

Human-centric Cyber Social Computing Model for Hot Event Detection and Propagation

Lei-lei Shi, Lu Liu, *Member, IEEE*,

Yan Wu, Liang Jiang, Muhammad Kazim, Haider Ali and John Panneerselvam

Abstract—Microblogging, networks have gained popularity in recent years as a platform enabling expressions of human emotions, through which users can conveniently produce contents on public events, breaking news, and/or products. Subsequently, microblogging networks generate massive amounts of data that carries opinions and mass sentiment on various topics. Herein, microblogging is regarded as a useful platform for detecting and propagating new hot events. It is also a useful channel for identifying high-quality posts, popular topics, key interests, and high-influence users. The existence of noisy data in the traditional social media data streams enforces to focus on human-centric computing. This paper proposes a Human-centric Social Computing (HCSC) model for hot event detection and propagation in microblogging networks. In the proposed HCSC model, all posts and users are pre-processed through Hypertext Induced Topic Search (HITS) for determining high-quality subsets of the users, topics, and posts. Then, a Latent Dirichlet Allocation (LDA) based multi-prototype user topic detection method is used for identifying users with high influence in the network. Further, an influence maximization is used for final determination of influential users based on the user subsets. Finally, the users mined by influence maximization process are generated as the influential user sets for specific topics. Experimental results prove the superiority of our HCSC model against similar models of hot event detection and information propagation.

Index Terms—Social computing, event detection, event propagation, human-centric.

1 INTRODUCTION

CYBER technology with its wide range of computing and communication devices has a significant impact on our daily life. Cyber-enabled online services [1], [2], [3], [4], [5], [6], [7] have become an integral part of daily activities. Microblogging is one of the popular cyber-enabled online services which aid sharing and disseminating information anytime and anywhere [8], [9], [10], [11]. Being an information sharing platform [9], [12], [13], [14], microblogging also attracts many users on social media to establish friendships [15], exchange ideas, and to promote products. As a consequence, such activities in microblogging generate an enormous amount of data [16] with rich semantic content and structure. The availability of such data has created an opportunity for deeper research investigations in the fields of topic detection [1], [2], influence maximization [17], and information diffusion [18]. Hot event detection and propagation in social media are one of the recently emerging research areas, which predominantly focuses on identifying highly influential users in microblogging networks. High-influence users can be exploited to achieve wider and quicker dissemination of hot events. Hot event detection can be utilised in a wide range of applications such as product promotion [3], friend recommendation [19], and rumour control [20] etc.

Specifically, microblogging hot event detection and propagation emphasize finding most relevant influential speakers in order to propagate popular topics to enhance the effectiveness of hot events dissemination. An efficient hot event detection and propagation should necessarily incorporate the following three steps.

1.1 Choice of Posts

Microblogging generates a large number of posts daily. The ability to find high-quality posts is one of the most important requirements in the hot events dissemination process. Selection of these high-quality posts can exert better influence during hot event detection and propagation.

1.2 Choice of Topics

Usually, the information generated in microblogging can be clustered under different topics. Determination of the suitable topic for every piece of information is another significant requirement in the process of hot event dissemination. Selecting the most appropriate topic areas for information can drastically improve the dissemination effect.

1.3 Spreaders of Topics

Furthermore, spreaders in the network usually play a significant role during the dissemination of hot events. Usually, the most influential users with direct access to relevant target receivers are regarded as the best spreaders in the network. Highly influential spreaders utilizing the inherent power of diffusion of social networks can potentially improve the efficiency of the hot event detection and propagation.

With this in mind, this paper proposes a Human-centric Social Computing (HCSC) model for hot event detection and propagation to mine the most influential spreader sets for popular topics. The proposed model carefully incorporates the aforementioned three

- *Lei-lei Shi, Yan Wu, and Liang Jiang are with The School of Computer Science and Telecommunication Engineering, Jiangsu University, China and also with Jiangsu Key Laboratory of Security Technology for Industrial Cyberspace, Jiangsu University, China*
- *Lu Liu is with the Department of Informatics, University of Leicester, UK. Email: l.liu@leicester.ac.uk (corresponding author: Lu Liu)*
- *Muhammad Kazim, Haider Ali and John Panneerselvam are with The Department of Electronics, Computing, and Mathematics, University of Derby, UK.*

integral requirements to achieve the maximum possible dissemination effect. The proposed HCSC model contains the following four decisive stages. First, all posts and users in the microblogging network are pre-processed using Hypertext Induced Topic Search (HITS) to identify high-quality subsets of users, topics, and posts. Second, a Latent Dirichlet Allocation (LDA) based multi-prototype user topic detection method is applied to recognize the users with high influence under popular topics. Such identified users are used as the initial set of core users for hot event detection and propagation. Third, a two-stage topic popularity-based influence maximization model is deployed for determining the most influential users based on the user subsets. The first stage mines users with the highest topic hub value. The second stage uses this input data to spread event information by the topic popularity-based information propagation model and then mines users with the greatest influence under popular topics. Finally, the users mined by the two-stage process are used as the influential user sets for specific topics. Important contributions of this paper are summarized as follows.

- 1) We propose a Probabilistic Latent Semantic Analysis (PLSA) based hot event detection model named HEE-PLSA that detects hot events based on popular topics. HEE-PLSA filters topic, post, and user information to guarantee high-quality social media big data. HEE-PLSA applies influence maximization based on topic popularity to improve the accuracy and quality of the event detection process for efficient soft computing. The proposed approach creates a smaller, high-quality training dataset by selecting high-quality posts and/or users for popular user interests to reduce the impact of general user interests, garbage posts and ordinary users. An LDA based multi-prototype user topic detection method called Multi-Prototype-LDA (MP-LDA) is developed. The posts and users in each interest carry various weights to describe their degree of representation. The proposed MP-LDA model enables each topic to be represented by more than one post and a single user. The centrality value of posts and users is used for calculating prototype weights while user topic similarity is utilized to help in partitioning the user's topics in the microblogging network.
- 2) Based on the above user topic discovery, a two-stage user topic popularity based hot event propagation model is proposed to further reduce the time required by the classic greedy algorithm to complete rapid mining of the proper number of k key users. The first stage mines users with the highest user topic influence statically and the second stage uses the users identified from the first stage to spread information based on the user topic popularity based information propagation model, and then mine users with largest influence increment greedily. The number of key users is initially allocated in proportion to the size of the topic, whilst mining the key users for each topic in parallel. During the mining process for key users, a high influence greedy strategy is adopted to form an alternative set of key users and from which an optimization based on sub-
- modularity is used to select the key users whom can exert the largest effect.
- 3) Experiments conducted demonstrate the effectiveness of the proposed HCSC model in significantly improving the efficiency of hot event detection and propagation when compared to the existing state-of-the-art methods presented in [1], [8], [9], [10], [21], [22], [23].

The rest of the paper is organized as follows. Section 2 reviews the existing works in event detection and propagation. Section 3 presents our proposed HCSC model. Experimental results are discussed in Section 4 and Section 5 concludes this paper.

2 RELATED WORK

The problem of efficient hot event detection and propagation in social networks is gaining considerable attention in recent years. In general, existing works of event detection and propagation in the social network can be classified into the threshold and independent cascade models [17], [24], [25]. These models simulated the NP-hard problem of influence diffusion as a discrete optimization problem and used a greedy algorithm to determine an approximate solution. In these approaches, the probability of each user being influenced usually increases with their corresponding number of influenced neighbours. A greedy algorithm [24] has been shown to perform considerably better than most of the generous algorithms, random algorithms, and degree of the centrality-based algorithms. The performance of the former greedy algorithm has been claimed to achieve a near optimal solution within the range of $1 - 1/e$.

However, the deployment of the greedy algorithm identified three important problems [17]. First, the algorithm runs longer, as it purely adopts a greedy guiding approach in each step in the global network. Further, this method needs the implementation of the Monte Carlo method in every simulation graph of the global network based on the size of the sphere of influence set. It is worthy of note that modestly sized networks with hundreds of nodes would require a lengthy computation time so that the implementation of the greedy algorithm is not practically feasible for networks with thousands or millions of nodes. Second, it does not exploit the network topology properties such as the power-law distribution of nodes or community structure. Moreover, it does not incorporate the interactive information features such as social factors which differentiate between nodes and inconsistencies between node relationships. Third, not considering the influence of overlapping nodes is a noted drawback of the greedy approach. Each step in the greedy algorithm selects a local optimal node which may not be necessarily the global optimal node. Moreover, this approach of the greedy algorithm might lead to a final node selection, consisting only of locally influential nodes and not globally influential nodes. Such drawbacks of the greedy algorithm necessitate further research to improve the quality of event detection [25].

Goyal et al. [26] developed an optimized algorithm to improve operational efficiency in the event detection process. This algorithm reduced the duration of the core node mining process; however, the influence of the core node has been lower. Nguyen et al. [24] presented a new heuristic algorithm which individually

considered the propagation probabilities of nodes in the network along with the effect of multi-hop neighbours. Subsequently, it achieved a higher influence spread. Moreover, this work proposed a realistic network model with non-uniform propagation probabilities between nodes. However, both these methods do not explicitly consider the popularity and quality of users which makes it hard to propagate the hot events to suitable users.

Liu [17] investigated the influence of maximization and proposed a probabilistic solution. Sepehr et al. [25] proved that selecting the most influential nodes is an NP-hard problem. This work developed an efficient Viral Cascade Probability Estimation (VICE) method that leverages a special importance sampling approximation to achieve high accuracy even in the case of a small probability of influence. Aslay et al. [27] formulated the problem of selecting influential individuals as an optimization problem and developed an algorithm for a diffusion model and improved the efficiency and scalability of the influence maximization algorithms. Zhou et al. [18] used node topic distribution to estimate the node activation probability at the topic-level and proposed an influence maximization algorithm called TopicRank to mine the influential node set under specific topic based on event probability. Some other researchers [1], [20], [22] have also developed algorithms based on this work and improved the time efficiency of mining influential users.

However, in these studies, the influence of a single user on all other users is the same i.e. the activation probability of a user to activate other nodes always stays constant. This is not applicable to the real-world scenario where an individual may be considered highly influential among other peers when discussing the subject economy but completely unknown by peers in the area of sport. In other words, it is rare for any individual to be considered as an expert in multiple fields. Thus, the influence of a user in a social network is likewise related to both the user and the topic while the influence of a given user is different for different topics. A few other works proposed in [22], [28], [29] suffer common limitation, such that they only considered the topic influence on user activation probability and ignored the popularity degree of the topic, links between posts, and the diffusion power of users. Consequently, such drawbacks reflected in their lower efficiency and the improper number of final mining core users.

Few research studies integrated topic popularity degree scoring, user topic detection, and influence maximization to enhance the efficiency of event detection while enlarging the sphere of influence of hot event propagation. Each of the aforementioned studies has not considered the importance of hot event topic information content and user preferences. Literature in the context of event detection demonstrates that the influence of users in the social network is related to the relationship between users and topic while the influence of the same user is dissimilar under different topics [30]. Furthermore, these studies in [11], [23], [31], [32] focus on the effect of topic influence on user activation probability and ignored utilizing the popularity of topics, links between posts, and diffusion power of the users. This leads to the waste of users influence resulting in lower efficiency and the improper number of final mining core users.

To address this, we propose an efficient human-centric soft computing model for Hot Event Detection called HCSC that includes a PLSA based hot events detection model HEEPLSA, a LDA

based multi-prototype user topic detection method, and a two-stage topic popularity based influence maximization for final determination of influential users based on user subsets. We efficiently and accurately mine an appropriate number of most influential spreader sets for hot events propagation and ordinary users are removed during the selection process. Moreover, there is no need to predefine a proper number of topics manually, as our approach automatically detects hot events and identifies influential spreaders under popular topics.

3 A HUMAN-CENTRIC CYBER SOCIAL COMPUTING MODEL

This section details our proposed human-centric social computing model for event detection and propagation.

3.1 HEE-PLSA model

The proposed HEE-PLSA model selects the number of popular topics using the TD-HITS [8] approach while hot events are detected according to the PLSA [28] model ensuring high-quality posts and high-influence users under popular topics.

HITS creates a smaller high-quality training dataset by extracting high-quality posts and influential users from a large pool of posts and users with high efficiency and accuracy. Then we use a topic-decision method to determine the appropriate number of topics and to discover key posts from a large number of posts. Our proposed HEE-PLSA model exploits the TD-HITS method to identify high influence users during hot event detection. Fig. 1 demonstrates the process of hot event detection and identification of high influence users.

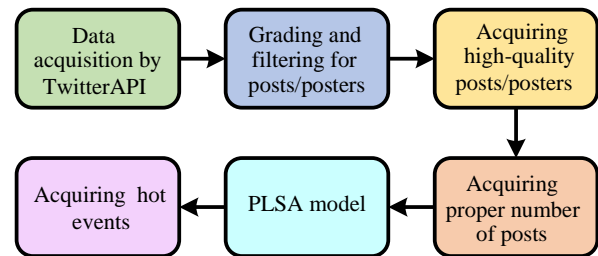


Fig. 1. The processing model of HEE-PLSA

3.2 LDA based multi-prototype user topic detection model

Existing event detection models hardly distilled the popular topics, which results in low quality of posts and users being discovered under popular topics. Therefore, topic filtering method [8] is essential for determining the importance of users under popular topics.

The MP-LDA model is composed of two modules. After identifying the topic scores using the TD-HITS method, the topics are clustered by using the LDA topic model in the post-network and then the Gibbs sampling [8], [9] method is used for user topics detection. Then the initial influential spreaders in influence maximization are discovered based on their hub value in the microblogging network and their local features in the user-user network using TS-LDA [8]. Fig. 2 illustrates the process involved in user topic detection and discovery of initial influential spreaders.

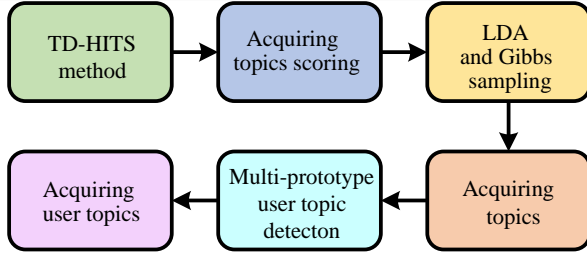


Fig. 2. MP-LDA model

3.3 A two-stage user topic influence maximization model

The two-stage user topic influence maximization model consists of the following two stages. First, a topic popularity-based information propagation based on Independent Cascade Model [17]. Second, a topic popularity-based influence maximization method to determine the proper number of spreaders set with the size of K for popular topics from the microblogging network through which the hot events can be propagated to more users. Fig. 3 depicts the process involved in user topic detection and discovery of influential spreaders in user topics.

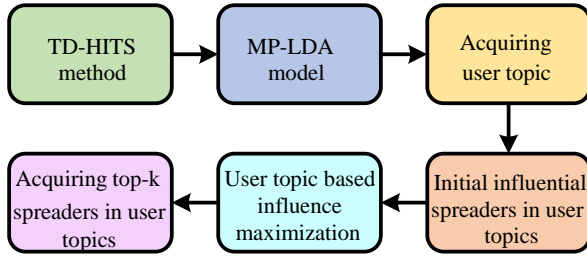


Fig. 3. Two-stage topic popularity-based influence maximization model

Definition 1 (Microblogging Network G): Let V be the set of users in a topic popularity based microblogging network, G , E denotes the set of edges in G , and T shows the set of topics in G . Then topic popularity based microblogging network can be expressed as $G(V, E, T)$, Where $V = \{V_1, V_2, \dots, V_n\}$, $|V| = n$, $E = \{< u, v > | u, v \in V\}$. A link between u and v , it shows an edge between them otherwise $T = \{1, 2, \dots, z\}$ and $|T| = z$.

Definition 2 (Sphere of influence δt): If S is the set of users with K size, then $\delta(S|t)$ denotes the sphere of influence of users set S for topic t i.e. the number of users activated after S propagates information about topic t .

3.4 Mining high influence users under popular topics

In the initial stage, we choose the initial influential spreaders while not considering the information propagation characteristics of the microblogging network. The second stage uses the spreaders obtained from the first stage to spread information, and then iteratively mine top- k spreaders with the largest topic influence increment among the remaining influential users in a greedy way. In other words, an increment in the sphere of influence value of the spreader set after adding a spreader u is ensured to achieve

the maximum effect. Therefore, the topic popularity based influence maximization can be described as follows: given a Microblogging network $G(V, E, T)$, a parameter K and a topic t , then the goal is to find a user set $S(|S| = K)$ in G that can result in a maximum value of $\delta(S|t)$, i.e. $\max S$.

$$S = \max\{\delta(S|t), |S| = K, S \in V, t \in T\} \quad (1)$$

Topic popularity based influence maximization tends to find a high influence user set with K and t from $G(V, E, T)$ through which the information can be propagated to more high-influence users. In order to discover these users, we first model the social network $G(V, E)$ using the topic model HEE-PLSA and obtain the topics T and topic distribution of the users $\psi_{V_i}^T = (\phi_{V_i}^1, \phi_{V_i}^2, \dots, \phi_{V_i}^z)$, $(V_i \in V)$. Then according to the obtained topic popularity based microblogging network $G(V, E, T)$, a topic popularity-based influence maximization algorithm is proposed.

In the influence maximization, the selection of the initial propagation users depends on the information propagation trend and the sphere of influence. High-quality posts can draw more attention from high-influence users spreading and/or broadcasting such posts in the microblogging network. Developing the ability to identify influential spreaders effectively and efficiently in the microblogging network is a major challenge. A large number of important evaluation methods have been proposed to address this problem such as degree centrality [33], clustering coefficient centrality [34], and betweenness centrality [8], [26]. However, degree centrality and clustering coefficient centrality of spreaders can only characterize the local information of networks. The complexity of computing betweenness centrality is very high due to the need to calculate the shortest path. Tweet composition between users and posts in the microblogging network cannot be directly analysed by the application of these centrality methods alone. Other related theories describe that high-influence users are responsible for a large number of high-quality posts. We propose a novel solution to this problem. We believe that the initial influential spreader should satisfy a global condition and a network topology condition within a certain period i.e. a high global feature and a high local feature. Here, we expect that high local feature users will trigger an early and rapid accumulation of contagious transmissions among a large number of candidate users. Finally, the global importance of users can be acquired by the hub value of users along with their posts. Subsequently, high global feature users will spread information, ideas, or rumours much faster than ordinary users.

In general, a user with more neighbour connections is highly likely to be an initial influential spreader in microblogging networks. Using this as inspiration, we propose a novel influence measure by considering the hub value and the degree value of user interaction. Moreover, the method proposed in this paper improves the accuracy of the identification of initial influential spreaders that can further improve the accuracy of the identification of final influential spreaders and the influential scope of influence maximization.

3.5 Topic popularity based hot event information propagation model

This section explains the traditional information propagation i.e. independent cascade (IC) model [17] and our proposed topic popularity-based information propagation model.

IC is the most widely adopted information propagation model used to simulate information propagation in social networks. In the IC model, an active node can activate and/or inactive neighbour nodes in accordance with the activation probability between them. A higher activation probability depicts the level of ease in activating the nodes.

The activation probability in the IC model is generated randomly. However, the activation probability of nodes is related to the social relationships among the nodes and topics in process of information propagation. It is worth noticing that nodes have a different activation probability for different topics. Therefore, we use a topic popularity-based information propagation model to calculate the node activation probability $P_{u,v}^t$ for specific topics to simulate the information propagation in social networks in a more realistic way. $P_{u,v}^t$ is influenced by social connections between the nodes. Closely related to the social connections between nodes, the greater connection times imply a more intimate relationship between nodes and have a higher activation probability. Therefore, user intimacy can be used to represent the degree of intimacy between nodes. Moreover, $P_{u,v}^t$ is also significantly affected by the user's topic popularity. Information is usually quickly propagated between two highly popular users. Therefore, the topic popularity can influence $P_{u,v}^t$ of two users.

Definition 3 (User intimacy): User intimacy, $C_{u,v}$ denotes the frequency of connection between nodes u and v . It can be obtained from the ratio of the connection times of u and v to the connection times of u and other nodes given as follows:

$$C_{u,v} = \frac{R_{u,v}}{\sum_{i=1}^n R_{u,V_i} + \sum_{i=1}^n R_{v,V_i}}, (u,v, V_i \in V) \quad (2)$$

where $R_{u,v}$ shows the connection times of nodes u and V_i , $R_{u,v}$ represents the connection times of nodes u and v .

Definition 4 (Topic popularity):

Topic popularity $TP_{u,v}^T$ shows the popularity degree of two user's topic calculated as follows.

$$TP_{u,v}^T = \frac{Authority_{u,v}^T}{Authority_{max}^{T_i} + Authority_{min}^{T_i}}, (u,v \in V, T_i \in T) \quad (3)$$

where $Authority_{u,v}^T$ denotes the authority of a key post T , $Authority_{max}^{T_i}$ represents the biggest authority of key posts, and $Authority_{min}^{T_i}$ shows the smallest authority of key posts in topics.

Concisely, $P_{u,v}^t$ is influenced by the user intimacy $C_{u,v}$ and $TP_{u,v}^T$. Thus, $P_{u,v}^t$ of a user u to v for specific t can be calculated as follows:

$$P_{u,v}^t = C_{u,v} \times TP_{u,v}^T \quad (P_{u,v}^t \in [0,1]) \quad (4)$$

The propagation process of the TPIP model is similar to the IC model. Each user has only a chance to activate their neighbour users while the user's activation process is independent of each other. The difference is that the user's activation probability of TPIP is dissimilar under different topics which is more in line with the information propagation of microblogging networks.

4 EXPERIMENTS AND RESULTS

We evaluate the efficiency of our proposed HCSC model against four existing topic models including PLSA [28], LDA [8], BEE [23], and EVE [10] to compare the efficiency of HCSC model. We generated our data sets from Twitter (<http://twitter.com/>) via TwitterAPI consisting of 1,500,000 posts and 36,052 users.

4.1 Popular topics detection

Fig. 4 shows the authority value and minimum distance of each post. Fig. 4 can be used to identify the proper number of topics, located in the upper right quadrant, which plays a key role in the spreading of hot events under a chosen specific topic community. TABLE 1 shows high-quality posts ranked by their minimum distance. It is evident from TABLE 1 that our proposed HCSC model can detect the top- k high-quality posts according to their authority value efficiently and effectively. The parameter k is set to 10 according to the number of posts, with the authority value being far greater than others in the right upper quadrant in Fig. 4. At the same time, it can be seen from Table 2 that the key users under popular topic communities can also be detected by our model, which further improves the efficiency of hot event detection and guarantees the high quality of data for efficient soft computing.

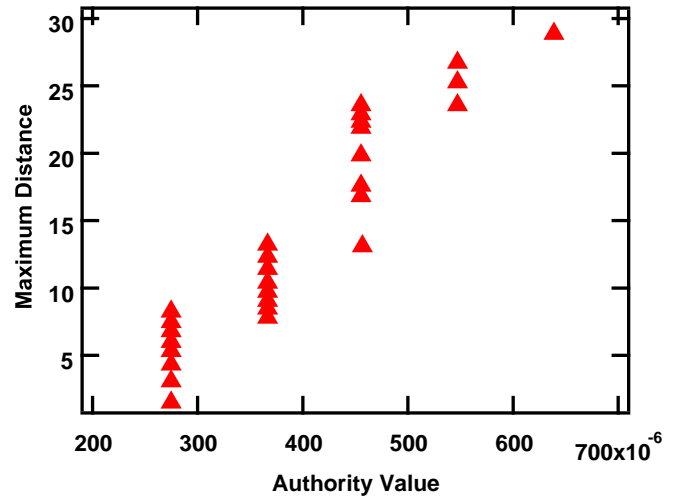


Fig. 4. The number of topics from the HCSC model

TABLE 1 Minimum Distance and Authority of Posts

Post ID	Authority Value	Minimum Distance
681693469564383232	0.000636248	28.79236010
681697568456192001	0.000636248	28.79236010

681695337304702976	0.000545355	26.73948391
681697928268910593	0.000545355	25.29822128
684205783525888002	0.000545355	23.53720459
681696803629219840	0.000454463	23.53720459
681695402928648193	0.000454463	23.51595203
681699684168015873	0.000454463	23.00000000
681697799730249728	0.000454463	22.51666050
681697033523077122	0.000454463	22.49444376

TABLE 2
Key Posts Under Popular Topics

Post ID	Popular Topic
681693469564383232	Basketball
681697568456192001	Basketball
681699684168015873	Basketball
681697799730249728	Basketball
681697033523077122	Basketball
681695337304702976	Music
681697928268910593	Music
681696803629219840	Music
684205783525888002	Economy
681695402928648193	Emotion

4.2 Initial influential spreaders analysis

The degree and hub values of users for topics according to the results achieved from the HEE-PLSA model can help distinguish the importance of users under each popular topic which is better than the BEE and EVE models, as listed in TABLE 3. We can also determine the number of influential users for each popular topic from TABLE 4 by setting a different number of initial influential users. The proper number of initial influential spreaders can also be discovered in Fig. 5 which is better than the baseline approaches. With an increase in the number of initial influential users, the sphere of influence reaches 3,266 when the number of initial influential users is 8 and saturates then after. Hence, the appropriate number of popular topics is set to 4 i.e. ($K = 4$). It verifies the correctness of the appropriate number of popular topics identified in TABLE 4 based on the key posts. The top 10 initial influential spreaders and the popular topics they belong to are listed in TABLE 5, which plays an important role in a specific topic.

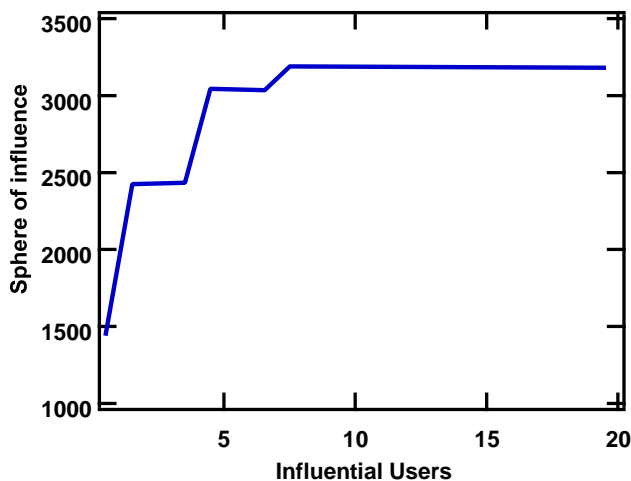


Fig. 5. The proper number of initial influential spreaders

TABLE 3
Degree and Hub Values of Top 10 Influential Users Under Topics

User ID	Hub Value	Degree	Topic
339283603	0.003429355	24535	Basketball
1679619506	0.003233392	2869	Basketball
3693887599	0.003135411	334	Music
933364430	0.002253576	1157	Basketball
4068440360	0.00186165	377	Emotion
1000421510	0.001665687	1458	Music
2168821905	0.001567705	21973	Emotion
3254047099	0.001567705	489	Emotion
2310175028	0.001273761	1778	Music
863205451	0.000979816	44	Conflict

TABLE 4
Number of Users in Popular Topics

Popular Topic	Number of Users
Basketball	1436
Economy	998
Music	589
Emotion	165

TABLE 5
Top 10 Initial Influential Spreaders Mining and relation with Popular Topics

User ID	IF	Popular Topic
339283603	0.051440325	Basketball
1679619506	0.03233392	Basketball
3693887599	0.03135411	Music
933364430	0.020282184	Basketball
1000421510	0.01303155	Music
4068440360	0.011757792	Emotion
1367531	0.011757792	Economy
3254047099	0.011659809	Emotion
2310175028	0.010973935	Music
2168821905	0.010973935	Emotion

4.3 Final sphere of influence

The contrast of the final sphere of influence aids influential spreader discovery. It can be noticed from TABLE 6 that the final sphere of influence of the proposed HCSC model is approximately the same or even better than the IC model. Thus, the influential spreaders discovered by our HCSC model is more appropriate compared to the IC model. This is because of the fact that the activation probability of the nodes considers the topic popularity and intimacy between the nodes in our proposed model.

TABLE 6
Final Sphere of influence of Influence Maximization

User ID	IC	HCSC
339283603	1407	1401
1679619506	1124	1385
3693887599	1014	1089
933364430	1093	1084
1000421510	983	994
4068440360	843	841
1367531	583	593
3254047099	336	338
2310175028	139	143
2168821905	122	126

4.4 Influence maximization precision and efficiency analysis

In order to verify the sphere of influence and operation efficiency of the proposed HCSC model, we run all the baseline algorithms under the same experimental setup and computing platform. Each experiment is repeated 5 times and average values are considered for evaluation. We then compare the similarity of top K users with those discovered by the other two models along with the run time of the studied algorithms under specific topics. The experimental results are summarized in TABLE 7, for $K = 10$, all the top 10 users detected by two models are almost the same. The HCSC model considers the impact of the topic popularity subsequently, the number of influential users under a specific topic is more accurate than the IC model. The HCSC model selects enough nodes with high topic diffusion power as the influential nodes, such that their sphere of influence nearly covers the entire topic area. However, the IC model does not consider these parameters and further produces an improper number of selected nodes under specific topics. Besides, it can be observed from TABLE 8 that the HCSC model characterizes the shortest running time with an appropriate number of influential nodes. The HCSC model preprocesses data to remove users who might not affect the results and to directly mine users characterizing the highest topic diffusion power. Thus, the HCSC performs faster than the IC model. In summary, the proposed HCSC model is superior compared to the IC model in terms of time efficiency whilst possessing the same sphere of influence.

TABLE 7
Final Influential User Comparison

User ID	IC	Topic	HCSC	Topic
1	339283603	Basketball	339283603	Basketball
2	1679619506	Basketball	1679619506	Basketball
3	3693887599	Music	3693887599	Music
4	933364430	Basketball	1000421510	Music
5	1000421510	Music	1367531	Economy
6	4068440360	Emotion	2358523992	Economy
7	1367531	Economy	4068440360	Emotion
8	3254047099	Emotion	2168821905	Emotion
9	2310175028	Music	6158440324	Emotion
10	2168821905	Emotion	1736523801	Emotion

TABLE 8
Comparison of Time Efficiency in Influence Maximization

Method	Time (minutes)			
	TF-HITS	HEE-PLSA	Influence Maximization	Total
IC	0	0	102.55	102.55
HCSC	15.2	32	24.68	71.88

4.5 Event-detection precision and efficiency

To compare the precision and efficiency of our HCSC model with PLSA, LDA, BEE, and EVE models, we performed experiments and calculated the precision as follows:

$$Precision = \frac{a}{b} \tag{5}$$

where a represents the number of detected events matching real-life events and b shows the total number of the events detected by the same algorithm while K is the number of possible events.

TABLE 9 shows the precision results of the five methods (HCSC, PLSA, LDA, BEE, EVE) and TABLE 10 summarizes their time efficiency. The HCSC model can find 6 events when $K = 8$. However, the performance of the PLSA, LDA, BEE, and EVE models are very similar when $K > 6$ such that these models can detect all the hot events only by artificial selection. If $K = 1 - 4$, some hot events would remain undetected. If the $K= 4$ is greater than 8, e.g. 10, these models can detect all of the events but their time complexity quite high. Thus, our proposed HCSC model is accurate, efficient, and outperforms the baseline state-of-the-art models.

TABLE 9
Comparison of Precision in Hot Event Detection

Method	K=1	K=2	K=4	K=8	K=10
PLSA	1/1	2/2	4/4	0	6/10
LDA	1/1	2/2	4/4	0	6/10
EVE	1/1	2/2	4/4	0	6/10
BEE	1/1	2/2	4/4	0	6/10
HCSC	0	0	0	6/8	0

TABLE 10
Comparison of Time Efficiency in Hot Event Detection

Method	Time (Minutes)					
	HITS	Topic Filtering			EM	Total
		Method	Gibbs Sampling			
PLSA	0	0	0	26	26	
LDA	0	0	21.52	0	21.52	
EVE	2.21	0	0	18.62	20.83	
BEE	0	0	0	26	26	
HCSC	2.21	6.09	1.26	8.56	18.12	

5 CONCLUSIONS

This paper proposed an efficient human-centric soft computing model (HCSC) model for hot event detection and propagation. HCSC integrated with the PLSA-based hot events detection model can create a high-quality training dataset from a large collection of users and posts containing noise. Moreover, our approach automatically detects the correct number of topics while efficiently identifying event related key posts with high precision. Furthermore, it also detects critical events by analysing the number of popular user topics and determines the influential spreaders linked to them using the LDA based multi-prototype user topic detection method. The HCSC approach utilizes both posts' and users' information which potentially enables a better understanding in a timely and accurate manner of the users involved in critical incidents by deploying the two-stage user topic popularity based hot event propagation model. Additionally, noisy posts and ordinary users are effectively removed from the selection thus eliminating the need to predefine the proper

number of topics manually, as our proposed model can effectively detect hot events whilst identifying the influential spreaders under popular topics. Experimental results demonstrated the effectiveness and efficiency of our HCSC model against five related approaches developed for guarantying the quality of posts, users, and popular topics discovery. Particularly, our proposed model excels in filtering out noisy data existing in the social network during hot event detection and propagation. Our proposed model suffers limitations when there are changes in the user topic and behaviour evolution during the event propagation. Therefore, future work will focus on user behaviour analysis and predictive analytics for further improving the efficiencies of our proposed hot event detection model.

ACKNOWLEDGMENTS

The research presented in this paper is supported by the National Key R&D Program of China (No. 2017YFE0117500), the Natural Science Foundation of Jiangsu Province under Grant BK20170069, and UK-Jiangsu 20-20 World Class University Initiative program, and UK-Jiangsu 20-20 Initiative Pump Priming Grant. Lu Liu is the corresponding author.

REFERENCES

- [1] Y. Fang, X. Chen, Z. Song, T. Wang, and Y. Cao, "Modelling propagation of public opinions on microblogging big data using sentiment analysis and compartmental models," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 13, no. 1, pp. 11–27, 2017.
- [2] F. Pein, I. Tecuapetla-Gomez, O. M. Schutte, C. Steinem, and A. Munk, "Fully automatic multiresolution idealization for filtered ion channel recordings: Flickering event detection," *IEEE transactions on nanobioscience*, vol. 17, no. 3, pp. 300–320, 2018.
- [3] F. Kalloubi, E. H. Nfaoui, and O. El Beqqali, "Harnessing semantic features for large-scale content-based hashtag recommendations on microblogging platforms," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 13, no. 1, pp. 63–81, 2017.
- [4] Y. Guo, L. Liu, Y. Wu, and J. Hardy, "Interest-aware content discovery in peer-to-peer social networks," *ACM Transactions on Internet Technology (TOIT)*, vol. 18, no. 3, p. 39, 2018.
- [5] L. Liu, N. Antonopoulos, M. Zheng, Y. Zhan, and Z. Ding, "A socioecological model for advanced service discovery in machine-to-machine communication networks," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 15, no. 2, p. 38, 2016.
- [6] L. Jiang, L. Shi, L. Liu, J. Yao, B. Yuan, and Y. Zheng, "An efficient evolutionary user interest community discovery model in dynamic social networks for internet of people," *IEEE Internet of Things Journal*, 2019.
- [7] H. Ali, U. U. Tariq, Y. Zheng, X. Zhai, and L. Liu, "Contention & energy-aware real-time task mapping on noc based heterogeneous mpocs," *IEEE Access*, vol. 6, pp. 75 110–75 123, 2018.
- [8] L. Shi, Y. Wu, L. Liu, X. Sun, and L. Jiang, "Event detection and identification of influential spreaders in social media data streams," *Big Data Mining and Analytics*, vol. 1, no. 1, pp. 34–46, 2018.
- [9] L.-I. Shi, L. Liu, Y. Wu, L. Jiang, and J. Hardy, "Event detection and user interest discovering in social media data streams," *IEEE Access*, vol. 5, pp. 20 953–20 964, 2017.
- [10] X. Sun, Y. Wu, L. Liu, and J. Panneerselvam, "Efficient event detection in social media data streams," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*. IEEE, 2015, pp. 1711–1717.
- [11] G. Tong, W. Wu, S. Tang, and D.-Z. Du, "Adaptive influence maximization in dynamic social networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 25, no. 1, pp. 112–125, 2017.
- [12] L. Jiang, L. Shi, L. Liu, J. Yao, and M. A. Yousuf, "User interest community detection on social media using collaborative filtering," *Wireless Networks*, pp. 1–7, 2019.
- [13] L.-I. Shi, L. Liu, Y. Wu, L. Jiang, and A. Ayorinde, "Event detection and multi-source propagation for online social network management," *Journal of Network and Systems Management*, pp. 1–20, 2019.
- [14] Y. Wu, C. Yan, L. Liu, Z. Ding, and C. Jiang, "An adaptive multilevel indexing method for disaster service discovery," *IEEE Transactions on Computers*, vol. 64, no. 9, pp. 2447–2459, 2015.
- [15] G. Xu-Rui, W. Li, and W. Wei-Li, "Using multi-features to recommend friends on location-based social networks," *Peer-to-Peer Networking and Applications*, vol. 10, no. 6, pp. 1323–1330, 2017.
- [16] P. Yan, "MapReduce and semantics enabled event detection using social media," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, no. 3, pp. 201–213, 2017.
- [17] L. Liu, B. Qu, B. Chen, A. Hanjalic, and H. Wang, "Modelling of information diffusion on social networks with applications to wechat," *Physica A: Statistical Mechanics and its Applications*, vol. 496, pp. 318–329, 2018.
- [18] D. Zhou, W. Han, and Y. Wang, "Identifying topic-sensitive influential spreaders in social networks," *International Journal of Hybrid Information Technology*, vol. 8, no. 2, pp. 409–422, 2015.
- [19] S. Huang, J. Zhang, D. Schonfeld, L. Wang, and X.-S. Hua, "Two-stage friend recommendation based on network alignment and series expansion of probabilistic topic model," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1314–1326, 2017.
- [20] Z. He, Z. Cai, J. Yu, X. Wang, Y. Sun, and Y. Li, "Cost-efficient strategies for restraining rumor spreading in mobile social networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2789–2800, 2017.
- [21] A. Aldhaheri and J. Lee, "Event detection on large social media using temporal analysis," in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2017, pp. 1–6.
- [22] J. Huang, M. Peng, H. Wang, J. Cao, W. Gao, and X. Zhang, "A probabilistic method for emerging topic tracking in microblog stream," *World Wide Web*, vol. 20, no. 2, pp. 325–350, 2017.
- [23] J. Li, Z. Tai, R. Zhang, W. Yu, and L. Liu, "Online bursty event detection from microblog," in *Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*. IEEE Computer Society, 2014, pp. 865–870.
- [24] D.-L. Nguyen, T.-H. Nguyen, T.-H. Do, and M. Yoo, "Probability-based multi-hop diffusion method for influence maximization in social networks," *Wireless Personal Communications*, vol. 93, no. 4, pp. 903–916, 2017.
- [25] A. Sepehr and H. Beigy, "Viral cascade probability estimation and maximization in diffusion networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 3, pp. 589–600, 2019.
- [26] J. Sohn, D. Kang, H. Park, B.-G. Joo, and I.-J. Chung, "An improved social network analysis method for social networks," in *Advanced Technologies, Embedded and Multimedia for Human-Centric Computing*. Springer, 2014, pp. 115–123.
- [27] C. Aslay, L. V. Lakshmanan, W. Lu, and X. Xiao, "Influence maximization in online social networks," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 775–776.
- [28] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth Conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.
- [29] K. Tamura, K. Hirahara, H. Kitakami, and S. Tamura, "Parallel processing of burst detection in large-scale document streams and its performance evaluation," *GSTF Journal on Computing (JoC)*, vol. 2, no. 4, 2018.
- [30] S. Su, Y. Wang, Z. Zhang, C. Chang, and M. A. Zia, "Identifying and tracking topic-level influencers in the microblog streams," *Machine Learning*, vol. 107, no. 3, pp. 551–578, 2018.
- [31] Y. Yilmaz and A. O. Hero, "Multimodal event detection in Twitter hashtag networks," *Journal of Signal Processing Systems*, vol. 90, no. 2, pp. 185–200, 2018.
- [32] Y. Wang, Q. Fan, Y. Li, and K.-L. Tan, "Real-time influence maximization on dynamic social streams," *Proceedings of the VLDB Endowment*, vol. 10, no. 7, pp. 805–816, 2017.
- [33] P. G. Sun and Y. Yang, "Methods to find community based on edge centrality," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 9, pp. 1977–1988, 2013.
- [34] M. G. Campiteli, A. J. Holanda, L. D. Soares, P. R. Soles, and O. Kinouchi, "Lobby index as a network centrality measure," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 21, pp. 5511–5515, 2013.



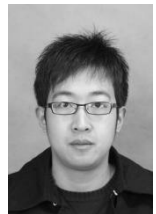
Lei-lei Shi received the B.S. degree from Nantong University, Nantong, China, in 2012, and the M.S. degree from Jiangsu University, Zhenjiang, China, in 2015. He is currently working towards the Ph.D. degree at the School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China. His research interests include Event Detection, Data Mining, Social Computing and Cloud Computing.



Lu Liu is the Professor and Head of Department of Informatics at the University of Leicester, UK. Prof. Liu received his Ph.D. degree from the University of Surrey and M.S. degree from Brunel University. Prof. Liu's research interests are in areas of data analytics, service computing, cloud computing, Artificial Intelligence and the Internet of Things. He is the Fellow of British Computer Society (BCS).



Yan Wu received the M.S. degree from Shandong University of Science and Technology, Qingdao, China, in 2009, and the PhD degree from Tongji University, Shanghai, China, in 2014. He is currently a Lecturer with the School of Computer Science and Telecommunication Engineering at Jiangsu University, China. His research interests include formal methods, service-oriented Computing, and Cloud Computing.



Liang Jiang received the B.S. degree from Nanjing University of Posts and Telecommunications, China, in 2007, and the M.S. degree from Jiangsu University, Zhenjiang, China, in 2011. He is currently working towards the Ph.D. degree at the School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China. His research interests include OSNs, Computer Networks and Network Security.



Muhammad Kazim received his BSc degree in Information and Communication Systems Engineering and MSc degree in Computer and Communication Security from National University of Sciences and Technology Pakistan in 2011 and 2014, respectively. He is currently a KTP associate and PhD candidate at the University of Derby. His research interests include Cloud

Computing, computer networks and computer security.



Haider Ali received the master degree in electronic systems design engineering from Manchester Metropolitan University, U.K. He is currently pursuing the Ph.D. degree with the Department of Electronics, Computing and Mathematics from University of Derby (UoD). He was a Lecturer at COMSATS University Islamabad, Abbottabad Campus, Pakistan from 2011 to 2016. His research area of

interest is electronic and biomedical systems design, image processing, and embedded systems.



John Panneerselvam is a Lecturer in Computing at the University of Derby, United Kingdom. John received his PhD in computing from the University of Derby in 2018 and an MSc in advanced computer networks in 2013. He is an active member of IEEE and British Computer Society and a HEA fellow. His research interests include cloud computing, fog computing, Internet of Things, big data analytics, opportunistic networking and P2P computing. He has won the best paper award in IEEE International

Conference on Data Science and Systems, Exeter, 2018.