Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 642 – 651

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Similarities of Frequent Following Patterns and Social Entities

Kyoji Kawagoe[a]*, Carson Kai-Sang Leung[b]

[a] *Ritsumeikan University, Kusatsu, Shiga, JAPAN*
[b] *University of Manitoba, Winnipeg, MB, CANADA*

Abstract

Social network sites such as Twitter and Facebook are used for sharing knowledge and information among users. As social networks grow larger, it becomes difficult for a user to find frequently followed group of social entities. Recently, the frequent following pattern (FFP) mining concept and method were proposed to extract patterns of the relationship between a set of following entities and their most frequently followed entities. In this paper, we propose two similarity definitions: FFP similarity and FFP-based Entity (FbE) similarity. These similarities can be used to recommend a new appropriate social entity to a "read-only-user". In other words, these similarities can be defined only with followed-and-following (F-F) relationships and without additional information such as entity characteristics or entity access logs. To the best of our knowledge, this is the first attempt to define these similarity definitions for social entity recommendations. Some examples show the effectiveness of our similarity definitions by checking their satisfaction of established requirement.

Keywords: Social network; frequent following pattern; similarity; social entity; recommendations.

1. Introduction

Social networks which are composed of many social entities such as individuals, organizations, user-groups, corporations, and their associations, have been widely used to acquire and share knowledge and information among social entities. The exchange of information between social entities is supported not only by social network

* Corresponding author. Tel.: +81-77-561-3904; fax: +81-77-561-5203.
E-mail address: kawagoe@is.ritsumei.ac.jp

doi:10.1016/j.procs.2015.08.200

infrastructures but also by a social computing environment including social behavior studies and social information mining.

Among the many social network sites, Twitter is one of the most popular services that allow users to write and read short messages called “tweets”. A Twitter user can follow other users to possibly receive their favorite tweets. It is very important for a user to know the most popularly followed users, referred to as “followees”, in order to receive interesting, up-to-date or useful information from them. For any Twitter user, it might be effective and important to follow some group of users that many of his or her friends are currently following. In addition to Twitter, many social network services, such as Facebook and Google+, allow users to add others as friends, link to their personal pages, and create or join user groups.

Because the number of social network service users grows rapidly by the day, most of them needs to carefully select their followees, because little adequate information can be given in short messages enough for the selection, and browsing messages from a huge number of social entities is time-consuming. Many data mining methods used for social networks have been proposed [2-15], including topic discovery [2], community detection [3], graph-mining [4], popular friend discovery [5], influential friend discovery [6], and [7]. However, no methods aim to recommend a social entity suitable for so-called lurkers (i.e., "read-only-users") who never send tweets but only read others’. It should be noted that there is little information used for recommendation to these read-only-users because they tend to send few messages and provide little profile information.

The common characteristics of links that are provided by many social network services are non-mutuality in connections. That is, even if user U1 is a follower or a subscriber of another user U2, it is unknown whether the user U2 is also a follower or a subscriber of the user U1. U2 may or may not be a follower of U1. Therefore, the followed-and-following relationship between a follower and followees is quite different from the conventional friendship relationship.

In this paper, we propose two novel similarity methods, Frequent Following Patterns (FFP) [1] and FFP-based Entity (FbE) similarities in order to recommend a social entity suitable for read-only-users with no textual information but only using non-mutual followed-and-following (F-F) relationships. Because no tweets have been sent by read-only-users, collaborative filtering [17] and content-based filtering [18] are not effective in follower recommendation for Read-Only-Users. Moreover, the existing graph mining methods, such as matrix factorization [19] and graph matching [20] methods, cannot be used for this recommendation, because neither ranking data nor any kinds of features are available. FFP similarity is a similarity between two patterns extracted from FFP mining [1]. With this similarity, it can be calculated how much similarity exists between two FFPs are. The FbE similarity is a similarity between two social entities calculated from FFP similarity values for the related entities. Using the FbE similarity, a particular social entity can receive some recommended social entities obtained by calculating the FbE values of the entity and then by extracting those which have FbE values over some given threshold.

Our important contributions are: (i) the first proposal regarding FFP similarity to calculate the similarity between two FFPs and (ii) the first proposal regarding social entity similarity using the FbE similarity obtained from using only F-F relationships, with no additional information.

In this paper, the concept of FFP is briefly described in Section 2. In Sections 3 and 4, FFP similarity and FbE similarity are proposed using their formal definitions as well as examples. After describing comparative analysis and related works in Sections 5 and 6, conclusions are given in Section 7.

2 Frequent Following Pattern (FFP)

The following pattern is one of following-and-followed (F-F) relationships, which is composed of a set of followers and their set of followees. For example, in the case that a social entity E1 is following another social entity E2, there is one F-F relationship between E1 and E2. In this relationship, E1 and E2 are referred to as a follower and followee, respectively. The entity E2 follows many followees in addition to E2, while E2 receives many messages from followers including E1. It is natural that E2 is not following E1. An FFP is represented by a set of followers and a set of followees, subject to the conditions where the number of followers is not less than the given minimum follower values and the number of followees is also not less than the given minimum followee value.

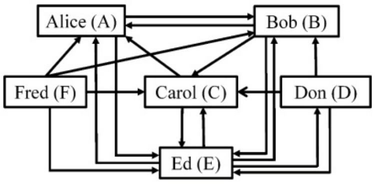

Fig. 1. FFP example [1]..

Fig. 1 shows an example of several entities and the F-F relationships among them. There are six entities, from A to F, and 18 F-F relationships among them, represented as directed edges. If both the minimum followers and the minimum followees are set to two, the following five FFPs can be extracted from the five F-F relationships. These are labelled sequentially from FFP1 to FFP5, using the FFP representation as followers → followees.

FFP1:{EF}→{ABC}, FFP2:{BF}→{ACE}, FFP3:{DF}→{BCE}, FFP4:{BCF}→{AE}, FFP5:{BEF}→{AC}.

For example, the first FFP1, {EF}→{ABC}, means that both entities (E and F) follows three entities (A, B, and C). This pattern meets the conditions because the numbers of followers and followees are two and three, respectively. As in Fig. 1, there are three edges directed from entity E to other entities (A, B, and C). There are also three edges directed from entity F to the same entities (A, B, and C). In FFP1, all the possible combinations between two nodes (E and F) and three nodes (A, B, and C) exist in these F-F relationships shown in Fig. 1. Although it is true that {EF}→{AB} is an FFP where {AB} is a subset of this FFP followees {ABC}, only the maximum frequent pattern {EF}→{ABC} among the related FFPs can be selected. Other FFPs are extracted in the same way. Note that FFP5 differs from FFP1 although the followee set {AC} of FFP5 is a subset of the set {ABC} of FFP1, because the followers set {EF} of FFP1 is different from the follower set {BEF} of FFP5. We treat FFP1 and FFP5 as different FFPs. We use the notations {A} and {AB} to represent an entity set containing only A and an entity set containing both A and B, respectively.

Suppose that there exists a set of FFPs, extracted from a large number of F-F relationships using a FFP mining method, denoted by $F = \{f_i\}$. For each FFP f_i in F, a pattern is represented by a follower set S_i and a followee set $T_{i,}$ as f_i: $S_i \rightarrow T_i$, where S_i is a subset of all social entities. E and T are also subsets of all social entities E, that is, $S_i \subseteq E$ and $T_i \subseteq E$. We use {X} to show an entity set, $X \subseteq E$, and also use {XY} to show the union of the set X and Y, $\{XY\} = X \cup Y \subseteq E$.

There are several important concepts on FFP to be described here, which are (i) FFP basic properties and (ii) FFP closure. With respect to FFPs, the following basic properties are defined.

- If {X} → {Y} and {X} → {Z} are FFPs, {X} → {YZ} is also FFP.
  For example, the minimum number of followees is set to one in Fig. 1. In this case, by matching that {X} = {BCF}, {Y} = {A}, and {Z} = {E}, if {BCF} → {A} and {BCF} → {E} are FFPs, then {BCF} → {AE} is also FFP.
- If {X} → {YZ} is FFP, {X} → {Y} and {Y} → {Z} are also FFPs. For example, For example, the minimum number of followees is also set to one in Fig. 1. If {BCF} → {AE}, then {BCF} → {A} and {BCF} → {E} are also FFPs.
- If {X} → {Z} and {Y} → {Z} are FFPs, {XY} → {Z} is also FFP.
  For example, the minimum number of followers is set to one in Fig. 1. In this case, by matching {X} = {E}, {Y} = {F}, and {Z} = {ABC}, if {E} → {ABC} and {F} → {ABC} are FFPs, then {EF} → {ABC} is also FFP.
- If {XY} → {Z} is FFP, {X} → {Z} and {Y} → {Z} are also FFPs.
  For example, the minimum number of followees is also set to one in Fig. 1. If {EF} → {ABC} is FFP, then {E} → {ABC} and {F} → {ABC} are also FFPs.
- It is not true that if {X} → {Z} is FFP, {XY} → {Z} is FFP. There are many cases when this property does not hold. In the figure 1, although {EF} → {ABC} is FFP, {DEF} → {ABC} is not FFP.

- It is also not true that $\{X\} \rightarrow \{Z\}$ is FFP, $\{X\} \rightarrow \{YZ\}$ is FFP. There are many cases when this property does not hold. In the figure 1, although $\{EF\} \rightarrow \{ABC\}$ is FFP, $\{EF\} \rightarrow \{ABCD\}$ is not FFP.
- It is not true that $\{X\} \rightarrow \{Y\}$ and $\{Y\} \rightarrow \{Z\}$ are FFPs, $\{X\}\rightarrow \{Z\}$ is FFP. There are many cases when this property does not hold. In Fig. 1, by setting both the minimum followees and followers are set to one, $\{D\} \rightarrow \{B\}$ and $\{B\} \rightarrow \{A\}$ are FFPs, but $\{D\} \rightarrow \{A\}$ is not FFP.

On consideration of these properties, because neither the usual transitivity rule nor the augmentation rule hold in FFP, finding FFP closure is very difficult. Hence, the FFP closure is defined as follow using the above properties.

Given a FFP f_i, $S_i \rightarrow T_i$, in F, a closure of F with respect to f_i, denoted as $F^+(f_i)$, is defined, as follows. $F^+(f_i) = \{S_i' \rightarrow T_i' \mid |S_i'| \geq$ (given minimum number of followers), $|T_i'| \geq$ (given minimum number of followees), S_i' are following T_i', $S_i \cap S_i' \neq \{\}, T_i \cap T_i' \neq \{\}\}$. Finally, the closure F^+ of F is defined as all the union of $F^+(f_i)$, that is, $F^+=\cup_{fj\in F} F^+(f_i)$.

For example, when FFP1:$\{EF\} \rightarrow \{ABC\}$ is given, a closure of F with respect to this FFP, is $F^+(\{EF\} \rightarrow \{ABC\}) = \{\{EF\} \rightarrow \{ABC\}, \{EF\} \rightarrow \{AB\}, \{EF\} \rightarrow \{BC\}, \{EF\} \rightarrow \{AC\}$, in the case that both the minimum followers and followees are two. After obtaining all the $F^+(f_i)$ for the five FFPs as in the above example, $F^+= \{\{EF\} \rightarrow \{ABC\}, \{EF\} \rightarrow \{AB\}, \{EF\} \rightarrow \{BC\}, \{EF\} \rightarrow \{AC\}, \{BF\} \rightarrow \{ACE\}, \{BF\} \rightarrow \{AC\}, \{BF\} \rightarrow \{CE\}, \{BF\} \rightarrow \{AE\}, \{DF\} \rightarrow \{BCE\}, \{DF\} \rightarrow \{BC\}, \{DF\} \rightarrow \{BE\}, \{DF\} \rightarrow \{CE\}, \{BCF\} \rightarrow \{AE\}, \{BC\} \rightarrow \{AE\}, \{BF\} \rightarrow \{AE\}, \{BEF\} \rightarrow \{AC\}, \{BE\} \rightarrow \{AC\}\}$.

3 FFP similarity

3.1 Preliminary definitions and some requirements

FFP similarity is a similarity function $SIM_{FFP}(f_i, f_j)$ between two patterns f_i and f_j. By using this FFP similarity, we can calculate the amount of similarity between two FFP. For instance, when $\{XY\} \rightarrow \{PQ\}$, $\{XZ\} \rightarrow \{RS\}$, and $\{VW\} \rightarrow \{PQ\}$ are given as FFPs, it would be difficult to know whether $\{XY\} \rightarrow \{PQ\}$ is more similar to $\{VW\} \rightarrow \{PQ\}$, rather than to $\{XZ\} \rightarrow \{RS\}$, without FFP similarity calculation. Before describing the FFP similarity definition, some requirements for this FFP similarity are listed below.

1. (RQ1) $0.0 \leq SIM_{FFP}(f_i, f_j) \leq 1.0$
   The value of FFP similarity is between 0.0 and 1.0 for simplicity.
2. (RQ2) $SIM_{FFP}(f_i, f_i) = 1.0$
   The self-similarity value is always 1.0.
3. (RQ3) $SIM_{FFP}(f_i, f_j) = SIM_{FFP}(f_j, f_i)$
   The similarity value is the same even if the order of two FFPs is described in reverse.
4. (RQ4-1) $SIM_{FFP}(f_1, f_2) = 1.0$ when $f_{1:}\ \{S\} \rightarrow \{T_1\} \in F$ and $f_{2:}\ \{S\} \rightarrow \{T_2\} \in F$, or (RQ4-2) $SIM_{FFP}(f_1, f_2) = 1.0$ when $f_{1:}\ \{S\} \rightarrow \{T_1\} \in F$ and $f_{2:}\ \{S\} \rightarrow \{T_1T_2\} \in F$
   This requirement indicates that, when the follower set of two FFPs is the same, the similarity value of these two FFPs is the highest.
5. (RQ5-1) $SIM_{FFP}(f_1, f_2) = 1.0$ when $f_{1:}\ \{S_1\} \rightarrow \{T\} \in F$ and $f_{2:}\ \{S_2\} \rightarrow \{T\} \in F$, or (RQ5-2) $SIM_{FFP}(f_1, f_2) = 1.0$ when $f_{1:}\ \{S_1\} \rightarrow \{T\} \in F$ and $f_{2:}\ \{S_1S_2\} \rightarrow \{T\} \in F$
   This requirement also indicates that when the followee set of two FFPs is the same, the similarity value of these two FFPs is the highest.
6. (RQ6-1) $SIM_{FFP}(f_1, f_2) \geq SIM_{FFP}(f_1, f_3)$ when f_1: $\{S_1\} \rightarrow \{T_1\} \in F$, f_2: $\{S_1S_2\} \rightarrow \{T_2\} \in F$, f_3: $\{S_2\} \rightarrow \{T_3\} \in F$, $S_1 \cap S_2=\{\}$, $T_1 \cap T_2=\{\}$, and $T_1 \cap T_3=\{\}$, or
   (RQ6-2) $SIM_{FFP}(f_1, f_2) \geq SIM_{FFP}(f_1, f_3)$ when f_1: $\{S_1\} \rightarrow \{T_1\} \in F$, f_2: $\{S_1S_2\} \rightarrow \{T_2\} \in F$, f_3: $\{S_2\} \rightarrow \{T_3\} \in F$, $S_1 \cap S_2=\{\}$, $T_1 \cap T_2=\{\}$, and $T_1 \cap T_3=\{\}$.
   This requirement is related to the meaning of a magnitude in the FFP similarity values. The similarity between one FFP and another FFP which includes a common follower or the common followee set, is greater than the similarity between one FFP and another FFP with no common social entities.

3.2 FFP similarity

We propose the following FFP similarity $\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j)$ to meet the above requirements. Our proposed definition of $\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j)$ is $\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j) = \text{sim-c}(f_i, f_j)$, where (i) sim-c(fi, fj) = $\max\{s(s_i, s_j), s(t_i, t_j)\}$, (ii) $s(e_i, e_j) = |e_i \cap e_j|/\min\{|e_i|, |e_j|\}$, and (iii) f_i and f_j are FFPs of $s_i \rightarrow t_i$ and $s_j \rightarrow t_j$, respectively.

We shall next explain this definition in terms of the requirements. First, the requirements from RQ1 to RQ3 are obviously met from the definition of $\mathrm{SIM}_{\mathrm{FFP}}$. Second, because $s(s_i, s_i)$=1.0 and $s(t_i, t_i)$=1.0, $\mathrm{SIM}_{\mathrm{FFP}}$ satisfies both RQ4-1 and RQ5-1. $\mathrm{SIM}_{\mathrm{FFP}}$ also satisfies both RQ4-2 and RQ5-2. For the last requirements RQ6-1 and RQ6-2, $\mathrm{SIM}_{\mathrm{FFP}}(f_1, f_3)$ always becomes 0.0 because $S_1 \cap S_2=\{\}$ and $T_1 \cap T_3=\{\}$ and $\mathrm{SIM}_{\mathrm{FFP}}(f_1, f_2)$ is always 1.0. Although $\mathrm{SIM}_{\mathrm{FFP}}(f_1, f_2)$ and $\mathrm{SIM}_{\mathrm{FFP}}(f_1, f_3)$ satisfies RQ6-1 and RQ-2 in the same way, the magnitude in these FFP similarity values may be meaningless. To make these last two requirements meaningful, the following $\mathrm{SIM}_{\mathrm{FFP}}$(fi, fj) extension called EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(fi, fj) is presented. EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j) = \max\ \{\text{sim-c}_{\,fi'\in F+(fi),\, fj'\in F+(fj)}\ (f_i', f_j')\}$, where (i) $\text{sim-c}(f_i, f_j) = \max\{s(s_i, s_j), s(t_i, t_j)\}$ and (ii) $s(e_i, e_j)=|e_i \cap e_j|/\min\{|e_i|, |e_j|\}$.

The difference between $\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j)$ and EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j)$ is whether the closure is used or not in calculating the similarity value. With this extension of the $\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j)$, the RQ6-1 and RQ6-2 are satisfied and gets meaningful, because $\mathrm{SIM}_{\mathrm{FFP}}(f_1$, f3) is no more equal to 0.0 in the case when there exist $f_i'\in F^+(f_i)$ and $f_j'\in F^+(f_j)$ such that $|s'_i \cap s'_j|\neq\{\}$ or $|t'_i \cap t'_j|\neq\{\}$, where f_i': $s'_i \rightarrow t'_i$ and f_j': $s'_j \rightarrow t'_j$.

We show here EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_1, f_2) \geq$ EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_1, f_3)$. EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_1, f_2)$ is always equal to 1.0. From the EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j)$ definition, there are a pair of $f_1'\in F^+(f_1)$, and $f_3'\in F^+(f_3)$, such that sim-c(f_1', f_3') has the maximum value. Although f_1: $\{S_1\} \rightarrow \{T_1\} \in F$, f_3: $\{S_2\} \rightarrow \{T_3\} \in F$, $S_1 \cap S_2=\{\}$, and $T_1 \cap T_3=\{\}$, there is a possibility when $S_1' \cap S_2' \neq \{\}$, for f_1': $\{S_1'\} \rightarrow \{T_1'\}$, and f_3': $\{S_2'\} \rightarrow \{T_3'\}\in F^+(f_3)$, where $T_1' \cap T_3'=\{\}$. In this case, $1.0 \geq$ EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_1, f_3) > 0.0$.

It is necessary to extend this definition of FFP similarity, EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j)$, from a standpoint where EXT-$\mathrm{SIM}_{\mathrm{FFP}}$ (f_i, f_j) is always equal to 1.0 when f_1: $\{S_1\} \rightarrow \{T_1\} \in F$, f_2: $\{S_1S_2\} \rightarrow \{T_2\} \in F$, $S_1 \cap S_2=\{\}$, $T_1 \cap T_2=\{\}$. There are several ways to extend it to solve this situation. For example, the arithmetic mean can be used, instead of the max-operation in sim-c(f_i, f_j). However, we replace the max-operation as the geometric mean because EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j)$ should be 0.0 when $s(s'_i, s'_j)$=1.0 and $s(t'_i, t'_j)$=0.0 for any f_i': $s'_i \rightarrow t'_i$ and f_j': $s'_j \rightarrow t'_j$. Therefore, the final definition of EXT-$\mathrm{SIM}_{\mathrm{FFP}}$ (f_i, f_j) is EXT-$\mathrm{SIM}_{\mathrm{FFP}}(f_i, f_j) = \max\{\text{sim-c}_{\,fi'\in F+(fi),\, fj'\in F+(fj)}\ (f_i', f_j')\}$, where (i) $\text{sim-c}(f_i, f_j) = s(s_i, s_j)\times s(t_i, t_j)$ and (ii) $s(e_i, e_j)=|e_i \cap e_j|/\min\{|e_i|, |e_j|\}$.

3.3 FFP example

We calculate the FFP similarity values using the example shown in Fig.1, where the five FFPs are extracted. To calculate EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP1, FFP2), F^+(FFP1) and F^+(FFP2) need to be obtained beforehand. As explained before, F^+(FFP1) = F^+({EF} → {ABC}) = {{EF} → {ABC}, {EF} → {AB}, {EF} → {BC}, {EF} → {AC}}. In the same way as F^+(FFP1), F^+(FFP2) can be easily calculated as F^+(FFP2)= F^+({BF} → {ACE}) = {{BF} → {ACE}, {BF} → {AC}, {BF} → {AE}, {BF} → {CE}}. We then calculate sim-c values for all the combinations of FFP pairs in F^+(FFP1) and F^+(FFP2). The maximum value 0.5 is obtained in the case of sim-c({EF}→{AC}, {BF}→{ACE}) or sim-c({EF}→{AC}, {BF}→{AC}). Therefore, we obtained EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP1, FFP2)=0.5. Because the rest of EXT-$\mathrm{SIM}_{\mathrm{FFP}}$ for other pairs can be calculated in the same way, the results are the following.

EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP1, FFP3)=0.5	EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP1, FFP4)=0.25	EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP1, FFP5)=1.0
EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP2, FFP3)=0.5	EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP2, FFP4)=1.0	EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP2, FFP5)=1.0
EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP3, FFP4)=0.25	EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP3, FFP5)=0.25	EXT-$\mathrm{SIM}_{\mathrm{FFP}}$(FFP4, FFP5)=0.5

4 FFP-based Entity (FbE) similarity

4.1 Requirements

Before defining FbE, we introduce some requirements for FbE similarity. The FbE similarity should be defined to meet the following requirements. $\mathrm{SIM}_{\mathrm{FbE}}(e_i,e_j)$ is a similarity between two social entities $e_i\in E$ and $e_j\in E$, where E is the set of all social entities.

7. (RQ7) $0.0 \leq SIM_{FbE}(e_i, e_j) \leq 1.0$
   The value of FbE similarity is between 0.0 and 1.0 for simplicity.
8. (RQ8) $SIM_{FbE}(e_i,e_i)=1.0$
   The self-similarity value is always 1.0.
9. (RQ9) $SIM_{FbE}(e_i, e_j) = SIM_{FbE}(e_j, e_i)$
   The similarity value is the same even if the order of two social entities is described in reverse.
10. (RQ10) $SIM_{FbE}(e_1, e_2) = 1.0$ when e_1 and e_2 have the same common followees or have the same common followers.
    In the case when, if e_1 follows some followees, then e_2 also follows the same followees, and vice versa; it is assumed that e_1 and e_2 are the most similar.
11. (RQ11) $SIM_{FbE}(e_1, e_2) \geq SIM_{FbE}(e_1, e_3)$ when e_1 and e_2 share a common follower or followee, and e_1 and e_3 have neither common followees nor common followers.
    In this case, it is meant that e_1 is more similar to a social entity e_2, who has at least one common follower or followee, than to a social entity e_3 who does not have any common followers and followees.
12. (RQ12) $SIM_{FbE}(e_1, e_2) \geq SIM_{FbE}(e_1, e_3)$ when both e_1 and e_2, and e_1 and e_3, share no common followers or followees, but there exists at least one common social entity who is overlapped with a set of followers or followees of e_1 and those of e_2. In this requirement, it is meant that e_1 is more similar to a social entity e_2, who has at least one common social entity whose followers or followees are overlapped with followers or followees of e_1 and e_2, than another completely non-related social entity e_3. For example, when {AB}→{EF} and {BC}→{G}, there are no direct relationships between A and C. However, as B is the common followers who follow E and F as A, the similarity between A and C is more similar than A and another social entity besides B, C, E, F, and G.

4.2 FbE similarity

We propose the FbE similarity $SIM_{FbE}(e_i, e_j)$ to meet the above requirements. The following related similarities, $SIM_{FbE1}(e_i, e_j)$ and $SIM_{FbE2}(e_i, e_j)$, are defined. Our proposed definition of $SIM_{FbE}(ei, ej)$ is the maximum value of these two similarity values.

- $SIM_{FbE1}(e_i, e_j) = Avg(EXT\text{-}SIM_{FFP} (f_i, f_j))$ for all pairs of $f_i \in F$ and $f_j \in F$, such that $e_i \in s_i$, $e_j \in s_j$, f_i: $s_i \rightarrow t_i$, f_j: $s_j \rightarrow t_j$
- $SIM_{FbE2}(e_i, e_j) = Avg(EXT\text{-}SIM_{FFP} (f_i, f_j))$ for all pairs of $f_i \in F$ and $f_j \in F$, such that $e_i \in t_i$, $e_j \in t_j$, f_i: $s_i \rightarrow t_i$, f_j: $s_j \rightarrow t_j$
- $SIM_{FbE}(e_i, e_j) = \max\{SIM_{FbE1}(e_i, e_j), SIM_{FbE2}(e_i, e_j)\}$.

$SIM_{FbE1}(e_i, e_j)$ and $SIM_{FbE1}(e_i, e_j)$ are FbE similarities from the follower and followee viewpoints, respectively. Therefore, $SIM_{FbE1}(e_i, e_j)$ defines how much similarity exists between two entities, e_i, e_j, as followers, whereas $SIM_{FbE2}(e_i, e_j)$ shows how similar the two entities are as followees.

In the above definitions, we employ the average calculation of the $EXT\text{-}SIM_{FFP}$ values for all the FFP pairs including either of two given entities. The arithmetic mean function is selected as the average function because of its simplicity and conventional usage for the average calculation. Any other aggregate function, such as max or geometric mean, may be possible.

This SIM_{FbE} definition is more described in terms of the requirements. First, the requirements from RQ7 to RQ9 are obviously met from the definition of SIM_{FbE}. Second, because e_1 and e_2 have the same common followees or have the same common followers, either $\{e_1e_2\} \rightarrow c$ or $c \rightarrow \{e_1e_2\}$ holds. This FFP is denoted as f_1. $SIM_{FbE}(e_1,e_2)=1.0$ because $EXT\text{-}SIM_{FFP}(f_1,f_1)=1.0$. Therefore, SIM_{FbE} satisfies RQ10. In the same way as RQ10, SIMFbE also satisfies RQ11. For the last requirements RQ12, because there exists a common social entity c whose followees are overlapped with a set of common followees of e_1 and e_2, this is a situation when $\{e_1e_k\} \rightarrow \{de\}$ and $\{e_2\} \rightarrow \{d\}$ hold, where e_k is another entity or entity set. Therefore, $SIM_{FbE}(e_1,e_2)$ is never equal to 0.0 becomes $EXT\text{-}SIM_{FEP}(f_1,$

f_2)=1.0, where f_1: e_1→d and f_2: e_2→d. The follower-overlapped case is the same as the above followee-overlapped case.

Table 1. FbE (FbE1:FbE2) similarity for the example data.

	A	B	C	D	E	F
A	1.0	-	-	-	-	-
B	0.53(0.0:0.53)	1.0	-	-	-	-
C	0.70(0.0:0.70)	0.83(0.83:0.66)	1.0	-	-	-
D	0.0(0.0:0.0)	0.33(0.33:0.0)	0.25(0.25:0.0)	1.0	-	-
E	0.69(0.0:0.69)	0.71(0.71:0.50)	0.60(0.28:0.60)	0.38(0.38:0.0)	1.0	-
F	0.0(0.0:0.0)	0.72(0.72:0.0)	0.60(0.60:0.0)	0.50(0.50:0.0)	0.63(0.63:0.0)	1.0

4.3 FbE example

We calculate the FbE similarity values using the example shown in Fig.1. There are six entities, A to F, in Fig.1. We calculate the values for all the pairs of the six entities. The result is shown in Table 1.

For example, in the previous example, F= {{EF}→{ABC}, {BF}→{ACE}, {DF}→{BCE}, {BCF}→{AE}, {BEF}→{AC}} is obtained. A does not appear in an entity set on the left side of any FFP. Therefore, SIM_{FbE1}(A, B)=0.0. However, A appears in an entity set on the right side of four FFPs, FFP1, FFP2, FFP4, and FFP5. B also appears in an entity set of the right side of two FFPs, FFP1 and FFP3. All the combinations of four FFPs and two FFPs are generated. The generated combinations are (FFP1, FFP1), (FFP1, FFP3), (FFP2, FFP1), (FFP2, FFP3), (FFP4, FFP1), (FFP4, FFP3), (FFP5, FFP1), and (FFP5, FFP3). For each combination, the pre-calculated EXT-SIM_{FFP} (f_i, f_j) values are used to obtain SIM_{FbE2}(A, B). The result of the average is 0.53.

The result of calculating $SIM_{FbE}(e_i, e_j)$, $SIM_{FbE1}(e_i, e_j)$, and $SIM_{FbE2}(e_i, e_j)$ for all pairs of entities e_i, $e_j \in$ {A,B,C,D,E,F} is shown in Table 1. It is observed from Table 1 that, from the followee viewpoint, C and E are the two entities most similar to A. From the follower viewpoint, the most similar entity to B is C, and so on.

5 Discussions

5.1 Preliminary comparative study

We will show the superiority of our similarity definition to other possible similarity definitions. Three methods can be applied to define the FFP similarity: are Jaccard coefficient, the Maximum Common Subgraph (MCS) [14], and he Graph Edit-distance [15].

The Jaccard coefficient is frequently used to compare two item sets. The Jaccard coefficient J between two item sets, A and B, is defined as J=|A∩B|/|A∪B|. In order to apply this coefficient to define the FFP similarity, we need to use the coefficient both for the follower and followee cases. Therefore, given two FFPs, f_1: s_1→t_1 and f_2: s_2→t_2, $J_1(f_1,f_2)$ and $J_2(f_1,f_2)$ can be defined as $J_1(f_1,f_2)$ =$|s_1 \cap s_2|/|s_1 \cup s_2|$ and =$|t_1 \cap t_2|/|t_1 \cup t_2|$, respectively. The similarity $J(f_1,f_2)$ can then be calculated using an appropriate aggregate function, such as max or mean, for J_1 and J_2.

The MCS is a method to find the maximum subgraph among common subgraphs between two given graphs [14]. A similarity function SM based on this MCS is defined as $SM(g_1,g_2)=|MCS(g_1,g_2)|/(|g_1|+|g_2|-|MCS(g_1,g_2)|)$, for given two graphs g_1 and g_2. Because FFP can be decomposed into a set of entities and of direct-edges between two entities, each FFP is considered to be a graph. $SM(f_i,f_j)$ can then be calculated using the MCS.

The Graph edit-distance is an extension of the conventional edit distance, which is obtained by minimizing the total of the operation costs that are necessary to transform one graph to another. Many methods of calculating the distance have been proposed. Operation cost determination especially plays an important role in calculating the distance. In this paper, we describe the simplest definition of the graph edit distance, which is the sum of the number of operations, including node/edge insertion and node/edge deletion. We assume that these operation costs are the same. The similarity value is calculated simply by taking the reciprocal of the edit distance value plus one.

In Table 2, we show the results of requirement satisfaction assessment of FFP similarity for EXT-SIM$_{FFP,}$ SIM$_{FFP}$, Jaccard, MCS, and graph edit-distance. As shown in this table, no methods other than EXT-SIM$_{FFP}$ meet all the FFP similarity requirements.

As for the entity similarity definition, three methods apart from SIM$_{FbE}$ are possible to use for defining the entity similarity: use of the binary vector model [21]; use of the structural similarity [25]; SimRank [26]; and use of Linked Data Semantic Distance (LDSD) [16].

Table 2. Requirement satisfaction assesment of EXT-SIM$_{FFP}$ and other methods.

	EXT-SIM$_{FFP}$	SIM$_{FFP}$	Jaccard	MCS	Edit-dist.
RQ1	Yes	Yes	Yes	Yes	Yes
RQ2	Yes	Yes	Yes	Yes	Yes
RQ3	Yes	Yes	Yes	Yes	Yes
RQ4	Yes	Yes	Yes	No	No
RQ5	Yes	Yes	Yes	No	No
RQ6	Yes	Yes, but always 1.0>0.0	Same as SIM$_{FFP}$	No	No

In the binary vector model, each entity is represented as a binary vector whose entry is set to one if the entity follows another entity related to this entry; otherwise, the entry is set to zero otherwise. The cosine similarity is then used to calculate the similarity.

In the structural similarity, the similarity, $\sigma(a,b)$, between two entity nodes, a and b, is defined as $\sigma(a,b)=|\Gamma(a)\cap\Gamma(b)|/\mathrm{sqrt}(|\Gamma(a)|\ |\Gamma(b)|)$, where $\Gamma(a)$ is a neighborhood of a node [21].

SimRank is a well-known method used to calculate the similarity between two nodes of a graph from the neighboring viewpoint [26].

In the LDSD, the semantic distance between two nodes can be easily calculated even when there exists a huge number of nodes such as in Linked Open Data. The LDSD distance definition is $LDSD(n_i,n_j) = 1/(1+c_d(n_i,n_j)+c_d(n_j,n_i)+c_{ii}(n_i,n_j)+c_{io}(n_i,n_j))$, where $cd(n_i,n_j) = C_d(n_i,n_j) / (1+\log(C_d(n_i,n))$, $c_{ii}(n_i,n_j) = C_{ii}(n_i,n_j)/ (1 + \log(C_{ii}(n_i,n))$, and $c_{io}(n_i,n_j) = C_{io}(n_i,n_j) / (1+\log(C_{io}(n_i,n))$ [16].

Table 3. Requirement satisfaction assessment of SIM$_{FbE}$ and other methods.

	SIM$_{FbE}$	Binary vector	Structural sim.	SimRank	LDSD
RQ7	Yes	Yes	Yes	Yes	Yes
RQ8	Yes	Yes	Yes	Yes	Yes
RQ9	Yes	Yes	Yes	Yes	Yes
RQ10	Yes	Yes	Yes	Yes	No
RQ11	Yes	Yes	Yes	Yes	No
RQ12	Yes	No	No	No	No

In this definition, $C_d(n_i,n_j)$ is the number of direct edges directed from n_i to n_j and $C_d(n_i,n)$ is the number of edges from n_i to any node. $C_{ii}(n_i,n_j)$ is the number of nodes, where each node is linked with at least two edges, one from the node to n_i and one from the node to n_j. $C_{ii}(n_i,n)$ is also the number of nodes, where each node is linked with at least two edges, one from the node to n_i and one from the node to any node. $C_{io}(n_i,n_j)$ is the number of nodes, where each node is linked with at least two nodes, one from n_i to the node and one from n_j to the node. $C_{io}(n_i,n)$ is the number of nodes, where each node is linked with at least two edges, one from the node to n_i and one from any node. The similarity value is obtained only by calculating $c_d(n_i,n_j)+c_d(n_j,n_i)+c_{ii}(n_i,n_j)+c_{io}(n_i,n_j)$.

Table 3 shows the results of our requirement satisfaction assessment on the entity similarity for SIM$_{FbE}$, binary vector model, structural similarity, and LDSD. As shown in this table, the methods, the binary vector model, the structural similarity, and use of LDSD do not meet some of the entity similarity requirements, but SIM$_{FbE}$ meet all of the requirements.

5.2 Social entity recommendation

Although this paper focuses on the similarity of FFP and entities, it is important to describe briefly how the entity similarity can be used for social entity recommendation in social network services. Here, we describe only one possible, simple scheme for entity recommendation, but other methods can be proposed. Moreover, we describe the method only as one method to recommend appropriate social entities using the proposed similarity definitions. Therefore, the method is not considered from a viewpoint of computational efficiency.

Assume that there is one user e who would like to find followees $E_{followee}(e)=\{e_j\}\subset E$ appropriate to the user e. Before finding entities, FFPs are extracted. For each FFP f_k, $F^+(f_k)$ is then generated. Given an entity similarity threshold δ, a set of entities, $E_{follower}(e)$ for the entity e is then extracted, where $E_{follower}(e) = \{e_{k|}\ SIM_{FbE1}(e,e_k) \geq \delta, e_k \neq e, e_k \in E\}$. From the obtained $E_{follower}(e)$, $E_{followee}(e)$ can be extracted as $E_{followee}(e)=\{e_j|\ e_j$ is followed by e_k, e_j is not followed by e, $e_k \in E_{follower}(e)\}$. For example, for the previous example shown in Fig. 1, the FbE (FbE1:FbE2) similarity result obtained for this example is shown in Table 1. It can be seen that the entities similar to B, are (C, E, and F). Although B has already followed the followees, (C and F), E follows the entity D who is not followed by B yet. Therefore, D is recommended to B as a new followee candidate of B.

It is natural to introduce a ranking function when the number of $E_{followee}(e)$ is too large to recommend all followee candidates to the user e. It is beyond the scope of this paper to argue the details of the ranking function.

6 Related work

There are many studies on social mining in social network services, which are mainly categorized into two types: collaborative filtering [17,18,22], and topic detection [23,24].

Collaborative filtering is the information filtering method that can be extracted using rules obtained from a set of users preferences. The method can be easily applied for micro-blogging services, such as Twitter. Hannon et al. proposed a recommendation of Twitter users using this collaborative filtering and developed a recommender system called the Twittomender [17]. The system gathered a set of tweets and generated user profile databases from user‘s tweets, their followers, and their followees, stored as a vector space model [21]. They employed a TF-IDF to calculate the distance between two users or a user and a message. The (direct) following relationship is only represented by a binary vector, followed by combination with another vector constructed from various tweet messages. An indirect following relationship is not considered. In addition, it may be unclear whether TF-IDF can be applied well for short and noisy messages. Chen et al. proposed a collaborative personalized tweet recommendation method, which enables tweet recommendation to a Twitter user by using collaborative filtering [18,22]. They constructed various kinds of features for users and tweets. A set of new tweets were then ranked using their proposed ranking method. They did not intend to recommend useful follower or followee information to existing users, but rather tweets which would be useful to them.

Topic detection from information in social network services is used to detect a new topic upon which users frequently converse. There exist many services providing new detected topics using keyword-based topic detection, such as Trendistic (http://trendistic.com/) and Twopular (http://twopular.com/), as well as some research studies [23,24]. Zhao et al. used a clustering approach with which newly coming messages were clustered [23]. The clustered message graph, called an information flow pattern, is generated and a new topic event is extracted from the changes on the graphs. They proposed the graph similarity based on Dynamic Time Warping (DTW). Cataldi et al. proposed a different approach by using a modification of the famous PageRank algorithm [24]. After calculating the authority value for a user using the following relationships in the social network, a vector per keyword, called nutrition, is obtained with this value in a certain time interval. Then, after a topic graph is constructed, a new topic is detected. Although topic detection in social networks is an important application, these two existing methods for topic detection in social network services are basically combinations of existing techniques for information retrieval and machine learning.

7 Conclusion and future work

In this paper, we proposed a novel method to calculate two similarities, FFP and FbE, used in social network services. The FFP similarity is a similarity between two Frequent Following Patterns (FFPs), whereas FFP-based Entity similarity (FbEs) is a similarity definition that calculates the similarity between two social entities based on FFP. The FFP is a pattern composed of a set of followers and a set of followees who are followed by all the followers [1]. The FbE similarity can be calculated using these FFP values for two FFPs extracted only from Follower-and-Followee (F-F) structures in social networks. The main advantage of the proposed entity similarity is that a similarity value between two social entities can be obtained with no need for user profiles or user message logs. Therefore, even if a social network user is a novice or tends only to read messages, a set of other users similar to the user, determined by using the calculated similarity value can be extracted among a huge number of users. The similarity can be used for developing a method of recommendation of appropriate followees to a novice user. The preliminary comparative study in this paper indicated that the proposed similarities meet all the requirements necessary for the similarity definitions, whereas other possible candidates do not.

Because this paper focuses on describing the above two similarity definitions, the detailed recommendation scheme as well as actual efficient similarity calculation algorithms are not described in this paper. Moreover, as future work, we plan to extend our works in several directions, including detailed evaluation with a real dataset to show their actual effectiveness and efficiency, a recommender system to be developed based on the proposed similarities, and its application to other social mining functions, such as user clustering, social analysis, and trend extraction. We believe that the proposed similarities will become one of the key concepts for social mining and make the mining results more useful and practical by combining them with existing social mining methods that utilize user-profiles as well as user message logs.

Acknowledgements

This work was partially supported by JSPS 24300039 (Japan) and NSERC (Canada).

References

1. Jiang F, Leung CKS. Mining interesting “following” patterns from social networks. In: *DaWaK 2014*, p. 308–319.
2. Cuzzocrea A, Leung CKS, MacKinnon RK. Mining constrained frequent itemsets from distributed uncertain data. *FGCS* 2014; **37**:117–126.
3. Dhahri N, Trabelsi C, Ben Yahia S. A new approach for efficient events mining from social media RSS feeds. In: *DaWaK 2012*, p. 253–264.
4. Jiang F, Leung CKS. Stream mining of frequent patterns from delayed batches of uncertain data. In: *DaWaK 2013*, p. 209–221.
5. Jiang F, Leung CKS, Tanbeer SK. Finding popular friends in social networks. In: *CGC* (*SCA*) *2012*, p. 501–508.
6. Leung CKS, Jiang F. Frequent pattern mining from time-fading streams of uncertain data. In: *DaWaK 2011*, p. 252–264.
7. Leung CKS, Tanbeer SK. Mining popular patterns from transactional databases. In: *DaWaK 2012*, p. 291–302.
8. Leung CKS, Tanbeer SK. Cameron JJ. Interactive discovery of influential friends from social networks. *Soc. Netw. Anal. Min.* 2014; 4(1):art. 154.
9. Lin W, Kong X, Yu PS, Wu Q, Jia Y, Li C. Community detection in incomplete information networks. In: *WWW 2012*, p. 341–350.
10. Ma L, Huang H, He Q, Chiew K, Wu J, Che Y. GMAC: a seed-insensitive approach to local community detection. In: *DaWaK 2013*, p.297-308.
11. Tanbeer SK, Leung CK, Cameron JJ. Interactive mining of strong friends from social networks and its applications in e-commerce. *J. Org. Comp. & E. Com.* 2014; **24**(2-3):157–173.
12. Wei EHC, Koh YS, Dobbie G. Finding maximal overlapping communities. In: *DaWaK 2013*, p. 309–316.
13. Yu W, Coenen F, Zito M, El Salhi S. Minimal vertex unique labelled subgraph mining. In: *DaWaK 2013*, p. 317–326.
14. Bunke H, Shearer K. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters* 1998; **19**(3-4):255-259.
15. Xinbo G, Bing X, Tao D, Li X. A survey of graph edit distance. *Pattern Anal. Applic.* 2010; **13**(1):113–129.
16. Passant A. Measuring semantic distance on linking data and using it for resources recommendations. In: *AAAI SSS 2010*, p. 93-98.
17. Hannon J et al. Recommending twitter users to follow using content and collaborative filtering approaches. In: *ACM RecSys 2010*, p. 199-206.
18. Cheng Z, Caverlee J, Lee K. You are where you tweet: a content-based approach to geo-locating twitter users. In: *ACM CIKM 2010*, p. 759-768.
19. Chen K, Chen T, Zheng G, Jin O, Yao E, Yu Y. Collaborative personalized tweet recommendation. In: *ACM SIGIR 2012*, p. 661-670.
20. You GW, Hwang SW, Nie Z, Wen JR. SocialSearch: enhancing entity search with social network matching. In: *EDBT 2011*, p. 515-519.
21. Ricardo BY, Berthier RN, eds., *Modern information retrieval.* 2nd ed. Addison-Wesley; 2011.
22. Chen J, Nairn R, Nelson L, Bernstein M, Chi E. Short and tweet: experiments on recommending content from information streams. In: *ACM CHI 2010*, p. 1185-1194.
23. Zhao Q, Mitra P, Chen B. Temporal and information flow based event detection from social text streams. In: *AAAI 2007*, p.1501–1506.
24. Cataldi M, Di Caro L, Schifanella C. Emerging topic detection on twitter based on temporal and social terms evaluation. In: *ACM MDMKDD 2010*, art. 4.
25. Xu X, Yuruk N, Feng Z, Schweiger TAJ. SCAN: a structural clustering algorithm for networks. In: *ACM KDD 2007*, p. 824-833.
26. Jeh G, Widom J. SimRank: a measure of structural-context similarity. In: *ACM KDD 2002*, p. 538-543.