Decision Support Systems 81 (2016) 1–11

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A model-free scheme for meme ranking in social media

Saike He [a], Xiaolong Zheng [a,*], Daniel Zeng [a,b]

[a] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] Department of Management Information Systems, University of Arizona, Tucson, AZ 85721, USA

ARTICLE INFO

Article history:
Received 6 March 2015
Received in revised form 9 July 2015
Accepted 3 October 2015
Available online 13 October 2015

Keywords:
Meme ranking
Model-free scheme
Transfer entropy

ABSTRACT

The prevalence of social media has greatly catalyzed the dissemination and proliferation of online memes (e.g., ideas, topics, melodies, and tags). However, this information abundance is exceeding the capability of online users to consume it. Ranking memes based on their popularities could promote online advertisement and content distribution. Despite such importance, few existing work can solve this problem well. They are either daunted by unpractical assumptions or incapability of characterizing dynamic information. As such, in this paper, we elaborate a model-free scheme to rank online memes in the context of social media. This scheme is capable to characterize the nonlinear interactions of online users, which mark the process of meme diffusion. Empirical studies on two large-scale, real-world datasets (one in English and one in Chinese) demonstrate the effectiveness and robustness of the proposed scheme. In addition, due to its fine-grained modeling of user dynamics, this ranking scheme can also be utilized to explain meme popularity through the lens of social influence.

1. Introduction

Meme (pronounced "meem") was first coined by Richard Dawkin in analogy with gene in genetics four decades ago [1]. It is defined as "unit of conceptual replication" that identifies idea, topic or style that spreads from person to person within a culture. Like the natural selection of genes that confer 'differential reproductivity', memes also compete for our scarce individual and collective attention [2]. During this process, some of them quickly die out of popularity while others persist for a long period of time. In recent years, the advent of various social media platforms has lowered the cost of information generation, boosting the potential reach of each meme among online users. This information abundance is exceeding human capacity to consume it [3–5]. Therefore, an effective ranking scheme is imperative to focus human limited attention on the most important memes. Appropriate solutions for this issue would provide direct implications in refining online advertisement and content distribution. In online advertising, new revenue models could be developed to charge advertisers for the amount of attention that a meme will receive. In media outlets, ranking information can be used to highlight the most popular memes in realtime. These condensed results are especially beneficial in emergent situations where information fragments emerge at random moments, such as social events [6,7], public health [8,9], and political campaigns [10–12]. In light of such importance, meme ranking has attracted considerable research interests in various disciplines [2,13,14].

However, to the best of our knowledge, few existing studies provide an adequate solution for meme ranking task. Traditional bottom-up approaches attempt to construct various diffusion models in analogy to behavior replication [15–18], epidemic contagion [14,19–24], or competitive gaming [25,26]. Although these models can help to track the diffusion process and measure its fact on online users, their computational complexities are often comparatively high, sometimes even NP-hard [27]. Furthermore, oversimplifications made in these models, such as user homogeneity [2,16], static network structure [19], and finite interaction patterns [28,29], can lead to unrealistic or even misleading conclusions. Recently, several top-down approaches have been developed to characterize meme dynamics, which is critical in accessing the evolution and mutation of online memes [4]. These studies mainly focus on quantifying the topological centrality [30,31], content similarity [32], or user behaviors [13] based on large-scale datasets. This line of research has provided significant insights in understanding the trend of the web. Yet, lacking fine-grained modeling of user interactions in meme diffusion, they still cannot characterize meme dynamics well.

To solve the above challenges, in this paper, we elaborate a model-free ranking scheme that characterizes meme dynamics with few assumptions. Different from previous work, our ranking scheme is designed based on information theory and could capture complex meme dynamics without modeling its exact diffusion process. In addition, while most existing studies are concerned with aggregate measures for meme ranking, the scheme presented here allows more fine-grained characterization on information diffusion among online users. This key property enables us not only to rank meme at the macro level, but also to inquiry key factors determining meme popularity at the micro level. For evaluation, we have used two different genres of

* Corresponding author.
E-mail addresses: saike.he@ia.ac.cn (S. He), xiaolong.zheng@ia.ac.cn (X. Zheng), dajun.zeng@ia.ac.cn (D. Zeng).

http://dx.doi.org/10.1016/j.dss.2015.10.002

datasets: one from a Chinese microblogging system and the other one from an American political blog forum. Experimental results on these two datasets validate the efficiency and robustness of the scheme compared with several benchmark approaches. By examining two key factors pertaining to meme spreaders, we also uncover several principles governing meme popularity. These findings may provide both academic and industrial implications in understanding other new types of memes such as innovation [15], rumor [19], and viral marketing [14,28].

The remaining parts of the paper are structured as follows. Section 2 reviews existing studies most relevant to our task. In Section 3, the technical details for the proposed meme ranking scheme are represented. Section 4 gives the empirical results of our proposed scheme in comparison with several existing approaches. Finally, Section 5 concludes this paper with a summary and a discussion about future research directions.

2. Literature review

The original work regarding meme traces back to a theory proposed by Dawkin [1], who first coined the concept of meme. This concept is utilized to describe the potential process of information diffusion among online users, in analogy with gene in genetics. In the following part of this section, we will present the existing studies relevant to our work from two perspectives, including meme diffusion and meme ranking.

2.1. Meme diffusion

Existing studies concerning meme diffusion mainly focus on constructing various theoretical models from different views. These models can help us to uncover the potential evolutionary patterns of meme diffusion to a certain extent. Generally, these meme diffusion models can be roughly categorized into three groups, i.e., cascade models, epidemic models, and competitive models respectively.

2.1.1. Cascade models

One of the famous cascade models is proposed by Bikhchandani et al. [16], who explore social changes by assuming all users hold the same belief in behavior making. This assumption clearly does not hold in real-world situations. Kempe et al. [15] then study online innovation diffusion and try to maximize its influence among users by selecting a subset of key nodes. In their cascade model, dynamics of neighbor pairs are considered independently. In fact, user dynamics is highly interwoven. Models for multiple cascades have been studied by extending the existing independent cascade model. These models generally assume that the status of each node keeps intact once influenced by other nodes [17]. Myers and Leskovec [18] further infer social relations based on information propagation in latent social networks. Both the cascades and infections are postulated to be conditionally independent in their propagation model. One common drawback of all these work is that assumptions made in modeling clearly do not hold in real-world practice. In contrast, our model makes no explicit assumptions about information dynamics.

2.1.2. Epidemic models

The epidemical analogy of information to virus has opened a new perspective for investigating meme diffusion and evolution. This, in turn, leads to pervasive applications of compartmental models such as SI, SIR, and SIS [20,21,33]. The spread of rumors and the detection for its source are studied with classical susceptible-infected (SI) model [19]. This model heavily depends on the network structure, which keeps developing and evolving. Some researchers study meme dynamics in the context of personal publishing. Gruhl et al. [23] employ snapshot models to depict topic propagation in blogspace. Their models are designed to characterize dynamics for both the communities and users. Article memes are studied by expressing complex human dynamics in analogy with infection by a virus [22]. These studies often assume the background environment as constant, which is not very practical in real world situations. In another strand of research, Richardson and Domingos [14] seek to optimize viral marketing plans by mining knowledge-sharing websites. In their probabilistic models, only one type of marketing action is considered. This simplicity may run counter to actual marketing scenarios.

2.1.3. Competitive models

To study meme competition among public attention, Weng et al. [2] employ a parsimonious agent-based model. However, their model highly relies on the underlying network structure and does not account for the discrepancy in user interest. Wei et al. [34] try to predict meme prevalence by considering network structural and information propagation at the same time. They assume that all nodes are passive and can be characterized with the same propagation model. Further, mixture of meme effect on individual is forbidden. Such postulation may not reflect the real situation in many circumstances. Goldenberg et al. [28] try to understand personal communications in word-of-mouth marketing. However, their complex system modeling technique could only cope with two types of predefined social interactions.

2.2. Meme ranking

Though there is comprehensive work investigating meme diffusion, to the best of our knowledge, the existing studies concerning meme ranking are comparatively limited. In what follows, we present a brief survey for this line of research.

Ienco et al. [35] and Bonchi et al. [27] initially attempt to construct propagation models to rank memes, but find that these models are pragmatically unfeasible since their computational complexities are NP-hard. Consequently, they turn to employ several heuristic methods. However, these methods cannot distinguish the direction of information flow, which is crucial in determining user importance in meme diffusion. Different from their work, in this paper, we adopt an asymmetric measure that is capable of capturing the direction of information flow among users. In another line of research, Bauckhage [13] ranks memes according to their average daily activity. Since activity level is measured via relative value, the ranking result may be confounded by other memes beyond consideration. Thus, it is highly possible that meme activity increases while its portion drops due to the proliferation of unknown memes.

There also exist other studies trying to rank meme based on topological centrality measures, such as in/out-degree, and number of followers. Gloor [30] measures trends on the web based on betweenness centrality. This measure requires a complete collection of underlying network structure, which is impossible in most scenarios. PageRank is a centrality algorithm that has been used widely in network analysis and ranking related tasks. Rather than prioritizing authoritative blogs, Adar et al. [31] try to rank blogs from the perspective of information diffusion. To this end, they propose an iRank algorithm to rank blogs based on implicit link structure. Their approach requires additional resource to train a link predictor, whose performance highly relies on the quality of this resource. However, such resource is not always available in real world practice, thus limiting its applications on a wide scale. Gordevicius et al. [32] focus on ranking news stories. Instead of using hypertext-links, they construct implicit links based on content similarity. Their algorithm is computational expensive, since it is equivalent to obtaining the stationary distribution of a random walk over a whole graph. Besides, the ranking result varies based on the similarity measure used. In contrast, except for user behaviors, our approach does not need any extra information. In addition, it computational complexity is also acceptable.

Recently, studies on meme ranking turn to explore dynamic information. One of the significant studies is presented by Kwak et al. [36], who attempt to rank trending topics based on singleton, reply, mention,