Computer & Robotics

A Knowledge Management Approach to Discovering Influential Users in Social Media

Hosniyeh Safi Arian[a], Mohammad Jafar Tarokh [b,*]

[a] Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[b] Associate professor, IT Group - Faculty of Industrial Engineering, K. N. Toosi University of Technology Tehran, Iran

Received 29 June 2016; revised 21 September 2016; accepted 11 January 2017; available online 15 March 2018

Abstract

A key step for success of marketer is to discover influential users who diffuse information and their followers have interest to this information and increase to diffuse information on social media. They can reduce the cost of advertising, increase sales and maximize diffusion of information. A key problem is how to precisely identify the most influential users on social networks. In this paper, we propose a method to discover influential users based on knowledge management cycle that is called KMIU. The knowledge management cycle consists of several stages including capture, organize, storage, retrieval and mining stages. We try to analyze influential users in two micro bloggings networks as Facebook and twitter by KMIU method. The experimental results showed the proposed method maximize diffusion and has an accuracy 0.55. These maximization and accuracy are more than those of the previous methods.

Keywords: *Influential Users, Diffusion, Knowledge Management, Social Networks, Marketing.*

1. Introduction

Recently businesses renew the structural of productions, services and marketing tools in order to satisfy customer needs. They have to use marketing innovation strategy or advertising and improve the service for customers.

Companies have new strategy for attracting of their customers.one of new strategy is social media that it is a strong tool for marketing since the diffusion of information is very fast in social media and users can realize them in one minuet and also the growth of interaction is a sign a positive opportunity for advertisers and marketer in social media. Researchers can use user's data for analyzing behavioral of them in social media and also they inferred customer's interest to choice product. On the other hand these analyzing of these data can reduce the cost of advertising and segment customers accurately.

One of the advertising challenges increases enhancement and broad of advertisements in social media. Empirical studies of diffusion has two major difficulties, first influence spread is unobservable in social media and second the observational of data is complex in diffusion.

In this paper we identify influential users who disproportionately impact the likelihood that information will spread broadly [1]. In this paper for identifying them we used knowledge management system. Knowledge management systems support process of eliciting explicit or tacit knowledge from people, artifacts, or organizational

* Corresponding author. Email: mjtarokh@kntu.ac.ir

entities. Scholars have paid increasing attention to the role of knowledge in gaining competitive advantage leading to the emergence and development of the knowledge-based view of strategic management [2].

The rest of this paper is organized as follows. In the next section, some of the most important related works are reviewed. In section 3, the structure of proposed method for discovering influential users in social media .The evaluation of this method is discussed in section 4. Finally, last section presents conclusion.

2. Related Work

By the emergence of social networks in recent years, finding in Influential users has absorbed a considerable amount of attention from researchers in this area. In this section, we review the literature separately on finding influential users on SNWs.

Customer Networks studied to discover the problem of identifying influential customers. The network value of a customer in a social network is the profit due to additional sales to customers he or she may influence to buy [3]. Co-authorship data studied viral marketing problem using several commonly used diffusion models such as the Linear Threshold, Independent Cascade and General Threshold Models. They discover influence maximization problem is NP hard and also presented a greedy algorithm for influence maximization problem [1]. Large water distribution network studied the scalability of greedy approach of influence maximization this algorithm called CELF [4]. Two real life collaboration networks are studied to reduce running time of the greedy algorithm of [1] and its improvement by [4]; this algorithm is called Mix greedy [5].Real-world datasets studied to optimized computation and improved quality of seed selection by Vertex Cover Optimization and Look Ahead Optimization SIMPATH can active nodes more than another algorithms [6]. The mobile social network studied to reduce computational cost of greedy algorithm of influence maximization problem by using community-based approach. The authors emphasized that none of previous work to use community based approach for influence maximization and also greedy algorithm is not suitable for large mobile social network.so they proposed CGA algorithm, this algorithm has two features: 1) an algorithm for detecting communities based on information diffusion and 2) a dynamic algorithm to find influential nodes from these communities [7] .Yahoo! Meme and a prominent online news site studied the scalability issue of influence maximization problem by pruning the social network graph. The goal of which is to identify sub networks that preserve properties of a given network. This algorithm called SPINE [8].the researcher tend to acquire information on social network such as reading a blog, posting a picture, rating a movie etc.[3] The structure of action log and definition of terms, such as action propagation and user influence graph studied in [9] and These terms are frequently used in the work done [10] frequent pattern approach by [9]to discover leaders (or influential nodes) from social network graph and its action log [10].The work of [10] where they used action log to trace action propagation to learn influence probability that can be used for influence maximization algorithms such as greedy of [1] and [11].Dynamic of influence in different topics studied on three measures: in degree, retweet and mention [12] .Facebook studied on influential users based on users' activities on a social graph and present a method they empirically validated positive effects of influencers on spreading social games among Facebook users [13].Telecom Company studied on adopted centrality measures in selecting influencers. They selected customers from a telecom company with different centrality measures and evaluated these measures according degree, hubs, page rank in the network reached by these initial customers [14].Social marketing messages investigated on level of popularity and they used number of LIKE in popularity of message on Facebook fans [15]. Yahoo investigated Temporal features are time-related properties, such as the time of creation temporal [16]. The summarized of this section showed table 1.

Table. 1. Representative Studies on influential users

Reference	Dataset	Methods
[1]	Co-authorship	Greedy algorithm presented for influence maximization problem
[4]	Large water distribution network	The scalability of greedy algorithm improved by CELF algorithm
[5]	real life collaboration networks	Reduce to run time of the greedy algorithm by mix greedy algorithm
[6]	Real world	optimizing the computation and improving the quality of seed selection by SIMPATH algorithm
[7]	mobile social network	The CGA algorithm presented by community-based approach
[9]	Yahoo! Meme and a prominent online news	scalability issue of influence maximization problem by pruning the social network by SPINE algorithm
[8]	Face book	Discovering influential users based on

		Degree centrality, number of groups belonged to, pages linked, and updated per day
[12]	Yahoo! Question	Discovering influential users based on Content characteristics, User attributes Patterns of interaction among users
[13]	Face book	Discovering influential users based on Content of the messages and Media type of the post
[14]	telecom company	Discovering influential users based on degree, hubs, page rank
[15]	Facebook fans	Discovering influential users based on Number of like measurer
[16]	Yahoo	Discovering influential users based on temporal features as time

3. Proposed Method

In this section, we proposed a method for discovering influential users based on their actions. When users registered on social networks, they initiated to interact with other persons, they try to find their friends in virtual world, join to diversity groups and follow other persons, and they generate content such as picture, URL linked, news. They share contents between friends. We used knowledge management cycles for discovering influential users in proposed method.in the following, we explain our method, this method called KMIU.

3.1. Knowledge Management Cycle

The knowledge management consist of five phases: Capture, Organization, Storage, Retrieval and Mining that is showed in figure 1 [17]. Successful knowledge management systems can create and capture new knowledge by combining existing knowledge in new and interesting ways.

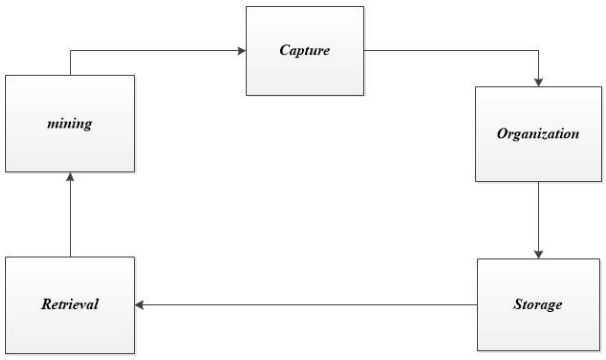

Fig. 1. Knowledge management cycle

3.1.1.Capture Stage

Knowledge capture systems support process of eliciting, explicit or tacit knowledge from user's profile, user's interaction or user's actions .Rely on mechanisms and technologies to support externalization and internalization.

Knowledge capturing social network consist of image, contents and linked and other user's data. The Knowledge capture social network solution is able to deliver measurable cost saving through reduction in the storage costs and administration associated with rapidly expanding data in databases.

All of the interactions captured based on figure 2 .this stage includes receive, analysis,

Release to workflow and segment of interaction.

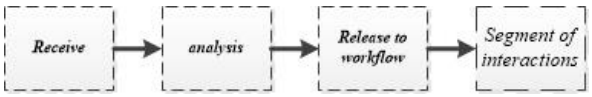

Fig. 2. The process of storage knowledge

3.1.2.Organization Stage

In this stage knowledge divided to three organization consist of: structural, semi structural and UN structural showed in figure 3.

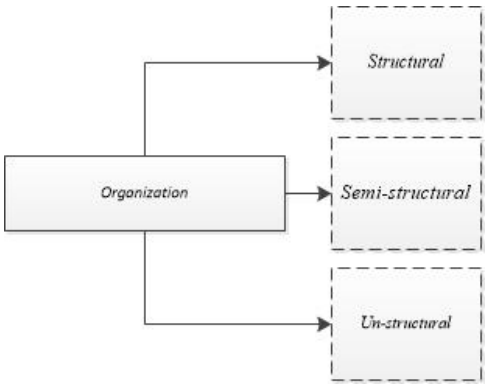

Fig. 3. The organization of knowledge

Structural knowledge resides in a fixed field within a record or file Structure, [18]. Structural knowledge is based on a model. This model consist of what fields of knowledge will be stored and how that knowledge will be stored. Structural knowledge has the advantage of being easily entered, stored, queried and analyzed. In the real world, data are incomplete, noisy and Inconsistent. The task of data preprocessing consist of Data cleaning, data integration, transformation and reduction Structure,[18] . Unstructured knowledge is all those things that can't be so readily classified and fit into a field same as images [19]. Semi structural knowledge is between structural knowledge unstructured knowledge [19] .The preprocess of unstructured knowledge showed in figure 4. Text cleaning Consist of stemming, removal of punctuations, removal of expressions, split attached words, removal of URLS and decoding data.

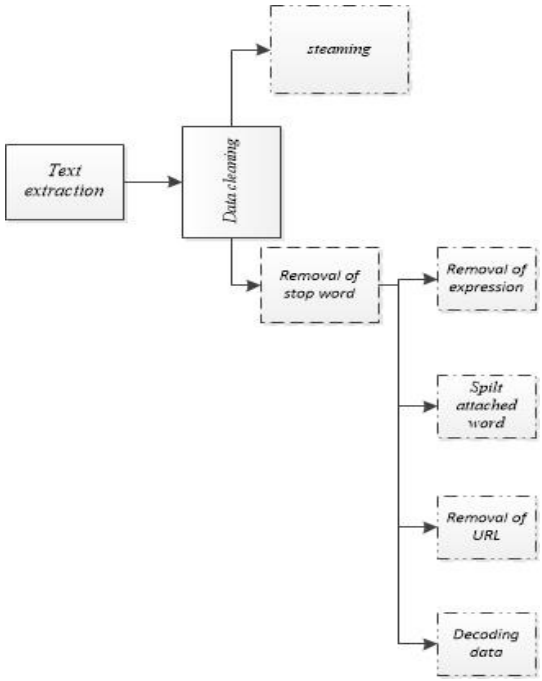

Fig. 4. The preprocess of UN structural knowledge

3.1.3.Knowledge Stage

There are diversity conceptual models for knowledge storage. In this paper we use Bayesian Net diagram for this stage. Bayes' nets use Bayesian probabilities as a way of capturing the uncertainty associated with different sets of events. Bayes' nets embed those probabilities into a diagram that captures how things are causally related [20].

3.1.4.Retrieval Stage

The main purpose is to retrieve what is useful while leaving behind what is not.

Structural, semi structural and un-structural information are retrieved by Vector Space Model [21]. The basic idea is first create a vector space, whose dimensionality is equal to the number of terms are in the corpus. Each document is mapped to a vector, whose component reflects the corresponding term's weight in that document. This weight is calculated based on term-frequency in that document called TF that calculated by equation (1) [21] and the term's important factor called IDF that calculated by equation (2) [22].

$$tf_{m,p} = \frac{\text{freq}_{m,p}}{\max_l\left(\text{freq}_{l,p}\right)} \quad (1)$$

In equation (1) where $\text{freq}_{m,p}$ is the raw frequency of term i appearing in post p and $\max_1(\text{freq}_{1,p})$ is the number of times the most frequent index term, l, appears in post j.

$$idf_m = \frac{\log N_p}{\log N_m} \quad (2)$$

In equation (2) Where Np is the total number of posts and n_m is the number of posts in which term m appears.

Finally, the query itself is also mapped to a vector and the similarities between query and documents are calculated according to some similarity function. The results are output in a similarity ranked order.

3.1.5.Mining Stage

In this stage we want to providing interfaces to allow your human subject matter experts to discover new relationships in massive data sets. First we identify important features in improving diffusion, Second we discover influential users.

3.1.5.1. Important Features in Improving Diffusion

Users interact with each other on social networks. These interactions consist of like, mention, share, hashtag and another feature that can improve diffusion on social networks. The effect of these features are different, some features are important more than another features. Multiple-regression is used for discovering features and their importance. Multiple-regression analysis is an appropriate method for developing a prediction model and analyzing the relationship between the features and influence score in equation (3) [23]:

$$y = b_0 + b_1(x_1) + b_2(x_2) + .. b_p(x_p) \quad (3)$$

Y is response variable, x_p is predictor variables and b_p is clustering coeffient between response variable and every predictor variables in equation (4) [23]:

$$\left(\frac{r_{y,x_p} - r_{yx_{p+1}} r_{x_p x_{p+1}}}{1 - (r_{x_p x_{p+1}})^2}\right)\left(\frac{SD_x}{SD_y}\right) \quad (4)$$

SD_x And SD_y are Standard deviation for x and y.r is clustering coefficient between y and x_p.

3.1.5.2. Discovering Influential Users

The goal of this stage identifies influential users based on the important of their actions. To identify influential users, we computed influence score of user's actions and also we fit regression tree model based on greedy optimization process recursively partitions the feature space [24]. In figure three user's actions are important in influence score of diffusion than another

user's actions that consist of number of followers (>μ) the number of posts (>π) and number of mentions (>α).these user's action are constraints in regression tree, where the left (right) child is followed if the condition is satisfied (violated). Leaf nodes give the predicted influence the corresponding partition. In Figure 3 For instance a user have more than 130 followers and they generate post more than 500 and they mentions by another followers more than 210, almost certainly they are influential users with probability 0.9715.

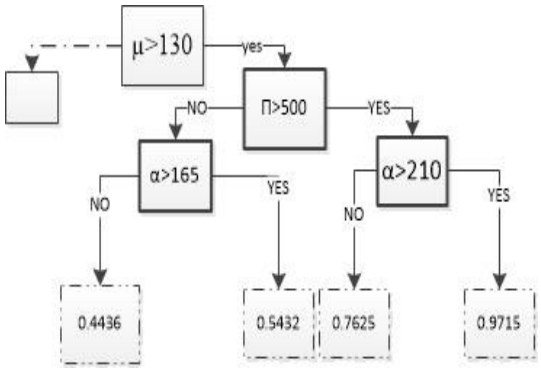

Fig. 5. The example of regression tree for discovering influential user

4. Evaluation

4.1. Network Datasets

- *face book*

The Facebook dataset consist of friends list. Facebook data was collected from survey participants using this Facebook app. The features of dataset showed table 2.

Table. 2. Summaries of properties of face book dataset

Parameters name	Value of parameters
Number of users	215
Number of contents	1654
Average number of like	2495
Average number of Share	1120
Average number of comment	3256
Average number of mention	16854
Average number of friends	46530

- *Twitter*

Twitter is a social news website. We used to API Twitter to analyzing influential users. The features of dataset showed table 3.

Table. 3. Summaries of properties of Twitter dataset

Parameters name	Value of parameters
Number of users	124
Number of contents	985
Average number of like	1150
Average number of Share	695
Average number of comment	869
Average number of mention	465
Average number of friends	1654

4.2. Method Compared

KMIU is compared with method [13] and method [14]:

Method [13]: This method is for influential user discovering. The feature of this method consist of: Degree centrality, number of groups belonged to, pages linked, and updated per day.

Method [14]: This method is for influential user discovering. The features of this method consist of: Centrality (degree, hubs, page rank).

4.2.1.Performance Analyzing

Figure 6 shows the influence spreads of KMIU generated from Facebook dataset. They maximize diffusion in decision tree. The number of influential users is 29.they when influential users are least than 11, the performance of KMIU is very closely method [13] and method [14] more than 11 influential users.

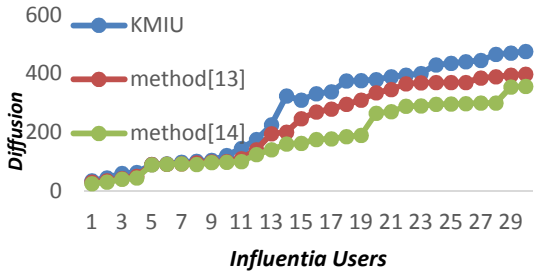

Fig. 6. Influence spreads of Kim et al. (2009) method and Kiss et al. (2008) method and KMIU method on Facebook Dataset

And also in figure 7 shows the influence spreads of KMIU generated from twitter dataset, the number of influential users are 14 they maximize diffusion in decision tree. KMIU has maximize influence than another two methods.

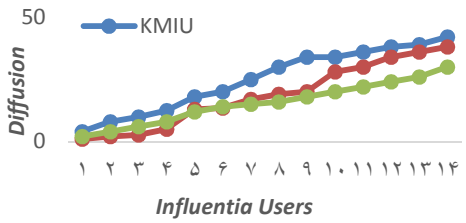

Fig. 7. Influence spreads of Kim et al. (2009) method and Kiss et al. (2008) method and KMIU method on twitter Dataset

4.2.2. Accuracy and F-Measure Analysis

We use accuracy and F1-measures to evaluate the performance of the method. The accuracy and f-measure calculated [24]:

$$\text{Accuracy} = \frac{\#\text{TP} + \#\text{TN}}{\#\text{TP} + \text{FP} + \text{FN} + \#\text{TN}} \quad (5)$$

$$recall = \frac{\#\text{TP}}{\#\text{TP} + \text{FN}} \quad (6)$$

$$\text{precision} = \frac{\#\text{TP}}{\#\text{TP} + \text{FP}} \quad (7)$$

$$\text{F} - \text{meauer} = \frac{2.\text{precision}.\text{recall}}{\text{precison} + \text{recall}} \quad (8)$$

In equation (5), (6) and (7) , variables are showed by table 4:

Table. 4. The variable of recall and precision

		Predicted condition	
True condition	condition positive	True positive	False Negative
	condition negative	False Positive	True negative

The experimental results are shown in Table 4. The results were obtained by 10-fold cross-validation. It showed that our KMIU method gives better and more robust prediction than method [13] and method [14]. The accuracy of KMIU method was 0.55 more than method [13] method [14].

Table. 5. Accuracy F1-measure of the KMIU method, method [13] and method [14]

	Method[13]	**Method[14]**	**KMIU method**
Accuracy	0.514560	0.546502	0556304
f_1-measuer	0.68054	0.690543	0.714651
f_2-measuer	0.58647	0.58765	0.598765
f_3-measuer	0.315158	0.460148	0.587279
F1-measuer	0.60102	0.610123	0.621495

5. Conclusion

As discussed, we consider influential users on social networks. We believe that influential users can maximize diffusion on social networks. Diversity methods introduced for detecting influential users and reviewed them.

We proposed a method based on knowledge management cycle that called KMIU. This method consist of: capture stage, organization stage, knowledge storage, retrieval storage and mining stage.in mining stage we used multiple regression for detecting important user's action on social networks and discovering influential users based on important actions.

In the empirical study, KMIU method consider with method [13] and method [14] in twitter dataset and Facebook dataset. In two datasets, the scale of diffusion KMIU is more than method [13] and method [14] and the accuracy of KMIU method is 0.55.

References

[1] Kempe, D.; Kleinberg, J.; Tardos, É., "Maximizing the spread of influence through a social network". In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 137-146, ACM (2003, August).

[2] Eisenhardt, K. M.; Santos, F. M., "Knowledge-based view: A new theory of strategy". Handbook of strategy and management, 1, pp. 139-164, (2002).

[3] Bonchi, F.; Castillo, C.; Gionis, A.; Jaimes, A., "Social network analysis and mining for business applications". ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 22, (2011).

[4] Leskovec, J.; Adamic, L. A.; Huberman, B. A., "The dynamics of viral marketing". ACM Transactions on the Web (TWEB), 1(1), 5 (2007).

[5] Chen, W.; Wang, C.; Wang, Y., "Scalable influence maximization for prevalent viral marketing in large-scale social networks". In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1029-1038, ACM (2010, July).

[6] Goyal, A.; Lu, W.; Lakshmanan, L. V., "Celf++: optimizing the greedy algorithm for influence maximization in social networks". In Proceedings of the 20th international conference companion on World Wide Web. Pp: 47-48, ACM (2011, March).

[7] Wang, Y.; Cong, G.; Song, G.; Xie, K., "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks". In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1039-1048, ACM (2010, July).

[8] Mathioudakis, M.; Bonchi, F.; Castillo, C.; Gionis, A.; Ukkonen, A., "Sparsification of influence networks". In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '11. ACM (2011).

[9] Goyal, A.; Bonchi, F.; Lakshmanan, L. V., "Discovering leaders from community actions". In Proceedings of the 17th ACM

conference on Information and knowledge management. pp. 499-508, ACM (2008, October).

[10] Goyal, A.; Bonchi, F.; Lakshmanan, L. V., "Learning influence probabilities in social networks". In Proceedings of the third ACM international conference on Web search and data mining. pp. 241-250, ACM (2010, February).

[11] Goyal, A.; Bonchi, F.; Lakshmanan, L. V., "A data-based approach to social influence maximization". Proc. VLDB Endow. 5, pp. 73–84 (2011).

[12] Cha, M.; Haddadi, H.; Benevenuto, F.; Gummadi, P. K., "Measuring User Influence in Twitter: The Million Follower Fallacy". ICWSM, 10 pp. 10-17, 30 (2010).

[13] Kim, E. S.; Han, S. S., "An analytical way to find influencers on social networks and validate their effects in disseminating social games". In Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in (pp. 41-46). IEEE (2009, July).

[14] Kiss, C.; Bichler, M., "Identification of influencers—measuring influence in customer networks". Decision Support Systems, 46(1), pp. 233-253, (2008).

[15] Yu, B.; Chen, M.; Kwok, L., "Toward predicting popularity of social marketing messages". In Social Computing, Behavioral-Cultural Modeling and Prediction (pp. 317-324). Springer Berlin Heidelberg (2011).

[16] Hong, L.; Dan, O.; Davison, B. D., "Predicting popular messages in twitter". In Proceedings of the 20th international conference companion on World Wide Web. pp. 57-58 ACM (2011, March).

[17] Kinney, T., "Knowledge management, intellectual capital and adult learning", Adult Learning, 4(1), pp. 2-5 (1998).

[18] Structure, M., "Meaning: Is "unstructured" data merely unmodeled". Intelligent Enterprise, March, 1 (2005).

[19] The Penn database group has structured and XML data project

[20] Friedman, N.; Geiger, D.; Goldszmidt, M., "Bayesian network classifiers". Machine learning, 29(2-3), pp. 131-163 (1997).

[21] Salton, G.; Wong, A.; Yang, C. S., "A vector space model for automatic indexing". Communications of the ACM, 18(11), pp. 613-620 (1975).

[22] Salton, G.; Wong, A.; Yang, C. S., "A vector space model for automatic indexing". Communications of the ACM, 18(11), pp. 613-620 (1975).

[23] Finkelstein, M.O.; "The judicial reception of multiple regression studies in race and sex discrimination cases". Columbia Law Review, pp.737-754, (1980).

[24] Salton, G., "Developments in automatic text retrieval." Science 253.5023 (1991): 974.