. The rapidly increasing amount of live information in social media streams demands the development of effective brand tracking techniques [7] for data gathering and media content analysis.

Brand tracking from social media streams has begun to attract research attention in recent years [14, 21]. The objective of brand tracking is to gather brand-related data from live social media streams. This is not a traditional search task due to several properties of social media data streams. First, social media posts tend to be short and conversational in nature, and thus the contents and vocabularies used in the posts tend to change rapidly. Hence, the use of a fixed set of keywords cannot guarantee the gathering of a sufficiently representative set of social media data. There is a need to track relevant microblogs using evolving keywords, key users and known locations. Second, the amount of social media data for a popular entity can be huge. For instance, the Super Bowl blackout at 2013 generated 231,500 tweets per minute, and the game got up to 24 million tweets in total. The traditional keyword-based data crawling methods [2,4,13] are limited in the coverage of all relevant data, and a multi-faceted approach is needed to ensure wider coverage of data from live streams. Third, the content of microblogs has become increasingly heterogeneous and multimedia. Recent statistics show that about 30% of microblog posts now contain images, and most such images do not have relevant text annotation. Hence, there is a need to analyze multimedia content when processing such microblogs.

[1]https://twitter.com/
[2]https://www.facebook.com/
[3]http://www.weibo.com

Figure 1: The framework of our brand tracking method.

The set of microblog streams gathered in such a multi-faceted approach will contain a large representative set of relevant posts, but also a lot of irrelevant ones. The next task is to analyze the set to filter out those irrelevant. As the data contains multimodal data, including text, images, locations and user data, a multimodal hypergraph based approach is employed to perform the filtering.

Overall, the proposed method contains four steps, namely, data gathering by using text, seed gathering, extended data gathering, and noisy data filtering, as shown in Figure 1. We first use brand-related keywords to crawl a set of related posts. Second, from the text-based results we select a set of seed microblogs by using the visual logo information. More specifically, we analyze both text and visual content to find a seed set that has relevance in both text and visual angles simultaneously. The posts in seed set are therefore highly relevant to the target brand. We then use the seed set to mine the social context (active users and known locations) and visual context. These context are then used as the basis to perform extended data gathering. Given the final set of data, a learning-based data filtering step is conducted to filter out noise. We evaluated the proposed brand tracking method on a microblog dataset on brand/product information from Sina Weibo containing 3 million microblogs with 1.2 million images (the Brand-Social-Net dataset). The experimental results demonstrated the effectiveness of the proposed method in gathering a more complete set of relevant brand-related data. As part of this research, we released the dataset in order to promote research in multimedia multimodal social media analysis tasks.

2. RELATED WORK

Gathering data from social media streams has attracted some research efforts in recent years [14, 21]. Existing works mainly focus on query expansion technique. Chen et al. [2] introduced a tweets gathering method, in which the keywords, candidate topics and popular topics are jointly employed for data gathering. Massoudi et al. [13] introduced a topic expansion technique to gather relevant data, in which query expansion is performed to generate dynamic topics for the target. They also introduced quality indicators for microblog posts, i.e., reposts, followers, and recency. These indicators are combined to estimate the probability of a microblog post. Similarly, Weerkamp and de Rijke [23] proposed a credibility framework to gather microblog posts. Sakaki et al. [18] proposed a real-time event information gathering in Twitter, in which a large query set of the target event is employed for data crawling. In [16], an exploratory data gathering method, named TweetMotif, is proposed by using frequent keywords and subtopics. Zhou et al. [27] proposed to expand personalized queries for data gathering. Besides the target, the annotations and resources of the user are also taken into consideration for further data crawling. A tag-topic model is formulated in a latent graph to explore the text data from social media streams. Leung et al. [11] proposed to employ human judgment to generate semantic indexes.

It is noted that most of the existing methods rely on text-based technique. Given the conversational and multimodal nature of social media streams, such methods will be limited in terms of coverage of relevant data.

3. BRAND DATA GATHERING IN SOCIAL MEDIA STREAMS

In this section, we introduce our proposed brand data gathering approach from social media streams, which comprises four stages.

3.1 Data Gathering based on Text Feature

For brand tracking, the text-based method is first performed to generate the initial results for the target brand $\mathcal{B}$, denoted by $\mathcal{M}^t$. Here, for a given brand, the related keywords, such as the brand name and corresponding product names, are used to crawl brand-related from social media streams. For example, given a brand "Volks Wagen", besides the brand name itself, the related keywords include the product names, e.g., "Jetta" and "Magotan", and other extended keywords, such as "car" and "engine".

3.2 Seed Gathering and Analysis

The data gathering using brand-related keywords tend to contain a lot of noise, as the presence of brand names does not necessarily guarantee the relevance of posts. To address this problem, we need to examine other aspects of microblog posts. Fortunately, many microblog posts contain images. We can thus leverage on image content to find a subset of relevant microblogs (known as the seed set) that have high relevance in both text and visual contents.

Here we use the logo as the discriminative visual feature for brand. Given the text-based results $\mathcal{M}^t = \{\mathcal{M}^{t_w}, \mathcal{M}^{t_o}\}$, $\mathcal{M}^{t_w} = \{m_1^{t_w}, m_2^{t_w}, \ldots, m_{n_w}^{t_w}\}$ are the n_w microblogs with images, and the n_o microblogs without images are denoted by $\mathcal{M}^{t_o} = \{m_1^{t_o}, m_2^{t_o}, \ldots, m_{n_o}^{t_o}\}$. For $\mathcal{M}^{t_w}$, let $\mathcal{I}^t = \{I_1^t, I_2^t, \ldots, I_{n_w}^t\}$ denote the corresponding n_w images. The objective of logo detection is to detect the brand logo in each image $I_i^t \in \mathcal{I}^t$. In our work, we employ a cascaded classifier which is jointly trained using Adaboost and SVM [3].

Figure 2 illustrates the logo detection method employed in our work. For training, we manually labeled a set of positive sample images obtained from Google Image and Flickr. A set of negative images with no brand logo is also collected to generate the initial negative sample set and false positives. Here false positives refer to those negative samples that are falsely classified as positive. The training is a recursive process as in [22] by producing a cascaded classifier consisting of multiple node classifiers. At each training round, Adaboost is conducted to select a bunch of Harr features. Different from [22], the final node classifier is a linear SVM learnt by using these selected Harr features on current training samples. Each node classifier is concatenated sequentially to form the cascaded classifier, which further runs on the negative image set to search for the false positives. The training is terminated when the false positive rate is low enough. During detection, the image is divided into sub-windows at multiple scales and we use the sliding window method with 1

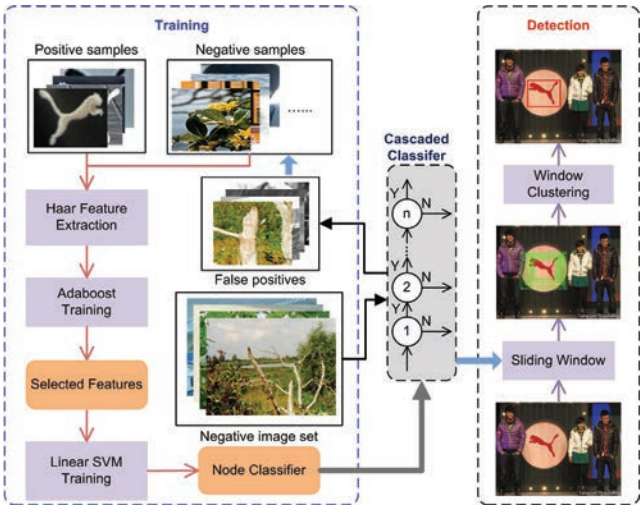


Figure 2: The logo detection pipeline in our work.

pixel stride on both the x and y directions for scanning. The sub-windows that are classified as positive are clustered to provide the final result representing the detected logos. During implementation, the training template is set to a very small size of, say, 24×18 pixels for the Puma logo. Thanks to the fact that each node of the cascaded classifier can eliminate a large amount of negative windows, the detection process is rather fast in practice.

All microblog images are then tagged as with or without logos with a property L. For the i-th image $I_i^t \in \mathcal{I}^t$, if it is detected as with the logo, then we set $L_i^t = 1$; otherwise, we set $L_i^t = 0$. As discussed before, these microblogs with relevant text and whose image has $L_i^t = 1$ are highly likely to be relevant to the brand and are used as the "Seed" set.

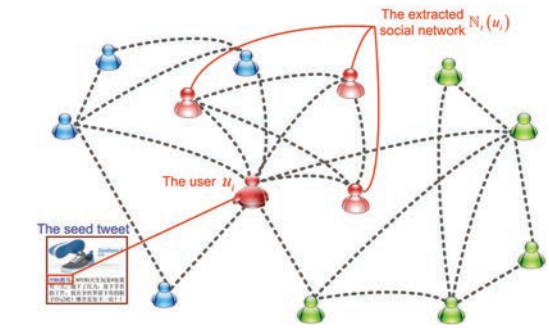
3.3 Extended Data Gathering

The result set $\mathcal{M}^t$ comes from text-based methods. To further explore the heterogeneous data in social media, we introduce the extended data gathering step to find more related microblogs beyond the text-based scope. In our extended data gathering procedure, both the social context and visual contents of the seed set are employed, which are introduced as below.

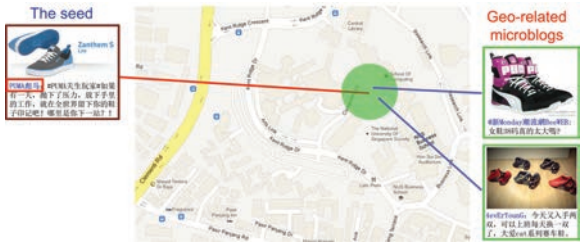
3.3.1 Social Context

In social media platforms, social context covers the social aspect of microblogs, such as the user name, post time, location, user comments, retweet activity, and relations between users, etc. Here we aim to mine accurate social context from the seed set for further relevant data gathering. In our work, we mainly focus on two types of social context, i.e., the key users and known locations extracted from the seed set. We briefly show these two types of extended information in Figure 3.

1. **The key users**.
   We define the key users as those who are active and influential with respect to the brand. Two types of key users are considered: (1) the authors of microblogs in the seed set; and (2) the users who have commented on the seed microblogs. These users are highly related to the seed microblogs and they have a high potential to post relevant microblogs again during a certain time period. For each author u_i of a seed, a time-constraint social network $\mathbb{N}_t(u_i)$ is extracted from their social connections $\mathbb{N}(u_i)$, and all the microblogs in $\mathbb{N}_t(u_i)$ are chosen as the candidates. For the users who have made comments, the microblogs from these users are also returned as the candidates.



(a) The user-based social context



(b) The known location-based social context

Figure 3: The illustration of extended data gathering by using social context.

2. **Known locations**.
   From the seed set, we want to identify possible geo locations with a high number of relevant seed microblogs. Such locations typically denote places with activities related to the brand, such as product launch, exhibition, etc. Therefore, other microblogs originating from these locations within a certain time period can potentially be relevant to the target brand too. Hence we gather all microblogs originated from nearby locations filtered by post time as a possible relevant set.

In our work, the time threshold for data selection is set as 1 day. By using the social context of the seed microblogs, we can obtain the social context-based microblog set, denoted by $\mathcal{M}^c = \{m_1^c, m_2^c, \ldots, m_{n_c}^c\}$.

3.3.2 Visual Content

Visual content is another essential information, which has increasing impact in social media streams. Similar visual content between two images can indicate close semantics in the corresponding microblogs. Here, we use the visual content of seeds as the basis for the second extended data gathering step to seek more potentially relevant microblogs. Figure 4 illustrates a visual content-based extended data gathering example.

As there are many duplicate images generated by re-post in social media platforms, we first perform seed image clustering to generate a group of unique images Λ for extended

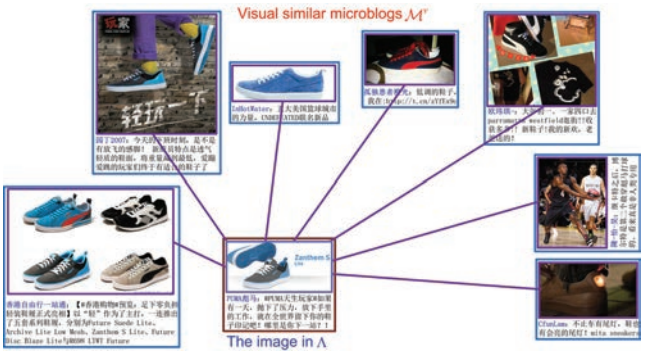


Figure 4: The illustration of extended data gathering by using visual content.

data gathering. Here we employ the hierarchical agglomerative clustering (HAC) method [19] for image clustering.

Next, we compare the contents of images in Λ with those posted within a fixed time period. For simplicity, we consider only a subset of images that are within the kNN distance of the images in Λ. Due to the high volume of data in social media streams, the set of images to be compared with these in set Λ is large, typically involve close to millions of images. For efficiency consideration, we build an efficient microblog image indexing system to achieve fast image matching. In the system, a spatial pyramid image feature [25] is extracted for each image, which is highly discriminative on spatial layout and local information. The dense sift feature is extracted for each image. A visual dictionary size of 1024 is learnt by sparse coding, and a spatial pyramid feature is generated by multi-scale max pooling. The spatial pyramid structure includes three levels and a 21504-D feature is generated for each image. We then generate a 32-bit Hash code for each image by using spectral hashing [24]. A 200-D feature is further extracted by using PCA for post-processing.

Now, given an image from Λ, the system first returns a set of results by using the Hash code. Next, the set is refined by using the PCA features. Finally, the results are ranked in terms of relevance to the image in Λ and the top n_i image are returned. The final set of extended visual search result obtained is denoted by $\mathcal{M}^v = \{m_1^v, m_2^v, \ldots, m_{n_v}^v\}$.

3.4 Noise Removal

In the previous stages, we have collected three different types of microblog candidates for a given brand $\mathcal{B}$, i.e., the text-based results $\mathcal{M}^t$, the social context-based extended data gathering results $\mathcal{M}^c$, and the visual content-based extended data gathering results $\mathcal{M}^v$. The use of extended data gathering also brings in a lot of noise. In this step, both the text information and visual content are investigated simultaneously to explore the relevance of these microblogs with respect to to the target brand $\mathcal{B}$, aiming to filter out noisy data.

To formulate the relationship among microblogs, a hypergraph structure is employed here. Hypergraph [26] has been employed in many data mining and information retrieval tasks [1, 5, 6, 9] due to its superiority in high-order relationship modeling. In the microblog hypergraph, a semi-supervised learning process is conducted for noisy data filtering. Figure 5 illustrates the noisy removal method.

Let $\mathcal{M} = \{\mathcal{M}^t, \mathcal{M}^c, \mathcal{M}^v\} = \{m_1, m_2, \ldots, m_n\}$ denote the

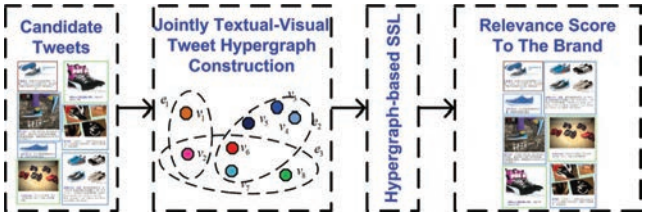


Figure 5: The joint text-visual microblog filtering procedure.

aggregated set of n candidate microblogs. A microblog hypergraph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{W}\}$ is constructed by using all microblogs in $\mathcal{M}$. In $\mathcal{G}$, each vertex $v \in \mathcal{V}$ denotes one microblog in $\mathcal{M}$. To investigate the correlation among microblogs, two types of hyperedges $\mathcal{E}$ are constructed; they are the text-based hyperedge $\mathcal{E}_{text}$ and the visual feature-based hyperedge $\mathcal{E}_{visual}$. For the text-based hyperedges, we first conduct text parsing for each microblog's text, and with a learnt codebook $\mathbf{D_{text}}$, each word is encoded into a code. Here we use only words with occurrence frequency of above a threshold σ ($\sigma = 10$ in this work) for hyperedge generation. Each microdblog m_i is represented by an $n_{c1} \times 1$ feature vector f_i^{text}, where $f_i^{text}(k, 1) = 1$ indicates that m_i contains the k-th word in the codebook. Each selected word generates a hyperedge, from which the microblogs in $\mathcal{M}$ that contain that word, i.e., $f_i^{text}(k, 1) = 1$, are connected. There are n_{c1} text-based hyperedges in total.

(a) Text-based hyperedges

(b) Visual-based hyperedges

Figure 6: Illustration for microblog hypergraph construction.

For visual content, the star-expansion method is employed to investigate the relevance among different microblog im-

ages. Each image is regarded as a center, and the top k nearest neighbor images are connected to this center, which generates one visual hyperedge. In our experiments, k is set as 5. There are n_{c2} visual feature-based hyperedges which is equal to the number of microblog images for processing. Figure 6 illustrates the hyperedge construction procedure. Altogether, there are $n_{c1}+n_{c2}$ hyperedges for $\mathcal{G}$. $\mathbf{W}$ is the diagonal matrix of the hyperedge weights. For each hyperedge $e_i \in \mathcal{E}$, the weight is set as $w(e_i) = \frac{1}{n_{c1}}$ and $w(e_i) = \frac{1}{n_{c2}}$ for the text-based and the visual-based hyperedges, respectively. The incidence matrix $\mathbf{H}$ of the microblog hypergraph $\mathcal{G}$ is generated by:

$$\mathbf{H}(v,e) = \begin{cases} 1 & \text{if } v \in e \\ 0 & \text{if } v \notin e \end{cases} \quad (1)$$

The vertex degree of a vertex $v \in \mathcal{V}$ is defined by:

$$d(v) = \sum_{e \in \mathcal{E}} w(e)\,\mathbf{H}(v,e), \quad (2)$$

and the edge degree of hyperedge $e \in \mathcal{E}$ is defined by:

$$\delta(e) = \sum_{v \in \mathcal{V}} \mathbf{H}(v,e). \quad (3)$$

We define two diagonal matrices $\mathbf{D_v}$ and $\mathbf{D_e}$ corresponding to $d(v)$ and $\delta(e)$ respectively as $\mathbf{D_v}(i,i) = d(v_i)$ and $\mathbf{D_v}(i,i) = \delta(e_i)$.

Our objective is to explore the correlation among all microblogs in the hypergraph structure. A semi-supervised learning procedure is conducted on the microblog hypergraph to minimize the empirical loss and the regularizer on the hypergraph structure simultaneously by:

$$\arg\min_{\mathbf{R}} \{\Psi + \lambda\Gamma\}, \quad (4)$$

where λ is a trade-off parameter; $\mathbf{R}$ is the to-be-estimated relevance vector of all microblogs to the brand; $\mathbf{Y}$ is the labeled vector by relevance estimation results in $\mathcal{M}^t$; Ψ is the regularizer on the hypergraph structure defined by:

$$\begin{aligned}\Psi &= \frac{1}{2}\sum_{e\in\mathcal{E}}\sum_{u,v\in\mathcal{V}} \frac{w(e)h(u,e)h(v,e)}{\delta(e)}\left(\frac{\mathbf{R}(u)}{\sqrt{\mathbf{D_v}(u,u)}} - \frac{\mathbf{R}(v)}{\sqrt{\mathbf{D_v}(v,v)}}\right)^2, \\ &= \mathbf{R}^T\left(\mathbf{I} - \mathbf{D_v}^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D_e}^{-1}\mathbf{H}^T\mathbf{D_v}^{-1/2}\right)\mathbf{R}\end{aligned} \quad (5)$$

and Γ is the empirical loss defined by:

$$\Gamma = \|\mathbf{R} - \mathbf{Y}\|^2. \quad (6)$$

Here we let $\Delta = \mathbf{I} - \mathbf{D_v}^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D_e}^{-1}\mathbf{H}^T\mathbf{D_v}^{-1/2}$, the solution for the above objective function can be achieved by:

$$\mathbf{R} = \left(\mathbf{I} + \frac{1}{\lambda}\Delta\right)^{-1}\mathbf{Y}. \quad (7)$$

By using the relevance score in $\mathbf{R}$, we can rank all microblogs in $\mathcal{M}$. The top results with high relevance scores are chosen as the relevant brand results.

4. THE BRAND-SOCIAL-NET DATASET

In this section, we introduce our microblog dataset (Brand-Social-Net[4]) on brand information.

[4]http://www.nextcenter.org/Brand-Social-Net/

4.1 Dataset

The microblog data was collected from Sina Weibo in June and July, 2012. This dataset consists of 3 million microblogs with 1.2 million images. Each microblog contains the text description, the image if available, the author information, posting time, geo location and user connections on Sina Weibo. This dataset includes 100 famous brands and 300 products, which are selected from automobile, sports, electronic products, and cosmetics domains. Figure 7 shows all 100 brand logos. In this dataset, there are about 1 million individual users.

Figure 7: The logos for 100 brands in the dataset.

For the 100 brands, the number of relevant microblogs ranges from 122 to 50,389, and the distributions of relevant microblogs for each brand are presented in Figures 8.

There are 20 brand/product-related events in this dataset. These events happened in June and July of 2012, and are listed in Table 1.

4.2 Reference Annotations

The dataset includes groundtruth on the relevance of each microblog to brands in terms of text/image, as well as the positions of objects/products/logos in each image. Each microblog is annotated by three volunteers, and the majority voting is employed to determine the final annotations.

- **Logo annotation**. For each image, a bounding box is used to identify the exact location of a logo, if presence.
- **Brand relevance annotation**. For each microblog, the relevance of the text content and the image content (if available) for each brand is annotated separately as 1 and 0.
  a) The text is annotated as $Br_t = 1$ if the text content is relevant to the brand; otherwise $Br_t = 0$.
  b) The image is annotated as $Br_i = 1$ if its content is relevant to the brand; otherwise $Br_i = 0$.

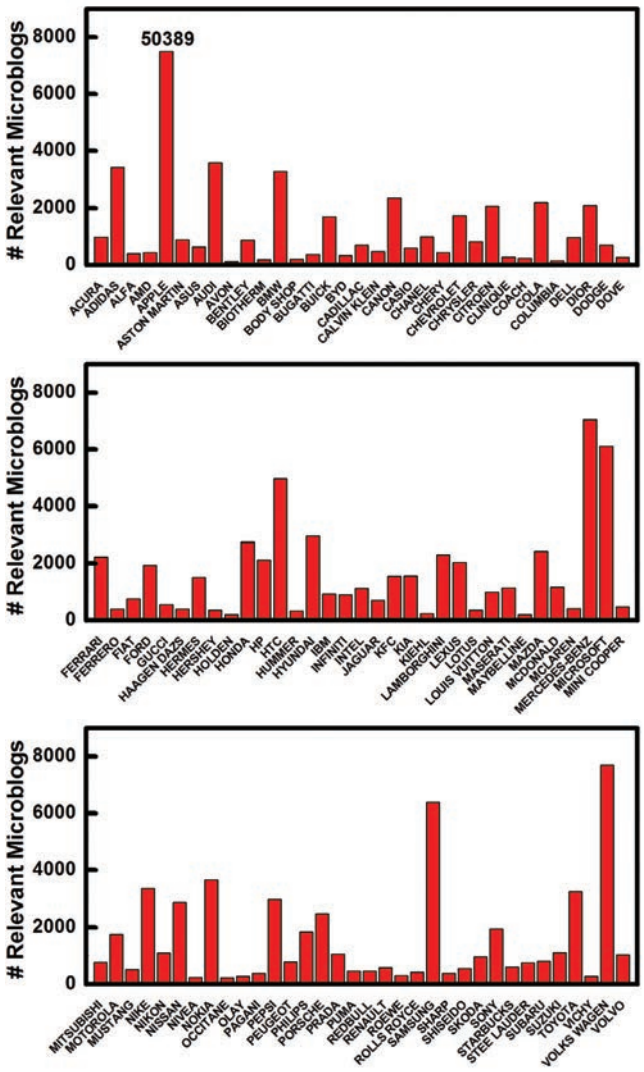


Figure 8: The number of microblogs for each brand.

Table 1: Events in the Brand-Social-Net dataset

The Apple Worldwide Developers Conference (WWDC)
Window 8 Release
Windows Office 2013 Release
Nokia Lumia Release
Pepsi Limited Edition Michael Jackson Cans
Sumsang Galaxy 3 I9300 Release
HTC ONE Release
Dior Addict Release
Hyundai MDAvante Release
CITROEN DS5 Release
Ferrari Berlinetta Release
Chrysler 300C Release
Honda CR-Z Release
Honda Elysion Release
Mazda CX-5 Release
Audi Q3 Release
Highlander 2012 Release
Shenzhen, Hongkong, Macao Auto Expo
Chongqing Auto Expo
Changchun Auto Expo

c) The microblog is annotated as $Br = 1$ if either the text content or the image content is relevant to the brand; otherwise $Br = 0$.

- **Product relevance annotation**. For each microblog, the relevance of the text content and the image content (if there is any) for each product is annotated separately as 1 and 0.

  a) The text is annotated as $Pr_t = 1$ if the text content is relevant to the product; otherwise $Pr_t = 0$.

  b) The image is annotated as $Pr_i = 1$ if its content is relevant to the product; otherwise $Pr_i = 0$.

  c) The microblog is again annotated as $Pr = 1$ if either the text content or the image content is relevant to the product; otherwise $Pr = 0$.

- **Object annotation**. If there are relevant objects for a given brand or product, the bounding boxes of these objects are labeled.

4.3 Challenging Tasks

The challenging tasks that can be performed on this dataset include, but not limited to, the following:

- Logo/Product/Brand detection and search task. This dataset includes 100 logos and 300 products, with the annotated groundtruth on the positions of logos/products and relevant objects. This task can be done using text, visual, social and combination of all features.
- Brand/Product data gathering task. One key challenge in social media platforms is how to gather representative set of data related to a brand or product.
- Social event analysis task. Over 20 brand-related events are defined for event detection and tracking research.
- Social media related research: This dataset contains social information to support research on sentiment analysis, social network analysis, key users and hot tweets/events analysis etc.

5. EXPERIMENTAL EVALUATION

To evaluate the performance of the proposed brand data gathering method from social media streams, we conduct experiments on the Brand-Social-Net dataset. The experimental settings and evaluations are provided in this section.

5.1 Experimental Settings

In our experiments, given a brand, the objective is to gather all relevant microblogs, i.e., $Br = 1$, to the given brand. The recall value is employed to evaluate the coverage of the gathered microblogs, and the Normalized Discounted Cumulative Gain (NDCG) [10] is used to measure the noisy data filtering performance. The parameter λ in Eq. (4) is set as 0.9. The number of selected images n_i is set as 100, and maximal number of returned images are 10,000 in our experiments. In the logo detection procedure, the average precision and recall are 0.743 and 0.383, respectively. As the logo detection results are regarded as the positive samples for microblog image relevance estimation, the precision is the most important criteria for further processing. A lower logo detection precision indicates more falsely detected

results leading to wrongly labeled samples for the following procedures. Thus high precision can guarantee that the selected images are highly related to the brand.

5.2 On the Coverage of Different Methods

Here we evaluate the coverage of different methods. For data gathering, the coverage is the most important performance regarding the brand. A higher coverage can lead to more useful content for further analysis. In our experiments, there are totally three data resources, i.e., the text-based results $\mathcal{M}^t$, the social context-based results $\mathcal{M}^c$, and the visual content-based results $\mathcal{M}^v$. The baseline method contains only the text-based results, and with extended gathering, $\mathcal{M}^c$ and $\mathcal{M}^v$ are included.

We first evaluate the overall coverage of different methods. The maximal coverage of keyword-based method can achieve a coverage of 60.12%, which is obtained by identifying whether there is any keyword in the text of the microblog. By utilizing extended data gathering based on social context, visual content and both, the data coverage is improved to 62.42%, 65.67% and 68.13%, respectively. Overall, the use of extended gathering method can lead to 13.32% improvement in data coverage as compared to just using the text based method as shown in Table 2.

Table 2: The comparison of data coverage with different variance of data gathering methods

Gathering Method	Coverage (%)	Improvement (%)
$\mathcal{M}^t$	60.12	-
$\mathcal{M}^t + \mathcal{M}^c$	62.42	3.83
$\mathcal{M}^t + \mathcal{M}^v$	65.67	9.23
$\mathcal{M}^t + \mathcal{M}^v + \mathcal{M}^c$	68.13	13.32

We also evaluate the coverage at top returned results of different methods. Here the coverage of top 100 to 1000 gathered results are compared in Figure 9. The proposed method could achieve a significant gain in the coverage of top returned results in comparison to the baseline method. With the social context-based results $\mathcal{M}^c$, $\mathcal{M}^t + \mathcal{M}^c$ obtains an improvement of 22.90%, 22.72%, 22.80%, 23.36%, 26.21%, and 20.60% for the recall depth of 100, 200, 300, 400, 500, and 1000 respectively as compared to baseline. For the visual content-based results $\mathcal{M}^v$, $\mathcal{M}^t + \mathcal{M}^v$, the improvements are 24.35%, 23.30%, 25.87%, 25.73%, 27.51%, and 21.96% respectively as compared to baseline. Specifically, the proposed method $\mathcal{M}^t + \mathcal{M}^c + \mathcal{M}^v$ obtains an improvement of 27.82%, 26.81%, 27.92%, 28.10%, 32.07%, and 26.90% for the recall depth of 100, 200, 300, 400, 500, and 1000 respectively as compared to baseline. This result demonstrates that the extended gathering method is effective on brand data gathering in social media streams.

5.3 On the Noisy Data Filtering

In this part, we evaluate the noisy data filtering performance. When multi-resources are employed by the extended gathering procedures, it brings in not only higher coverage of relevant data but also more noisy data. Therefore, noisy data filtering is essential for a better data gathering result. To evaluate the noisy data filtering performance, the NDCG values of top returned results are calculated to compare different methods. Figure 10 illustrates the comparison of all compared methods. As shown in Figure 10, the proposed

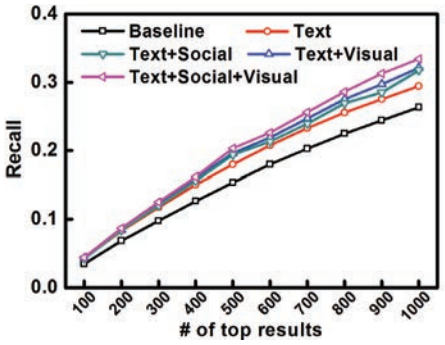


Figure 9: The recall performance of top returned results.

method with multi-faceted data resources can achieve better accuracy in the top results in comparison with the baseline method. We notice that the proposed method achieves an improvement of 16.18%, 15.24%, 13.81%, 13.15%, 12.21%, and 9.59% compared with the baseline method in terms of NDCG values at the depth of 100, 200, 300, 400, 500, and 1000, respectively.

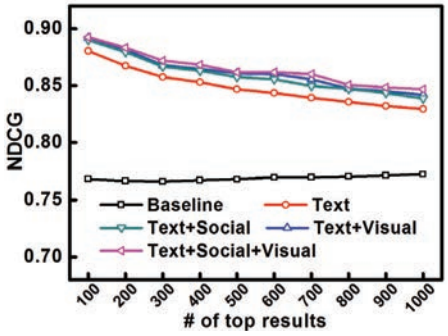


Figure 10: The performance of top returned results in terms of the NDCG measure.

6. CONCLUSION

The huge amount of live information in social media streams have led to high requirement for brand tracking technologies. In addressing this challenging task, we proposed a multi-faceted brand tracking method to gather representative data from large scale social media content. In our method, we took advantages of the heterogeneous data of social media content and proposed to mine a group of seeds first and then leveraged the social context and visual content of the seed set to gather more related posts from large scale noisy data. A noise filtering step is further conducted to filter out the noisy data in the returned results.

We have evaluated the proposed method on our microblog dataset (Brand-Social-Net), containing 3 million microblogs with 100 famous brands. Experiments on this dataset demonstrate that the proposed data gathering strategy can achieve better performance in comparison with the state-of-the-arts method.

To address the data harvesting task in social media platforms, there are still several future tasks. First, how to extract the visual context for the target objects is an important issue. The target objects may not explicitly appear in the visual content, while the visual context should implic-

itly help to uncover relevant visual content. Second, how to learn the social context from both the small seed set and the large data collection is important in both gathering more relevant data and in filtering noise. Third, the data filtering method comes with expensive computational cost. An effective and efficient data filtering algorithm is required when dealing with large scale live data.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

7. REFERENCES

[1] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He. Music recommendation by unified hypergraph: combining social media information and music content. In *Proceedings of MM*, 2010.

[2] C. Chen, F. Li, B. C. Ooi, and S. Wu. Ti: an efficient indexing mechanism for real-time search on tweets. In *Proceedings of the 2011 international conference on Management of data*, pages 649–660, 2011.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[4] M. Efron. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008, 2011.

[5] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai. 3D object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing*, 21(9):4290–4303, 2012.

[6] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, and X. Wu. Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing*, 22(1):363–376, 2013.

[7] S. Gaonkar, J. Li, R. R. Choudhury, L. Cox, and A. Schmidt. Micro-blog: sharing and querying content through mobile phones and social participation. In *Proceedings of the international conference on Mobile systems, applications, and services*, pages 174–186, 2008.

[8] C. Gu and S. Wang. Empirical study on social media marketing based on sina microblog. In *International Conference on Business Computing and Global Informatization*, pages 537–540, 2012.

[9] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas. Image retrieval via probabilistic hypergraph ranking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[10] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–466, 2002.

[11] C. H. Leung, A. W. Chan, A. Milani, J. Liu, and Y. Li. Intelligent social media indexing and sharing using an adaptive indexing search engine. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):47, 2012.

[12] G. Li, J. Cao, J. Jiang, Q. Li, and L. Yao. Brand tweets: How to popularize the enterprise micro-blogs. In *IEEE International Information Technology and Artificial Intelligence Conference*, volume 1, pages 136–139, 2011.

[13] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. *Advances in Information Retrieval*, pages 362–367, 2011.

[14] R. Nagmoti, A. Teredesai, M. De Cock, et al. Ranking approaches for microblog search. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010.

[15] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Searching microblogs: coping with sparsity and document quality. In *Proceedings of CIKM*, pages 183–188, 2011.

[16] B. O'Connor, M. Krieger, and D. Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[17] T. Rowlands, D. Hawking, and R. Sankaranarayana. New-web search with microblog annotations. In *Proceedings of WWW*, pages 1293–1296. ACM, 2010.

[18] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[19] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *Proceedings of KDD Workshop on Text Mining*, 2000.

[20] Y. Sui and X. Yang. The potential marketing power of microblog. In *International Conference on Communication Systems, Networks and Applications*, volume 1, pages 164–167, 2010.

[21] J. Teevan, D. Ramage, and M. R. Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44, 2011.

[22] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[23] W. Weerkamp and M. De Rijke. Credibility improves topical blog post retrieval. Association for Computational Linguistics (ACL), 2008.

[24] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. NIPS, 2008.

[25] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1794–1801, 2009.

[26] D. Zhou, J. Huang, , and B. Schokopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Proceedings of NIPS*, 2007.

[27] D. Zhou, S. Lawless, and V. Wade. Improving search via personalized query expansion using social media. *Information retrieval*, 15(3-4):218–242, 2012.