Southeast-Asian J. of Sciences, Vol. 7, No. 2 (2019) pp. 133-141

MODELING USER'S INTERESTS, SIMILARITY AND TRUSTWORTHINESS BASED ON VECTORS OF ENTRIES IN SOCIAL NETWORKS

Dinh Que Tran[1], **Thi Hoi Nguyen**[2]
and
Phuong Thanh Pham[3]

[1] *Department of Information Technology*
Posts and Telecommunications Institute of Technology (PTIT),
Hanoi, Vietnam

[2] *Department of Informatics,*
Thuongmai University, Hanoi, Vietnam

[3] *Department of Mathematics and Informatics,*
Thanglong University, Hanoi, Vietnam
E-mail: tdque@yahoo.com; hoi@gmail.com; ppthanh216@gmail.com

Abstract

The purpose of this paper is first to present vectorial representations of user's entries and interests in topics in social networks. Based on such vectorization of short texts, we propose three interest measures of users. And then we investigate the relationships among interest degrees, similarity and trustworthiness of users based on these measures. Some preliminary studies on these correlations are exhibited.

Key words: social networks, text processing, decision support, distributed systems, artificial intelligence, reliability.
2010 AMS Mathematics classification: 911D30, 91D10, 68U115, 68U35, 68M14, 68M115, 68T99.

1 Introduction

Social media has been becoming an important source of information to spread knowledge, trends, news, and services to users on Internet. The resources of *entries* have been elicited and analyzed to determine interest subjects and trust degrees of users. These issues have attracted a large number of research interests ([5] [6] [7] [3] [4] [8] [9]). Most of these studies make use of the vector model in some form for representing texts and classifying users.

Along with this approach, in this paper, we utilize the technique of tf-idf ([5] [6]) to compute the weight of word in a document for the vector representation of entries and topics as well. Based on such a vector model, we construct similarity measures and interest degrees. Then we study various methods for estimating trustworthiness of users via these interest degrees. We also investigate if there are correlations among similarity degrees of users, their own interests and trustworthiness. This paper is considered as an extension and a continuation of our previous researches ([12] [13] [14]).

The remainder of this paper is structured as follows. Section 2 describes vector representation of entries and topics. Section 3 presents models of user's interests based on similarity and correlation measures. Section 4 is devoted to formulating the similarity of users and their interests. Section 5 covers correlation between interests, similarity and trust computation. Conclusions are presented in Section 6.

2 Representing Entries and Topics in Vector

The vectorial model for representing texts by means of tf-idf has been widely used in various fields of the computer science such as the information retrieval and text mining ([2] [1]). This section is to reformulate the model in some formal way for the object of our paper. The purpose is to apply the approach to vectorizing entries and topics with word weights in texts. The n-gram technique for extracting a text into terms or words being applied in text analysis will not be reminded here. And from now on, in this paper, any document or text is always considered as a set of terms.

2.1 Vector Representation of Documents

Definition 1. *Given a collection of documents* $\mathcal{D} = \{D_1, \ldots, D_p\}$*, each of which is represented as set of terms or words* $D_i = \{d_{i1}, \ldots, d_{ip_i}\}$*. Let* $\mathcal{V} = \{v_1, \ldots, v_q\}$ *be a set of distinct terms in the collection. The weight of term* $d \in \mathcal{V}$ *w.r.t.* D_i *is defined as follows:*

$$w_d = tf(d, D_i) \times idf(d, \mathcal{D}) \tag{1}$$

where $tf(d, D_i)$ *is the number of times the term* d *appears in* D_i *and* $idf(d, \mathcal{D}) = log(\frac{\|\mathcal{D}\|}{1+\|\{D_i|d\in D_i\}\|})$.

Each D_i is then represented by means of a vector in weights of terms. For convenience in computation, the vector is normalized so that its length belongs to interval $[0, 1]$.

Definition 2. *Given a collection of documents* $\mathcal{D} = \{D_1, \ldots, D_p\}$*, each of which is a set of terms* $D_i = \{d_{i1}, \ldots, d_{iq_i}\}$*. Let* $\mathcal{V} = \{v_1, \ldots, v_q\}$ *be the set of distinct terms in the collection. Each* D_i *is then represented with a normalized* q *dimension vector* $\mathbf{w_i} = (w_{i1}, \ldots, w_{iq})$ *being called the weight vector of the document* D_i *w.r.t. the corpus* $\mathcal{D}$*.*

2.2 Vector Representation of Entries and Topics

In this paper, an *entry* is a short piece of text, briefly a short text, being dispatched from some user to make a description or post information/idea/opinions on an item such as a comment, a paper, a book, a film, a video, etc. These short texts will be used as resources for classifying users according to similarity of their entries or topic interests. This section is devoted to presenting the weighted vector representation of such entries and topics.

Denote $\mathcal{U} = \{u_1, \ldots, u_n\}$ to be a set of users on a social network. In some temporal interval, each user owns a set of entries in the form of short texts $E_i = \{e_{i1}, \ldots, e_{in_i}\}$, denote $\mathcal{E} = \{E_1, \ldots, E_n\}$. Suppose that $\mathcal{T} = \{T_1, \ldots, T_p\}$ is a set of topics, in which each topic is defined as a set of terms or words. From **Definition 2**, we can construct weight vectors for topics and user's entries as follows.

Definition 3. *Given a collection of topics* $\mathcal{T} = \{T_1, \ldots, T_p\}$ *in which each topic is defined as a set of terms or words. Let* $V_T = \{v_1, \ldots, v_q\}$ *be a set of* q *distinct terms in all* T_i*. A topic vector is a weight one w.r.t. each topic* T_i *being defined as follows*

$$\mathbf{t_i} = (w_{i1}, \ldots, w_{iq}) \tag{2}$$

where $w_{ik} = tf(v_k, T_i) \times idf(v_k, \mathcal{T})$*,* $v_k \in V_T$ *as defined from* **Definition 1**.

Definition 4. *Suppose that* e_{ij} *is an entry of terms dispatched by* u_i*. An entry vector w.r.t. topics* $\mathcal{T}$ *is a weight one being defined as follows*

$$\mathbf{e_{ij}} = (e_{ij}^1, \ldots, e_{ij}^p) \tag{3}$$

where $e_{ij}^k = tf(v_k, e_{ij}) \times idf(v_k, E_i)$*,* $v_k \in V_T$ *as defined from* **Definition 1**.

Thus, from **Definition 3** and **Definition 4**, we have a sequence of topic vectors $\mathbf{t_1}, \ldots, \mathbf{t_p}$ and a sequence of entry vectors $\mathbf{e_{i1}}, \ldots, \mathbf{e_{in_i}}$ w.r.t. topics $\mathcal{T}$

and entries $E_i = \{e_{i1}, \ldots, e_{in_i}\}$ dispatched by u_i. These vectors are utilized for constructing the model of user's interests based on similarity, which is presented in the next section.

3 Modeling Users and Interests based on Entries and Topics

Suppose that $\mathcal{E} = \{E_1, \ldots, E_n\}$ is the set of entries dispatched by users $\mathcal{U} = \{u_1, \ldots, u_n\}$. Denote $E_i = \{e_{i1}, \ldots, e_{in_i}\}$ to be entries given by u_i and $\mathcal{P}(E_i)$ to be a set of all subsets of E_i and $\mathcal{P}(\mathcal{E}) = \bigcup_i \mathcal{P}(E_i)$.

3.1 Similarity and Pearson Correlation Measures

For easily following the paper, this subsection presents two measures, which are widely used in classification techniques and clustering as well [2]. The one is based on the cosine of two vectors and the other one is the Pearson correlation measure.

Given two vectors $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$. Cosine similarity and Pearson correlation measures are defined respectively by the following formulas:

$$sim(\mathbf{u}, \mathbf{v}) = \frac{< \mathbf{u}, \mathbf{v} >}{\|\mathbf{u}\| \times \|\mathbf{v}\|} \tag{4}$$

where $< \mathbf{x}, \mathbf{y} >$ is a scalar product and $\|\mathbf{x}\|$ is the Euclidean length of a vector and

$$cor(\mathbf{u}, \mathbf{v}) = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i (u_i - \bar{u})^2} \times \sqrt{\sum_i (v_i - \bar{v})^2}} \tag{5}$$

where $\bar{u} = \frac{1}{n}(\sum_{i=1}^{n} u_i)$ and $\bar{v} = \frac{1}{n}(\sum_{i=1}^{n} v_i)$. It is clear that values of the function $sim(x, y)$ belong to the interval $[0, 1]$, whereas values of $cor(x, y)$ are in $[-1, 1]$. We may make use of the function $f(x) = \frac{(x+1)}{2}$ to bound values of function $cor(x, y)$ into the unit interval $[0, 1]$.

3.2 Interest Degrees of Users on Topics

Based on the above measures, we can define similar or correlation degrees among entries and topics. Denote

$$\alpha_{ij}^{k} = cor(\mathbf{e_{ij}}, \mathbf{t_k}) \tag{6}$$

to be correlation degrees of the entries e_{ij} given by u_i w.r.t. topics t_k. Each e_{ij} is then represented by correlation degrees $cor(e_{ij}, \mathcal{T}) =< \alpha_{ij}^{1}, \ldots, \alpha_{ij}^{p} >$.

Definition 5. *Given $0 < \epsilon \leq 1$. An entry e_{ij} is called ϵ-entry w.r.t. topic t_k if and only if $cor(\mathbf{e_{ij}}, \mathbf{t_k}) \geq \epsilon$.*

Before constructing user's interest degrees, we take an observation that

- When the amount of entries given by some user with the same topic increases, his interest degree in that topic does as well;
- When the number of users are concerned about some topic increase, the topic is more noticeable.

We can define user's interest degree as follows

Definition 6. *The function $int : \mathcal{U} \times \mathcal{P}(E) \times \mathcal{T} \to [0, 1]$ is called the interest one iff it satisfies the condition that $int(u, U, t) \leq int(u, V, t)$, for all $U, V \in \mathcal{P}(E_u)$ such that $U \subseteq V$.*

For simplicity in the presentation, we omit parameters U, V and denote the interest function on a topic to be $int(u_i, t)$. It is easy to prove the following proposition.

Proposition 1. *The functions defined by the following formulas are interest ones:*

(i) $intMax(u_i, t) = \max_j(cor(\mathbf{e_{ij}}, \mathbf{t}))$

(ii) $$intCor(u_i, t) = \frac{\sum_j cor(\mathbf{e_{ij}}, \mathbf{t})}{\|E_i\|}$$

(iii) $$intSum(u_i, t) = \frac{1}{2}\left(\frac{n_i^t}{\sum_{l \in \mathcal{T}} n_i^l} + \frac{n_i^t}{\sum_{u_k \in \mathcal{U}, l \in \mathcal{T}} n_k^l}\right)$$

where n_i^t is the number of ϵ-entries concerned about the topic t given by u_i.

These functions define user's interest degrees in various topics. They are utilized for constructing the similarity of users in their interests which is considered in the next section.

4 Similarity of Users and their Interests

4.1 Similarity of Users

Given two users u_i, u_j with sets of entries $E_i = \{e_{i1}, \ldots, e_{in_i}\}$ and $E_j = \{e_{j1}, \ldots, e_{jn_j}\}$, respectively. Let V_{ij} be a set of distinct terms occurring in E_i

and E_j. From **Definition 2**, we can construct vectors $\mathbf{e_{il}}$, $\mathbf{e_{jk}}$ and a sequence of similarity values $sim(\mathbf{e_{ik}}, \mathbf{e_{jl}})$. And then similarity of users in entries is defined as follows

Definition 7. *Given two users u_i, u_j with sets of entries $E_i = \{e_{i1}, \ldots, e_{in_i}\}$ and $E_j = \{e_{j1}, \ldots, e_{jn_j}\}$, respectively. Similarity of users in entry is defined as follows*

$$sim_{ent}(u_i, u_j) = \max_{k,l}(sim(\mathbf{e_{ik}}, \mathbf{e_{jl}})) \tag{7}$$

It is easy to see that

Proposition 2. *Given two users u_i and u_j with sets of entries $E_i = \{e_{i1}, \ldots, e_{in_i}\}$ and $E_j = \{e_{j1}, \ldots, e_{jn_j}\}$, respectively. We have the following equality*

$$sim_{ent}(u_i, u_j) = sim_{ent}(u_j, u_i) \tag{8}$$

4.2 Interest Similarity of Users

Denote $u_i^t = int(u_i, t)$ to be interest degree of u_i in topic t as proposed in **Proposition 1**. Then each peer u_i is defined as a vector of interests on various topics.

Definition 8. *Degrees of user's interest on all topics is defined as a vector*

$$\mathbf{u_i^t} = (u_i^1, \ldots, u_i^p) \tag{9}$$

in which u_i^k is the interest degree of user u_i in topics $t_k \in \mathcal{T}$ $(k = 1, \ldots, p)$.

Thus the following matrix represents interest degrees of users on topics

$$\begin{array}{c} \\ \mathbf{u_1^t} \\ \mathbf{u_2^t} \\ \vdots \\ \mathbf{u_n^t} \end{array} \begin{array}{cccc} t_1 & t_2 & \cdots & t_p \\ u_1^1 & u_1^2 & \cdots & u_1^p \\ u_2^1 & u_2^2 & \cdots & u_2^p \\ \vdots & \vdots & \ddots & \vdots \\ u_n^1 & u_n^2 & \cdots & u_n^p \end{array}$$

Based on this interest degree we can construct a similar measure in interests as follows:

Definition 9. *Similarity degree in interest of two peers u_i and u_j is defined as a cosine similarity of two vectors $\mathbf{u_i}$ and $\mathbf{u_j}$*

$$sim_{int}^t(u_i, u_j) = \frac{< \mathbf{u_i^t}, \mathbf{u_j^t} >}{\|\mathbf{u_i^t}\| \times \|\mathbf{u_j^t}\|} \tag{10}$$

in which $< u, v >$ is the scalar product, $\times$ is the usual multiple operation and $\|.\|$ is the Euclidean length of a vector.

5 Correlation of Trust, User Interests and Similarity

5.1 Trust based on User's Interests and Interaction

This subsection is to present an extension of the definition on topic trust estimation that has been proposed by ourselves ([13] [14]).

Definition 10 ([14]). *A function* $trust_{topic} : \mathcal{U} \times \mathcal{U} \times \mathcal{T} \to [0,1]$ *is called a topic trust function, in which* $[0,1]$ *is an unit interval of the real numbers. When given a source peer* u_i*, a sink peer* u_j *and a topic* t*, the value* $trust_{topic}(i,j,t) = u_{ij}^t$ *means that* u_i *(truster) trusts* u_j *(trustee) of topic* t *w.r.t. the degree* u_{ij}^t*.*

Definition 11 ([14]). *Experience trust of user* u_i *on user* u_j*, denoted* $trust^{exp}(i,j)$*, is defined by the formula*

$$trust^{exp}(i,j) = \frac{\|I_{ij}\|}{\sum_{k=1,k\neq i}^{m} \|I_{ik}\|} \tag{11}$$

where $\|I_{ik}\|$ *is the number of connections* u_i *with each* $u_k \in \mathcal{U}$*.*

Based on the degrees of interaction of user's interests, we can define the *experience topic trust* for sink peers of u_i as follows.

Definition 12. *Suppose that* $trust^{exp}(i,j)$ *is the experience trust of* u_i *on* u_j *and* $intX(j,t)$ *is the interest degree of* u_j *on the topic* t*. Then the experience topic trust of* u_i *on* u_j *of topic* t *is defined by the following formula:*

$$trust_{topic}^{exp}(i,j,t) = \gamma \times trust^{exp}(i,j) + \delta \times intX(j,t) \tag{12}$$

where $\gamma, \delta \geq 0,\ \delta + \gamma = 1$ *and* $intX(j,t)$ *is the interest function defined in* **Proposition 1** *.*

5.2 Correlation

This subsection is to investigate the correlation of similarity measures and trust estimation.

- Is there any relationship in trustworthiness between two users which are similar in topic interests?
- Is there any correlation between two same users with similarity in interest topic?

It is easy to see from **Definition 12**

Proposition 3. *Suppose that* $intX(j,t)$ *and* $intX(k,t)$ *are interest degrees of* u_j *and* u_k *in topic* t*, respectively. If*

(i) $intX(j,t) \geq intX(k,t)$ *and*

(ii) $trust^{exp}(i,j) \geq trust^{exp}(i,k)$

then $trust^{exp}_{topic}(i,j,t) \geq trust^{exp}_{topic}(i,k,t)$.

The following proposition shows that the more two users are similar, the more trustful they are.

Proposition 4. *For every* $\epsilon > 0$*, there exists* $\eta > 0$ *such that if* $sim^t_{int}(j,k) > \eta$ *then* $\|trust^{exp}_{topic}(i,j) - trust^{exp}_{topic}(i,k)\| < \epsilon$

Following are statements that need to be confirmed via experimental evaluation

Statement 1.
If $sim^t_{int}(i,j) \geq sim^t_{int}(i,k)$, then $trust^{exp}_{topic}(i,j,t) \geq trust^{exp}_{topic}(i,k,t)$ for all t.

Statement 2.
$sim_{ent}(i,j) \geq sim_{ent}(i,k)$ if and only if $sim^t_{int}(i,j) \geq sim^t_{int}(i,k)$ for all t.

6 Conclusions

This paper has presented the vectorial model for representing topics and entries dispatched by users in social networks. By means of such vectors, we have defined the measures of similarity, correlation of entries and topics. And then, we have constructed estimations of interest, similarity and trust degrees of users. We also show that there are relationships of measures in user's interest, similarity and trustworthiness. These studies should be investigated furthermore and conducted experimental evaluation as well. The research results will be presented in our future work.

References

[1] D. Manning, Prabhakar Raghavan, Hinrich Schutze, "Introduction to Information Retrieval", 2013.

[2] Bing Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Springer-Verlag Berlin Heidelberg, 2011.

[3] Bo Jiang1 and Ying Sha1, *Modeling Temporal Dynamics of User Interests in Online Social Networks,* Inter. Conference On Computational Science, **51** (2015), pp. 503-512.

[4] Jaeyong Kang and Hyunju Lee, *Modeling user interest in social media using news media and wikipedia,* Information Systems, Vol.65, April 2017, pp. 52-64.

[5] E. Gabrilovich and S. Markovitch, *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis,* IJCAI, 2007. Available at: https://www.aaai.org/Papers/IJCAI/2007/IJCAI07-259.pdf

[6] Hitesh Sajnani et al., *Multi-Label Classification of Short Text: A Study on Wikipedia,* Association for the Advancement of Artificial Intelligence, 2011. Avaiable at: https://www.ics.uci.edu/ hsajnani/Publications/AAAI2011.pdf

[7] A. Yildirim et al., *Identifying Topics in Microblogs Using Wikipedia,* March 18, 2016.

[8] C. De Booma et al., "Representation learning for very short texts using weighted word embedding aggregation, Pattern Recognition Letters", Elsevier 2016. Available at: https://arxiv.org/pdf/1607.00570.pdf

[9] Abhishek Gattani et al., *Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-Based Approach,* Proceedings of the VLDB Endowment, Vol.6, No.11, 2013.

[10] Wang, P., Hu, J., Zeng, HJ. et al., *Using Wikipedia knowledge to improve text classification,* Knowl Inf Syst (2009). Available at: https://doi.org/10.1007/s10115-008-0152-4

[11] Wanita Sherchan, Surya Nepal, and Cecile Paris, *A survey of trust in social networks.* ACM Computing Survey, **45**(4):47:147:33, August 2013.

[12] Manh Hung Nguyen and Dinh Que Tran, *A commbination trust model for multi-agent systems,* International Journal of Innovative Computing, Information and Control, **9**(6) (2013), 2405-2420.

[13] Dinh Que Tran and Phuong Thanh Pham, *Integrating interaction and similarity threshold of user's interests for topic trust computation,* Southeast Asian Journal of Sciences, **7**(01) (2019), pp. 28-35.

[14] Dinh Que Tran, *Computational topic trust with user's interests based on propagation and similarity measure in social networks,* Southeast Asian Journal of Sciences, **7**(01) (2019), 18-27.