Scalable Micro-blogging in Mobile Social Communities

Zunnun Narmawala
Institute of Technology
Nirma University
Ahmedabad, Gujarat, India
Email: zunnun80@gmail.com

Sanjay Srivastava
Dhirubhai Ambani Institute of
Information and Communication Technology
Gandhinagar, Gujarat, India
Email: sanjay_srivastava@daiict.ac.in

***Abstract*—Micro-blogging in Mobile Social Network can be very successful because of spatiotemporal properties of user interests and generated messages. Users can follow their interests and receive messages of their interests without getting overwhelmed instead of having to identify and follow users having similar interests. We propose a distributed and scalable micro-blogging protocol which exploits community structure and heterogeneous popularity of nodes. Control messages across communities are restricted by aggregating control information and by limiting storage of state information at hub/gateway nodes. The protocol dynamically estimates distance up to which interest in the message is high and restricts message forwarding accordingly. Simulation results show that our protocol performs far better than existing protocol. We also propose three novel models to generate user interest profiles synthetically.**

I. Introduction

Now-a-days all most everybody is carrying a smart phone equipped with Wifi and Bluetooth interfaces. These devices can form an opportunistic network without any fixed infrastructure. As these devices are carried by human beings, their mobility pattern and in-turn connection pattern is governed by human social behaviour. So, this network is also called as Mobile Social Network (MSN).

Micro-blogging, particularly Twitter, is very popular among Internet users. Users tag their tweets based on its content which is called as hashtag. Each user receives tweets from users he/she is following. One can follow a user but not a topic in the system. A user does not get all tweets having hashtags in which he/she is interested on his/her time line because huge number of tweets are generated for a hashtag across the Internet. User gets tweets having a hashtag only when he/she searches for it.

Micro-blogging can be a very promising application for MSN. Small messages posted by users can be opportunistically spread in the network without using any infrastructure. As co-located devices communicate with each other in this network, messages having local and temporal properties can be pushed to users based on their interests without overwhelming them. Users can specify their interests or they can be derived from their posts. Further, co-located devices generally have similar interests [1]. So, users can receive messages in which they are interested in with high probability.

Human beings belong to various social communities like friends, family, co-workers etc. [2]. Individuals belong to multiple communities [2]. Further, some people meet more people in the community (hub nodes) or visit other communities more often than others (gateway nodes). In this paper, we propose a novel micro-blogging protocol for MSN which exploits overlapping community structure and heterogeneous popularity of nodes. When two nodes come in contact, each node pushes to other node messages with tags in which the other node is interested in. To increase chances of delivery to interested nodes, all messages of a community are forwarded to hub and gateway nodes of the community. Hub nodes help in spreading messages in the community while gateway nodes help in spreading messages to other communities.

Nodes share their interest profile with community members and hub nodes of community. Hub and Gateway nodes share and accumulate aggregated interest profiles of communities. Based on aggregated interest profiles, threshold distance is estimated up to which interest is relatively high in a message. Message forwarding is restricted based on the threshold distance. As messages in which neighbour node is not interested are forwarded only to hub and gateway nodes, that too based on distance, forwarding overhead of the method is limited. Further, nodes do not share their interest profiles with nodes of other communities thereby limiting control overhead.

From the analysis of Twitter data, it is shown in [3] that large number of users are interested in a small number of tags. Further, users having similar interests are co-located [1] and interest in a topic decreases with distance. We propose three novel models to synthetically generate user interest profiles representing these properties.

Salient features of our protocol are following.

- We exploit social properties of human beings such as overlapping community structure and heterogeneous popularity of nodes for micro-blogging.
- We propose models for generating user interest profiles synthetically resembling real scenario.
- We limit forwarding of messages based on distance and keep the amount of state information independent of number of communities and network size to make our protocol efficient and scalable.

The rest of the paper is organized as follows: In section II, we discuss related work. Section III discusses user interest profile generation. Micro-blogging protocol is presented in section IV. In section V, we discuss simulation results. Conclusion is presented in section VI.

II. RELATED WORK

Content dissemination schemes based on publish/subscribe model are similar to our protocol. Socio-aware overlay [4] is based on non-overlapping community structure. It takes advantage of only hub nodes but not gateway nodes and actual routing between hub nodes of different communities is left to any standard unicast protocol. Further, modeling of user interest profile is not considered and there is no consideration for scalability. Socialcast [1] and ContentPlace [5] calculate a utility function and based on that decide where to place each message. We rely on hub and gateway nodes for all messages instead of detail information exchange and calculation for each message.

Allen et al. in [6] have proposed micro-blogging protocol for MSN. In their protocol, nodes maintain set of nodes having similar interest profile as friends and set of nodes with frequent contacts as familiar nodes. Messages are pushed to a friend node if it is interested or its friends are interested in the message. Further, messages are pushed to familiar node based on the aggregation of interest profiles of friend nodes and are pushed to stranger node randomly. Each node exchanges its interest profile with all the nodes it encounters. Further, friend nodes also exchange aggregated profile of their own friends. In our protocol, only community members exchange individual interest profiles and only aggregated profile of a community is shared with hub/gateway nodes of other communities. We compare the performance of our protocol with their work which is the only work available in the literature for micro-blogging in MSN.

III. USER INTEREST PROFILE GENERATION

A user is interested in a subset of all tags with different interest levels in each tag. List of tags along with interest levels in each tag is called as user interest profile. User generates messages as per his/her interest profile and is also interested in receiving messages as per this profile. We propose following three models for generating user interest profile synthetically.

A. Random Model

A given number of tags are randomly assigned to individual user irrespective of its community membership from the set of all tags such that number of users in the network per tag follow Zipf distribution. So, users interested in a tag are spread across different communities existing at different locations. The model represents interests which are not local. User interest levels in tags assigned to him/her are Zipf distributed with random exponent. In this model, we do not maintain any correlation between the level of interest of users in a tag with number of users interested in the tag.

B. Community Based Model

In this model, users interested in a tag are predominantly part of few communities only and only few users from remaining communities are interested in the tag. The model represents interests which are local. It is more realistic than the one proposed in [6] where set of tags of each community is independent of other sets. To generate user interest profiles based on this model, we shuffle set of all tags with different random seed for each community. Then, tags to individual users in a community are assigned using Zipf random variate from the shuffled sequence of tags for the community. As Zipf returns lower index values with high probability, tags at lower index values in the shuffled sequence are assigned to high number of users in a community. As different but overlapping sets of tags are at lower index values in shuffled sequences of different communities, this method generates desired profiles.

C. Distance Aware Community Based Model

It is an extension of the community based model. Let the density be the percentage of users in a community interested in a tag. In this model, for each tag in a community, density in other communities should be either proportional to the distance from the community or more than that. If it is less than required in a community then appropriate number of additional users are assigned the tag in that community. For brevity, we subsequently call the model as the distance based model.

In the the community based and the distance based model, user interest levels and number of users interested in the tag are correlated.

IV. MICRO-BLOGGING PROTOCOL

We describe various components of our protocol in following subsections.

A. Overlapping Community Structure

Overlapping community structure can be found in a distributed manner by using a simple extension of AD-SIMPLE algorithm [7]. AD-SIMPLE maintains only current community of a node. When a node travels to different community, members of the new community are added based on contact duration and previous community members are dropped through aging process. We propose to maintain these dropped community members as a separate community of the node. This simple extension can find different communities in which a node is a member instead of maintaining only current community.

B. Hub and Gateway Nodes

Locally popular or hub nodes come in contact with more number of nodes in a community than other nodes. Similarly, globally popular or gateway nodes visit multiple communities and come in contact with more number of nodes in the entire network than other nodes. We use the methods proposed in [8] to identify hub and gateway nodes from community structure itself without doing message flooding or forwarding of accumulated encounter information from other nodes. It orders nodes of a community based on their suitability as hub/gateway node. We use some percentage of most suitable nodes as hub/gateway nodes.

C. Control Messaging

When two nodes of a community encounter the first time, they exchange their user interest profiles. Upon subsequent encounters, they exchange change in interest profile since last encounter. Each node sends its interest profile to hub nodes of current community. Hub nodes compute average interest in each tag along with number of nodes of community interested in the tag. Hub and gateway nodes of the network exchange and accumulate this community wise aggregated information when they encounter each other. Substantial change in interest of community members in a tag is also propagated by community hub nodes to other hub and gateway nodes.

To keep amount of state information maintained at hub/gateway nodes independent of number of communities and network size, for each tag, aggregated information is stored only for given number of communities in which interest for the tag is relatively high. This also reduces the amount of control information sent between hub and gateway nodes of the network.

D. Threshold Distance Calculation

Threshold distance (D) for each tag from each community is the distance up to which messages originated from the community with the tag are forwarded to gateway nodes of the network with probability 1. Hub nodes of each community calculate threshold distance for each tag of community from accumulated aggregate interest profiles.

Each hub node calculates distances from its community to all other communities and sort communities based on their distances in increasing order. Let I_i^j be average interest in tag i in community j, N_i^j be number of nodes of community j interested in tag i, D_i^j be threshold distance for tag i from community j and M be number of communities. Let,

$$F_i^l = \frac{\sum_{k=1}^{k=l} I_i^k * N_i^k}{\sum_{k=1}^{k=M} I_i^k * N_i^k}, l = 1, 2, ..., M \quad (1)$$

Here $l = 1$ represents community of the hub node, $l = 2$ represents nearest community and so on. F_i^l represents ratio of aggregated interest in tag i till community l to total aggregated interest in tag i. Then threshold distance (D_i^j) is the distance of community l such that $F_i^l > Th$ where Th represents threshold interest which is a protocol parameter. So, the threshold distance (D) covers communities such that their normalized cumulative interest in a tag is greater than threshold interest (Th) which is between 0 and 1.

E. Probabilistic Forwarding

Actual messages are forwarded opportunistically when a node comes in contact with another node. If the neighbour node is neither hub nor gateway of current community then only messages of interest to the neighbour node are forwarded to it. Else, all messages of current community are forwarded. If the node is gateway node of any community and the neighbour is either hub or gateway node of current community then messages of other communities are forwarded as follows: If distance of the current node from community of message origin is less than threshold distance (D) then the message is forwarded. Else, it is forwarded probabilistically where probability decreases with distance.

We assume that a node will be able to send all messages needed to be sent as per protocol to a neighbour node upon each encounter based on the fact that in micro-blogging, message size is very small and average contact period is sufficient to send all messages over Wifi. We also do not consider limited buffer case because of small message size.

TABLE I. SIMULATION PARAMETERS

Number of nodes	200
Number of total tags	400
Number of tags of each node	30
Zipf exponent for tag popularity for the random model	0.5
Zipf exponent for tag popularity for other two models	2
Zipf exponent for user interest levels in tags	Randomly chosen between 0 and 2
Threshold interest Th for threshold distance calculation	1
Zipf exponent for density in the distance based model	1
Percentage nodes of each community as hubs	20 %
Percentage nodes of each community as gateways	20 %
Communication range	40 m
Message size	154 Bytes
Inter-packet arrival time	140 sec
TTL	400 min

TABLE II. COMPARISON WITH UTTERING

Model	Spread index		Efficiency	
	Our Protocol	Uttering	Our Protocol	Uttering
Random	0.71	0.29	0.30	0.13
Community Based	0.82	0.50	0.48	0.31
Distance Based	0.81	0.66	0.57	0.32

V. SIMULATION RESULTS

To evaluate the performance, we have implemented our micro-blogging protocol in ONE simulator. We use CAHM mobility model [8] which incorporates all properties of human mobility derived from real world mobility traces as well as from social network theory. For comparison, we have also implemented micro-blogging protocol for MSN proposed in [6] in ONE simulator. The protocol is called as 'Uttering'. We measure and compare 'efficiency' and 'spread index' of our protocol with Uttering. Efficiency is defined as ratio of total number of useful messages received by all nodes to the total number of messages received by all nodes. Spread index is defined as ratio of total number of users who have received useful messages to the total number of users who were interested in those messages. Simulation parameters are as per Table I unless otherwise mentioned. All readings are taken once network reached steady state. With a random seed, IHHW generated 14 overlapping communities for 200 nodes.

As shown in Table II, our protocol significantly outperforms Uttering for all types of user interest profiles. Spread index of our protocol is better because Uttering relies on other ordinary non-interested nodes to reach interested nodes while we rely on hub/gateway nodes for the same. Efficiency of our protocol is better because we forward messages to neighbour node even if it is not interested in the message only if it is hub or gateway and hub/gateway nodes are only small percentage of total nodes.

To see the effect of the threshold distance, we disable probabilistic forwarding. Variants with and without probabilistic forwarding are referred as 'Probabilistic' and 'Absolute' variants respectively. As shown in Fig. 1, as threshold interest increases in 'Absolute' variant, spread index increases and efficiency decreases as expected for all types of user interest profiles. Further, performance is far better for the community based and the distance based interest profile as compared to random interest profile. It is because, in these profiles, nodes interested in a tag are co-located while, in random interest profile, nodes interested in a tag are spread across communities.

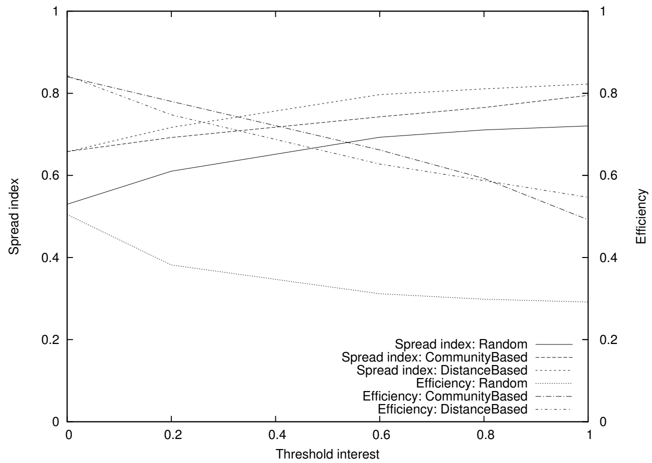

Fig. 1. Spread index and efficiency vs. Threshold interest (Th)

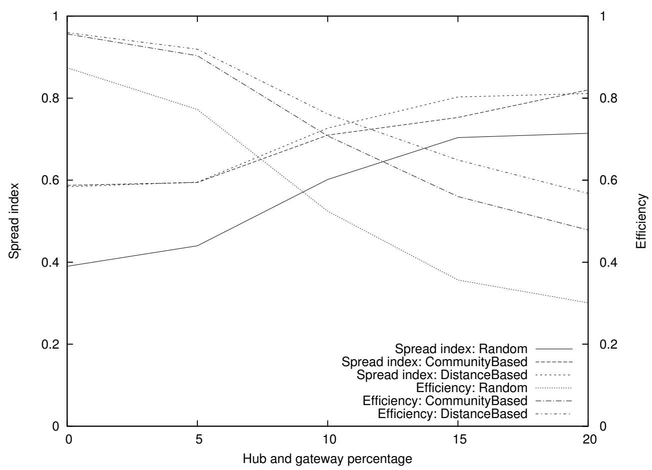

Fig. 2. Spread index and efficiency vs. Hub and gateway percentage

As shown in Fig. 2, as the percentage of nodes acting as hub and gateway nodes increases, spread index increases and efficiency decreases substantially. The figure is for 'Probabilistic' variant. Further, with an increase in hub and gateway nodes, the control message overhead is bound to increase. So, depending on the network traffic, number of hub and gateway nodes should be chosen. Further, with less number of hub and gateway nodes, load balancing can be done by rotating hub/gateway responsibility among eligible nodes.

Fig. 3 shows effect of limited number of entries of aggregate interest profile of a tag at hub/gateway nodes on the performance. As seen in the figure, for a tag, out of 14 possible entries for 14 communities, entries of only 2 communities which have high interest in the tag are sufficient. As for more than 2 entries, performance does not change; i.e. amount of information needed in our protocol is independent of number of communities and network size.

VI. Conclusion

In this paper, we propose micro-blogging protocol for MSN exploiting its overlapping community structure and heterogeneously popular nodes. The protocol restricts message forwarding based on distance and keeps a constant amount of state information to be efficient and scalable. We also propose three models to generate user interest profile synthetically out of which the distance aware community based model is most realistic. Simulation results show that spread index of our protocol is better than existing protocol (Uttering) by 18-59% and efficiency is better than Uttering by 35-56% for different user interest profile models. Simulation results also show that keeping only constant amount of state information does not degrade performance of our protocol.

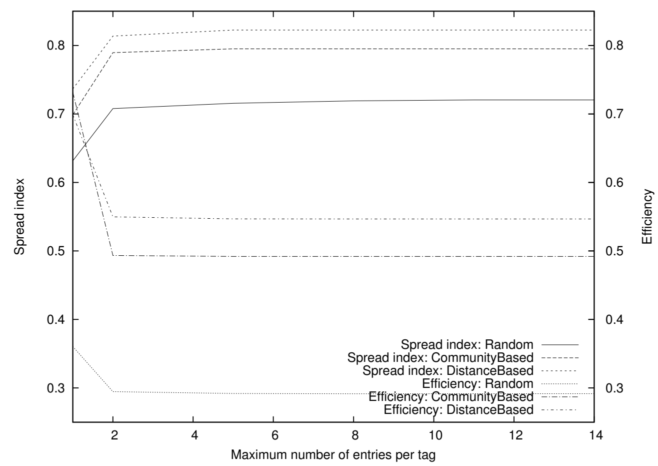

Fig. 3. Effect of a limited number of entries on performance

References

[1] P. Costa, C. Mascolo, M. Musolesi, and G. P. Picco, "Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks," *Selected Areas in Communications, IEEE Journal on*, vol. 26, no. 5, pp. 748–760, 2008.

[2] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[3] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*. Springer, 2011, pp. 338–349.

[4] E. Yoneki, P. Hui, S. Chan, and J. Crowcroft, "A socio-aware overlay for publish/subscribe communication in delay tolerant networks," in *Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems*. ACM, 2007, pp. 225–234.

[5] C. Boldrini, M. Conti, and A. Passarella, "Contentplace: social-aware data dissemination in opportunistic networks," in *Proceedings of the 11th international symposium on Modeling, analysis and simulation of wireless and mobile systems*. ACM, 2008, pp. 203–210.

[6] S. Allen, M. Chorley, G. Colombo, E. Jaho, M. Karaliopoulos, I. Stavrakakis, and R. Whitaker, "Exploiting user interest similarity and social links for micro-blog forwarding in mobile opportunistic networks," *Pervasive and Mobile Computing*, 2011.

[7] E. Borgia, M. Conti, and A. Passarella, "Autonomic detection of dynamic social communities in opportunistic networks," in *Ad Hoc Networking Workshop (Med-Hoc-Net), 2011 The 10th IFIP Annual Mediterranean*. IEEE, 2011, pp. 142–149.

[8] Z. Narmawala and S. Srivastava, "Community aware heterogeneous human mobility (cahm): Model and analysis," *Pervasive and Mobile Computing*, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574119214001977