Co-Media: Creating Active Interactions For Localized Community With Social Media

Ji-Dong Wang* Guangshuo Chen*†, Jia-Liang Lu*, Min-You Wu*
*Shanghai Jiao Tong University, China
†INRIA, France
*{wangjd, jlu, mwu}@sjtu.edu.cn, † guangshuo.chen@inria.fr

***Abstract*—We present Co-Media, a system designed to enhance offline and online social connections among individuals. Co-Media discovers localized communities from the social media (Sina Weibo), collects the media content, and detects popular topics among community members. The system exploits the common interests, so that community members could develop further discussions and strengthen their interactions. Through a-year running of Co-Media, the system is validated and our method worked effectively. We also demonstrate that most topics have strong relevance to the community beyond its localization property.**

I. Introduction

There is a tremendous growth of smartphone usages in recent years. Smartphones account for 88% of new mobile devices and 69% of global mobile traffic in 2014 [1]. It is predicted that more than half of all devices connected to the mobile network will be ”smart” devices in 2019 [1].

Sensors provide rich approaches for smartphones to interact with the environment. For example, a variety of applications are designed for mobile sensing, such as information reposting [2] and meeting arrangement [3]. Encouraging participation and collaboration of mobile sensing is also studied in recent years [4], [5], [6]. Smartphones are utilized by social media services to enrich their content of social media. Mobile applications of Facebook, Twitter and Weibo, can make rich media content (*e.g.* video, audio and text) with additional information (*e.g.* GPS coordinates). Moreover, mobile sensing can help individuals better interact with each other, *e.g.* friend recommendation using history trajectory[7].

We are interested in the topic of creating active interactions for a localized community. Individuals are naturally formed into localized communities, *e.g.* students, residents or employees working in the same location. These localized communities are stable since individuals tend to spend most of their time in a few locations [8]. Members of a localized community probably have common interests. More importantly, they have willing to enlarge their social circles by knowing others from the same community [7].

It is common that people spend a lot of time in the same area (living or working) without knowing each other, and it is difficult to guide strangers in friendship in real social life. We see the opposite situation on social network. Interactions on social network are more frequent, free, and cost less.

We consider to using online interactions to active local communities. One problem is that localized communities are hidden on social media networks. On registration of service, personal information like address is not obliged. Fortunately, smartphones bring additional information to the content of social media, which helps to discover localized communities on social network.

In this paper, we propose community-base social media (Co-Media), a system that attempts to promote interactions among individuals on both online and offline sides by discovering localized communities in social networks and exploiting community-based common interests from the content of social media. Co-Media is designed to discover communities from social networks. Individuals in the community of Co-Media live in the same target area. Co-Media collects the content of social media from these individuals, and then explores their interest by event detection. After detection, events are presented to residents of the target area by using localized installations.

The rest of this paper is organized as following. Section.II introduces related works. Section.III presents Co-Media system and emphasizes the major parts. Section.IV evaluates the performance. Section.V concludes the paper.

II. Related Works

In this section, we introduce related works of Co-Media. Co-Media is designed to find residents by discovering their geographical information from social media services, and aggregate events from the content of social media, which relevant to *community detection* and *text event detection*.

Fortunato makes a survey of community detection in graphs [9], in which he defines of the problem, discusses and compares the major approach of discovering community structure or clustering in graphs. Papadopoulos at el. presents the concept of community and the problem of community detection in social media [10], with the classification of existing approaches. Co-Media discovers the community by geographical information.

Event detection is to discover popular events from the content of social media. Kumaran and Allan presents an effective approach by using the techniques of text classification and name entities [11]. Phelan at el. proposes a real-time news recommendation approach for Twitter [12].

Co-Media contains localized installations, in order to public localized events and interact with individuals nearby. FliterMeet [2] is an approach for public information reposting,

978-1-5090-0243-6/15/$31.00 ©2015 IEEE

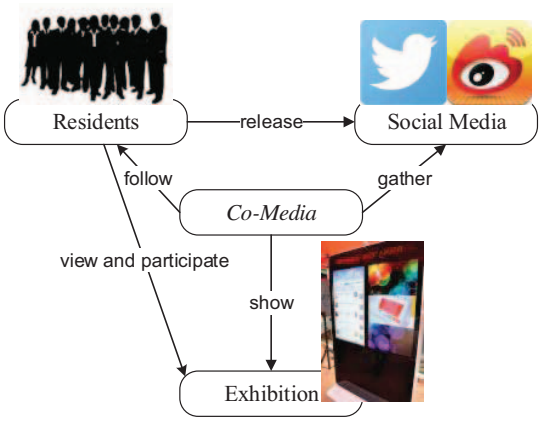

Fig. 1. The Co-Media Workflow.

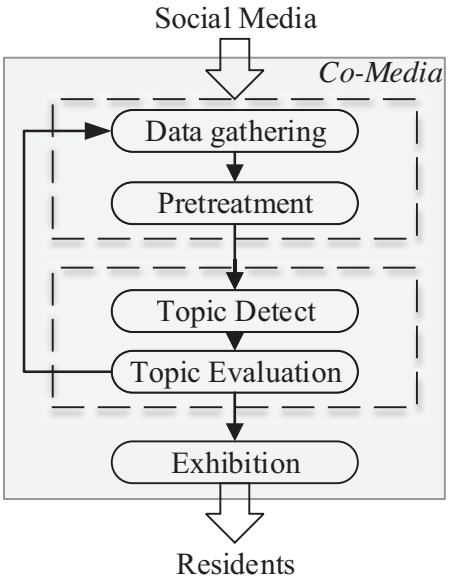

Fig. 2. The Co-Media Architecture.

which is designed to interact with varies of devices using different kinds of information. Co-Media is more targeted. The Co-Media workflow disturbs user's current experience of social media in a minimum level. The posted information in Co-Media comes from members of localized community and highly related to community.

III. Co-Media

In this section, we present the Co-Media system. Firstly, we give an overview of Co-Media by introducing the workflow and the architecture of the system. Then we illustrate three major processes in the Co-Media architecture, *i.e.* (1) social media data gathering, (2) topic detection and (3) topic evaluation.

A. System Overview

There are four roles in the Co-Media workflow: residents, social media, Co-Media system and exhibition as shown in Figure.1. *Residents* spend a large amount of time in the same area *e.g.* working or living nearby, and belong to a third-party social network such as Twitter or Weibo, where they generate content by using their devices like smartphones, *i.e. social media*. *Co-Media* is able to discover the community of residents, collects the content of social media and aggregates the popular topics in the community. These topics are presented to residents in the area reversely by deploying installations, *i.e. exhibition*. After residents find the interesting events through the exhibition, they will participate in the discussion and build new social network connection with others who are attracted by the same topic. Conversely, their interactions may contribute to new popular topic in the community.

B. Data Gathering

In our system, the data gathering is to collect microblogs from users who are physically located in the given area *i.e.* residents. We use an indirect way to find these users, since we have no access to their addresses. When using a smartphone, one posts microblogs containing GPS coordinates. With the GPS coordinate of the target area, we could obtain microblogs generated nearby. These microblogs come from both (1) residents and (2) randoms passing by. Hence we get our targets users after eliminating randoms.

In the data gathering process, we continuously follow users from microblogs at the target area. Alg.1 runs periodically to eliminate randoms from our following list. We design Alg.1 based on the assumption that a resident should post microblogs at the target area several times, but a random not. In our experiment, we update our following list and run the eliminating process at $5am$ every day.

With target users, we could get their microblogs simply from Weibo. For the accuracy of detection, we filter these microblogs by removing ones which (1) are less than 20 Chinese characters and (2) container over 70% of redundant characters (emojis, @username and links).

C. Topic Detection

After collecting microblogs from the target area, the next step is to extract events, which reveal the interests of people in the area. In Co-Media, we use clustering to detect events. These microblogs are categorized into groups by clustering. Each group links to an event, represented by the most typical microblog from the group. The event detection is repeated periodically, in order to find the live events in time.

In clustering, each microblog is divided into words by Chinese word segmentation [13]. Each word has a weight in terms of its meaning. There are some words which are common in the language and are usually filtered before or after natural language processing. Hereby words in the stop list [14] are filtered out before clustering.

Mathematically, a microblog C_i is modeled to a set of words as $C_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \cdots, (t_{in}, w_{in})\}$, where t_{ik} and w_{ik} represent the word and its weight respectively.

Algorithm 1 Eliminate randoms passing by from the following list.

Input:
U: the initial user set
μ: the upper bound
Output:
U_c: the user set after elimination
Main Procedure:

for u in U **do**
  $total \leftarrow 0$, $inArea \leftarrow 0$
  $W \leftarrow getMicroblogByUser(u)$
  for w in W **do**
    if w has GPS coordinate **then**
      $total \leftarrow total + 1$
      if w is in the target area **then**
        $inArea \leftarrow inArea + 1$
      end if
    end if
  end for
  if $total = 0$ **or** $inArea/total < \mu$ **then**
    Remove u from U
  end if
end for
return U

Hereby, Co-Media utilizes the Single-Pass clustering for the event detection [11] and obtains a serial of clusters, as Alg.2. Note that the similarity of two microblogs ($Similarity(C_i, C_j)$) in Alg.2, is calculated as following

1) get the union of two microblogs' words and their weight as $\{w_{i1}, w_{i2}, \cdots, w_{il}\}$ and $\{w_{j1}, w_{j2}, \cdots, w_{jl}\}$,
2) compute the similarity by

$$S(C_i, C_j) = \frac{\sum_{k=1}^{l} w_{ik} w_{jk}}{\sqrt{(\sum_{k=1}^{l} w_{ik}^2)(\sum_{k=1}^{l} w_{jk}^2)}}. \quad (1)$$

The word's weight makes impact on the accuracy of clustering. Co-Media utilizes the Term Frequency Inverse Document Frequency (TF-IDF) algorithm [11], in order to highlight meaningful words and downplay meaningless words in their weight. In the TF-IDF, a word is considered as expressing the core meaning of the microblog if it appears in a high frequency only in that microblog. The weight of word w_i is the product of the frequency of the word t_i in the microblog C_i (term frequency) as tf_i, and the inverse of the frequency of t_i in all microblogs (inverse document frequency) as idf_{ij}, *i.e.*

$$tf_i = \frac{n_i}{\sum_k n_k}, \quad (2)$$

$$idf_i = log \frac{|W|}{1 + |w \in W : t_i \in w|}, \quad (3)$$

$$w_{ij} = tf_{ij,i} * idf_{ij} \quad (4)$$

where n_i represents the number of t_i in the microblog C_i, $|W|$ represents the number of all microblogs, and $|w \in W : t_i \in w|$ represents the number of microblogs containing the word t_i.

Algorithm 2 Single-Pass clustering algorithm.

Input:
C: A serial of microblog text
Output:
E: A serial of clusters
Main Procedure:

for C_i in C **do**
  $S_{max} \leftarrow 0$
  $TempTopic \leftarrow null$
  if E is empty **then**
    New e
    $SetCenterText(e, C_i)$
    Put e into E
  end if
  for e in E **do**
    $C_j \leftarrow GetCenterText(e)$
    $S \leftarrow Similarity(C_i, C_j)$
    if $S > S_{max}$ **then**
      $S_{max} \leftarrow S$
      $TempTopic \leftarrow e$
    end if
  end for
  if $S_{max} > S_{th}$ **then**
    $Clustering(TempTopic, C_i)$
  else
    New e
    $SetCenterText(e, C_i)$
    Put e into E
  end if
end for
return E

D. Topic Evaluation

In order to reflect the residents' interests, it is necessary to rank the events after clustering. The ranking in Co-Media takes considerations of both the popularity and the community feature. Events are ranked by their community hot index CHI, which contains the hot index H and the community feature G. Mathematically, an event's CHI is formalized as $CHI_E = H_E * G_E$, in which

$$H_E = \sum_{w \in E} lg(follower(w) + 1) + comment(w) + 1 \quad (5)$$

and

$$G_E = \sum_{w \in E} g_w, \quad (6)$$

where $follower(\cdot)$ and $comment(\cdot)$ return the numbers of followers and comments, respectively. The g_w varies from 1 to 10, in terms of the correlation between the microblog and the target area. The strategy of setting g_w is customized in practice. For instance, the g_w is given a high value if the name of the area is directly mentioned in the microblog.

To each event, a microblog is selected in order to represent the event. The priority of the selection is that

1) from the official source,
2) from the original source,
3) having the least number of forwarding.

IV. Performance Evaluation

In this section, we present the results and evaluate the system.

A. Methodology

We deploy Co-Media in the Minhang campus of Shanghai Jiao Tong University, from April 2014 to April 2015. There are over $20,000$ students living in the campus of $2,822,903m^2$. The localized community hereby is considered as a group of students involved in Sina Weibo. We illustrate our results based on the data of April 2015.

From Sina Weibo, we collect information of users and media content as microblogs. Note that Sina Weibo allows us to obtain users around a given GPS coordinate. Hence we set the coordinate to the center of campus (30.032858, 121.447134) and discover users in the community. We then follow these users to obtain their microblogs. Only microblogs having a length $N >= 20$ are considered (about 85% as shown in Figure.3) as meaningful content.

The event detection process executes per day. μ in Alg.1 is set to 0.3. The g_w parameter in Eqn.6 should be related to the community. In the experiment, if a microblog mentions both the name of campus and university, its g_w is set to 10; if only the name of university, its g_w is set to 5; otherwise $g_w = 1$.

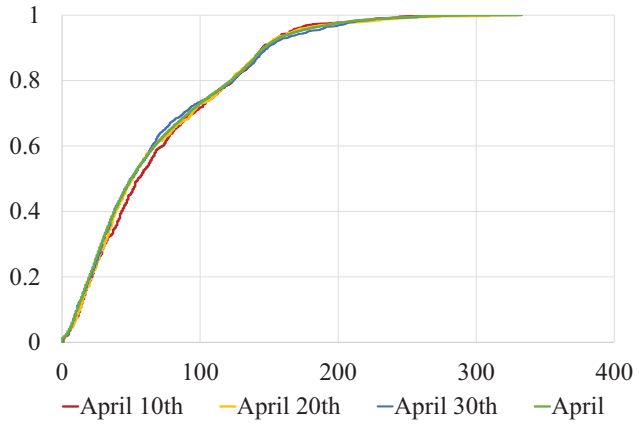

Fig. 3. CDF of lengths of microblog characters. In most cases, 85% of microblogs contain more than 20 characters. 60% contain over 50 Chinese characters.

B. Result and Analysis

In the experiment during the April 2015, Co-Media can detect about $4 \sim 10$ events per day. We classify these events into four categories: *community* (related to the community), *entertainment* (stars, TV shows or films), *advertisement*, and *other*.

We calculate the number of detected events by categories as shown in Figure.4. $30\% \sim 40\%$ of events are always related to the community. The community topics are about $30\% \sim 40\%$ in all topics every day. Figure.4 reveals that students are mainly interested on topics of their campus

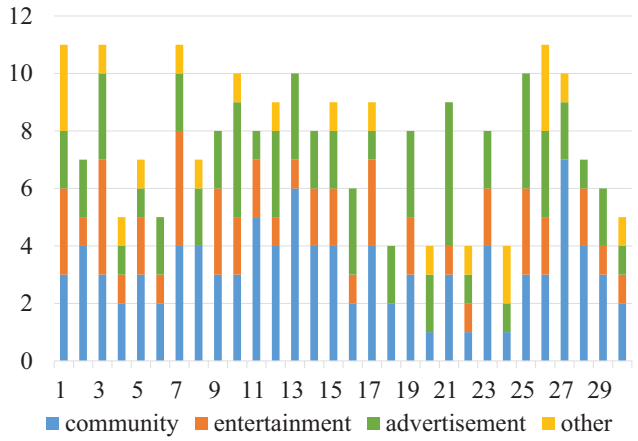

Fig. 4. Number of events grouped by categories(each day in April 2015).

(the community) and entertainment. Hence it is reasonable to promote interactions in the campus using these topics. Considerable *advertisement* events are detected by Co-Media, which are popular among members of the community. It means that social media advertising is widely observed and effective in the campus.

V. Conclusion

In this paper, we presented Co-Media, and introduced the design of Co-Media system and the implementation of data collection and event detection in the system. We validated the system through a-year running. The experiment showed that the idea of Co-Media was sound: our method worked effectively on detecting events about of community.

In the future, our interests would be (1) promoting better interactions between user devices and installations, and (2) discovering localized community more precisely.

Acknowledgement

This work was supported by the 863 project (No. 2015AA015802), Natural Science Foundation of China (NSFC) projects (No. U1401253), STCSM Project (No. 12dz1507400 and 13511507800).

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019 White Paper." http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html.

[2] B. Guo, X. Xie, H. Chen, S. Huangfu, Z. Yu, and Z. Wang, "FlierMeet: cross-space public information reposting with mobile crowd sensing.," *UbiComp Adjunct*, pp. 59–62, 2014.

[3] E. Miluzzo, N. D. Lane, K. Fodor, R. A. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell, "Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application.," *SenSys*, pp. 337–350, 2008.

[4] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing.," *MOBICOM*, pp. 173–184, 2012.

[5] Y. Zhang and M. van der Schaar, "Reputation-based incentive protocols in crowdsourcing applications.," *INFOCOM*, pp. 2140–2148, 2012.

[6] L. Duan, T. Kubo, K. Sugiyama, J. Huang, T. Hasegawa, and J. C. Walrand, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing.," *INFOCOM*, pp. 1701–1709, 2012.

[7] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, "Recommending friends and locations based on individual location history.," *TWEB*, vol. 5, no. 1, pp. 5–44, 2011.

[8] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Communications of the ACM*, vol. 56, pp. 74–82, Jan. 2013.

[9] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, pp. 75–174, Feb. 2010.

[10] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in Social Media," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 515–554, 2012.

[11] G. Kumaran and J. Allan, "Text classification and named entities for new event detection.," *SIGIR*, pp. 297–304, 2004.

[12] O. Phelan, K. McCarthy, and B. Smyth, "Using twitter to recommend real-time topical news.," *RecSys*, pp. 385–388, 2009.

[13] L. Zhou and D. Zhang, "NLPIR: a Theoretical Framework for Applying Natural Language Processing to Information Retrieval.," *JASIST*, vol. 54, no. 2, pp. 115–123, 2003.

[14] Baidu.com, "Stopwords." http://www.baiduguide.com/baidu-stopwords/.