MedCrav
Step into the World of Research

International Robotics & Automation Journal

Research Article

Open Access

CrossMark

Offering a new approach for extracting recurring conceptual links from social networks

Abstract

Massive amounts of data in social networks have made researchers look for ways to display a summary of the information provided and extract knowledge from them. Basic methods in the field used the criteria coming from graph theory, but new approaches are trying to take advantage of the realm of traditional data exploration and linked data. One of these new approaches is the approach called conceptual link approach, which was introduced to describe the social networks. In this approach, using the concept of contextual links, knowledge of the social network through a conceptual view of the dramatic structure is summarized as means of social networking. Conceptual perspective provides a summary of existing knowledge on a social network. In order to build this display, it is first needed to extract conceptual links from the intended network. However, extracting these links for networks with larger scale is very time consuming. In this paper, a new method for extracting frequent conceptual link from social networking is provided where by using the concept of dependency, it is tried to accelerate the process of extracting conceptual links. The proposed method will be able to accelerate this process if there are dependencies between data

Keywords: social network analysis, frequent conceptual link, data mining, graph mining, exploring social networks

Volume 4 Issue 1 - 2018

Hamid Tabatabaee

Department of Computer Engineering, Islamic Azad University, Iran

Correspondence: Hamid Tabatabaee, Department of Computer Engineering, Islamic Azad University, Iran, Email h_tabatabaee@mshdiau.ac.ir

Received: November 04, 2017 | **Published:** January 12, 2018

Introduction

Social network is a social structure that is composed of some agents (generally individuals or organizations) that are connected by one or more kind of dependencies, such as ideas and financial transactions, friends, relatives, Web links, spread of diseases (epidemiology). Social networks exist in different categories some of which could be found in.[1] The results of various studies indicate that the capacity of social networks can be used in many individual and social levels in order to identify problems and determine solutions, establishing social relationships, organizational governance, policy making and advising people on track to achieve the objectives. Social network analysis is a powerful diagnostic tool for analyzing the nature and pattern of communication among members of a particular group. Social network analysis helps imagine and analyse complex set of relationships between relevant factors as the maps (graphs or photographs) of connected symbols, and patterns within these categories, and it also helps calculate and review the exact size, shape and density of the network as a whole and calculate the position of each element within it. For example, in the science of epidemiology, social network analysis is used to help understand how patterns of human contact helps or prevents the spread of diseases such as HIV in a population. In addition to social network analysis is a useful tool for surveillance in high volume – for example, Total Information Awareness program has done detailed research on strategies to analyze social networks to determine whether citizens are political threats or not.

From a variety of social networks, online social networking has received attention among researchers. A key aspect of many online social networks is being data–rich, and therefore providing unprecedented challenges and opportunities in terms of knowledge discovery and data mining. One of the most important fields of study of traditional data mining is exploring the frequent pattern. In the field of complex data structures such as networks, the issue of exploring frequent items is discussed in form of finding a subset of nodes (sub–graphs) that occur frequently arises in a network known as graph mining. Although primitive methods in this field have been using measures deriving from graph theory,[2] new approaches known as social networks mining or simply link mining try to examine features of node in addition to the network structure to extract a new set of patterns[3–6] described a new approach as conceptual link to describe social networking. Conceptual link provides the knowledge about groups of nodes connected to each other in a dense social network, and through a reduced structure, which is called as conceptual view, leads to a meaning display of social network. However, the problem of extracting maximum frequent conceptual link is like extracting of frequent item sets[7] with NP–hard complexity.[8] In this study, we aim to provide a new approach to accelerate further in the extraction of these links. For this purpose, D–MFCLMin algorithm is presented that by using the concept of dependence between sets of items, and by pruning, the search space reduces the time required to extract frequent conceptual links. The paper will be structured as follows. In the next section, the concept of conceptual links is offered, and then in the third part, suggested methods for the extraction of frequent conceptual links are introduced. In the fourth part, we introduce the proposed method. In part five, test results are presented and finally in section six, conclusions and future works are presented.

Problem description and definitions

In the field of search for frequent conceptual links (FCL), a model is defined as "a set of links between the two groups of nodes, where the nodes in each group share common characteristics." When these patterns are found on the network with enough repetition, they are seen as frequent patterns and called FCL. More formally, assume that G = (V, E) is a network where V is the set of nodes and E is the set of edges with $E \subseteq V \times V$. V is defined as the relation $R(A_1,...,A_N)$

where each A_i is a trait. Thus, every vertex $v \in V$ is defined by the tuple $(a_1, ..., a_N)$ where $\forall k \in [1..N]$, $v[A_k] = a_k$, is the attribute value A_k in v. An item is a logical expression as A = x where A is an attribute and x is a value. Empty items are shown as ∅. A set of items is a combination of items, for example $A_1 = x$ and $A_2 = y$ and $A_3 = z$. A set of items, m, which is a combination of k non–empty item is called a k–item set and shown as $m^k \left(|m^k| = k\right)$. Suppose that m and sm are two sets of items. If $sm \subseteq m$, we say that sm is a subset of items and m is a superset of the items of sm. For example, $sm = xy$ is a subset of items from $m = xyz$. The all sets of t number of items made of V are shown with with I^t. Moreover, UI^t is defined as follows (all set of all sets of t number of items):

$$UI^t = \bigcup_{k=1}^{t} I^k \tag{1}$$

Suppose that G is a directed graph. Thus, for any item set m on UI^N, V_m is shown as a series of nodes in V that is consistent with the pattern m (literally meet their m) and defined as follows:

Set of links on the left m (LE_m): the set of links from E that starts from the nodes and satisfy m.

$$LE_m = \{e \in E; e=(a,b), a \in V_m\}$$

The set of links on the right (RY_m) m: the set of links from E that enter the nodes that satisfy m.

$$RE_m = \{e \in E; e=(a,b), b \in V_m\}$$

Definition 1– Conceptual links: suppose that m_1 and m_2 are two sets of items and V_{m_1} and V_{m2} are respectively a set of nodes in V that satisfy m_1 and m_2. $E_{(m_1,m_2)}$ is the set of links connecting the nodes in V_{m_1} to the nodes in V_{m_2},

$$E_{(m_1,m_2)} = LE_{m_1} \cap RE_{m_2} = \{e \in E; e=(a, b)\ a \in V_{m_1} \text{ and } b \in V_{m_2}\} \tag{2}$$

Definition 2: we call support $E_{(m_1,m_2)}$ a ratio of links in E that belongs to $E_{(m_1,m_2)}$.

$$supp\left(E_{(m_1,m_2)}\right) = \frac{|E_{(m_1,m_2)}|}{|E|} \tag{3}$$

Definition 3: It is said that there is FCL a nd we write m_1, m_2 if support $E_{(m_1,m_2)}$ is greater than the threshold value of at least β,

$$(\left(\sup p\left(E_{(m_1,m_2)}\right)\right) > \beta)).$$

Definition 4: Suppose UI^t is the set of all sets of items of maximum t value in V, the FL^t is defined as FCL extracted from the set of items.

$$FL^t = \bigcup_{m_1 \in UI^t,\ m_2 \in UI^t} \{E_{(m_1,m_2)}; \frac{|E_{(m_1,m_2)}|}{|E|} > \beta\} \tag{4}$$

Feature 1, being frequent: according to definition 3, if link (m_1, m_2) is frequent, the set of UI^t and RE_{m_2} meet the following condition:

$$|LE_{m_1}| > \beta \times |E| æ |RE_{m_2}| > \beta \times |E| \tag{5}$$

Definition 5– The conceptual sub link: suppose that two sets of items sm_1 and sm_2 are respectively sub–items of items m_1, m_2 in UI. Conceptual link (sm_1, sm_2) is called the sub link of (m_1, m_2), similarly (m_1, m_2) is called hyperlink (sm_1, sm_2) and written as

$$(sm_1, sm_2) \subseteq (m_1, m_2).$$

Feature 2, characteristics of the underlying closure: If a conceptual link *l* is frequent its entire sub links are frequent. Thus, if a link is not frequent none of its hyperlinks is frequent.

Definition 6– Maximum FCL: Assume that β has a given support threshold value, we say that the maximum f conceptual link (MFCL), any FCL is so that no hyperlink of å from l that is frequent exists. More formally:

$$\not\exists å \in FL^N \text{ so that } l \subset å \tag{6}$$

Previous work

Popular approaches of exploring social networking have been proposed to extract different forms of knowledge from these networks. Similar to the traditional field of data mining, exploring range of social networking addresses a wide range of tasks such as classification, clustering, search for recurring patterns or link prediction. Per se, these methods can be divided into two groups.[8]

a. Approaches based on predictive modeling that include techniques that analyze current and past facts to make predictive assumptions about future or unknown events.

b. Approaches based on descriptive modeling that cover a set of techniques whose aim is to summarize data by identifying some related features to describe how to organize things and how to make them real.

In this study, the focus is on descriptive approach of the social network. These approaches can be divided into 4 categories.

Link based clustering: (also known as Community Detection in research) That searches a dense groups of nodes and its aim is to analyze network to several linked components (communities) in such a way that nodes in each component have high–density connections, while nodes in different components have the lowest density of

Citation: Tabatabaee H. Offering a new approach for extracting recurring conceptual links from social networks. *Int Rob Auto J.* 2018;4(1):4–12.
DOI: 10.15406/iratj.2018.04.00084

the proposed methods in this category algorithm SLPA,[9] TopGC,[10] SVINET,[11] MCD,[12] CGGC,[13] CONCLUDE,[14] DSE[15] and SPICi[16] can be cited.

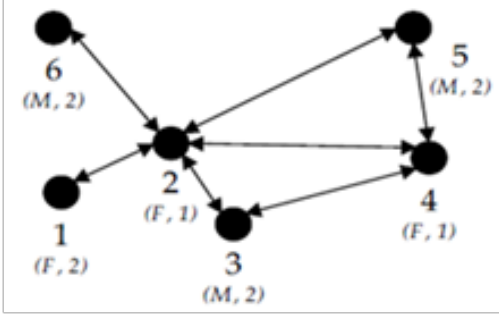
Hybrid clustering: That simultaneously considers attributes and the structure of the nodes to identify clusters. The aim of this new type of approaches is partitioning of the network for balanced between structural similarities and attributes so that nodes with common attributes are grouped in one partition and the nodes inside partition are densely linked. This type of approaches provides a more conceptual partition of the network that is not necessarily proportional to context. Of clustering methods SA–Cluster[4] and CESNA[17] can be cited.

Frequent sub–graph mining: That focuses on extracting infrastructure that occur frequently on the network. The most widely used definition of a pattern is as a connected sub graph.[18] Therefore, techniques that focus on the search for frequent patterns in social networks aim to identify sub graphs that occur frequently in a database or a very large network of networks, based on a minimum threshold value. Among the prominent methods in this category, Apriori–based algorithms[19] and pattern growth[20] can be cited.

FCL: Combines network structure information and features of node for providing knowledge about groups of nodes, which have more connections in a social network. Extracting MFCL creates a complexity similar to frequent item set, since it is proven that this complexity is NP–hard. Extracting all MFCLs from a social network may be a challenging problem and computationally severe. According to the definitions of the concept of conceptual links, we deal with the methods provided for extracting these links.

If search space is very broad, discovering all the frequent links in a network is very costly. In a simple approach, it is necessary to produce all set of possible items and then examine the frequency of each pair of them. To reduce this time, at the beginning, FLMIN algorithm[21] was proposed. This algorithm used a bottom–up approach by applying feature 2 to gradually reduce the search space to include a superset of items that will potentially exist in FCLs.

In Figure 1, the bottom–up performance levels of FLMIN algorithm are shown. Mark* means that the attribute can take any value. At first, the algorithm begins with the search for frequent links that include a set of one–item items (L1 in the picture section B). Then, using feature 2, we know that one superset of items from the item one–item set that are not in the conceptual link will not be in conceptual links (see L2). In general, at t, search space can be limited only to the subset of items that are in conceptual links of stage t–1. In[22] MAX–FL Min algorithm was presented. In this algorithm, the aim is finding MFCLs. Compared with previous algorithm, this algorithm only uses set of items that satisfy feature 1 to create links, and then they are checked for being frequent. In addition, in the process of examining the created link in order to add it to frequent links, this algorithm in addition to being frequent checks lack of existence of maximum frequent link compared to the current link. Moreover, if a frequent link is added to the result list (the list of maximum frequent links) all frequent sub–links will be deleted from the list. In,[5] H–MFCL Min algorithm was presented. In this algorithm, to accelerate the extraction of MFCLs, some of the item sets are filtered. The collection of deleted items includes items that are set by their respective number of nodes in the network less than the threshold α. α is an input parameter for the algorithm. The authors have assumed that FCLs exist between the sets of frequent items. In fact, this filtering is done with the argument that there is little likelihood that a collection of items with low frequency can attract a high proportion of links in the network and therefore by filtering these kinds of item sets, despite the reduction in the search space, certain information will not be lost from the final conceptual network.



L_1	Support
((M , *) , (F , *))	0.357
((M , *) , (* , 1))	0.357
((F , *) , (* , 1))	0.214
((F , *) , (* , 2))	0.428
...	...

L_2	Support
((M , 2) , (F , *))	0.357
((M , 2) , (F , 1))	0.357
((F , 1) , (* , 2))	0.428
((F , 1) , (M , *))	0.357
...	...

Figure 1 A sample of conceptual links extracted by FLMIN algorithm.[21]

The proposed algorithm

In this article, D–MFCL Min algorithm is suggested to extract conceptual links. By pruning the search space by applying the concept of dependency, this algorithm accelerates the extraction of conceptual links. In the following, first we introduce the concept of dependency, and then the proposed algorithms will be introduced and discussed.

Definition 7– Dependency: Suppose m^t and n^t are two sets of items. Say m^t is dependent on n^t and show it as $n^t \nearrow m^t$ if per $v \in Vm^t$, we have $v \in V_n t$. We show all dependencies of a set of items such as m^t in the form of D(m^t).

$$D(m^t) = n^t \mid n^{t'} m^t \tag{7}$$

Definition 8– A set of selected items: Assume that FL^t is the set of extracted FCL from set of items of maximum t–items. $LI^t_{sel}\left(RI_s el^t\right)$ is a set of set of items used to create these links.

Citation: Tabatabaee H. Offering a new approach for extracting recurring conceptual links from social networks. *Int Rob Auto J.* 2018;4(1):4–12.
DOI: 10.15406/iratj.2018.04.00084

$$\begin{aligned} LI_{sel}^{t} &= \left\{ m; E(m,n) \in FL^{t} \right\} \\ RI_{sel}^{t} &= \left\{ m; E(n,m) \in FL^{t} \right\} \end{aligned} \quad (8)$$

Feature 3: If item set n^{t} is not in any of the extracted FCLs in FL^{t} $\left(n^{t} \notin LI_{sel}^{t}\left(RI_{sel}^{t}\right) \right)$, then none of the sets of items that depend on it ($\{m^{t} \mid n^{t} \in D(m^{t})\}$) will be at FL^{t}.

Proving: Assume that m^{t} is one set of the items that depend on the set of items n^{t}, and suppose that n^{t} is not located in any FCL $\left(n^{t} \notin LI_{sel}^{t}\left(RI_{sel}^{t}\right) \right)$, so according to definition of FCL, for all series of items such as n_j $\left| LE_{n^t} \cap RE_{n_j} \right| < \times \left| \left(RE_{n^t} \cap LE_{m_j} < \beta \times E \right) \right.$ is established. Moreover, according to the definition 6 (dependency), we know that $V_{m^t} \subseteq V_{n^t}$, so we have $\left|LE_{m^t}\right| \le \left|LE_{n^t}\right|$ and $\left|RE_{m^t}\right| \le \left|RE_{n^t}\right|$, as a result $\left| LE_{n^t} \cap RE_{n_j} \right| < \times \left| \left(RE_{m^t} \cap LE_{n_j} < \beta \times E \right) \right.$ and therefore the above features is proven.

Definition9–parentsofitemset: For each itemset $(t>1)$ m^{t} two parents are shown as $parent1\left(m^{t}\right)$ and $parent1\left(m^{t}\right)$, $parent2\left(m^{t}\right) \in I^{t-1}$ so that $m^{t} = parent1\left(m^{t}\right).parent2\left(m^{t}\right)$

Definition 10, Dependency Level: For each item sets m, the dependence level is shown with DL (m) and defined as follows:

$$DL(m) = \begin{cases} 0 & if\ D(m) = \varnothing \\ \max\limits_{n \in D(m)} DL(n) + 1 & else \end{cases} \quad (9)$$

The proposed algorithm by developing the algorithm in[5] by applying feature 3 reduces the search space to extract FCL. The pseudo code for this algorithm is given below. Similar to the algorithm H–MFCL Min, input parameters are α and β that are respectively threshold value related to set of items and supporting links. The same as H–MFCL Min algorithm in,[5] in the first step (t = 1), single–item item set $LI_{cand}^{1}\left(RI_{cand}^{1}\right)$ are created according to the first and second features (features relating to the collection of eligible items) (lines 6 and 7). After creating these lists, the set of their items are ordered in terms of the amount of support and ascending order. Unlike H–MFCL Min, before the search for FCLs, in this algorithm in step t, the dependencies between items set in $tLI_{cand}^{t}\left(RI_{cand}^{t}\right)$ are obtained. For this purpose, a collection of t items of $LI_{cand}^{t}\left(RI_{cand}^{t}\right)$ are mutually attached and then, based on the amount of support set of items, the existence of dependency between two attached items is checked. In the absence of dependency, collection of attached items obtained is recorded as one of the candidate items on the list for the next step $LI_{cand}^{t+1}\left(RI_{cand}^{t+1}\right)$ (lines 25–11). This insertion is done in a way that the order of the list of items remains in ascending form in terms of the amount of support. After determining the dependencies among the items sets of step t, their dependence level (Relation 9) is calculated and then in $LI_{cand}^{t}\left(RI_{cand}^{t}\right)$ is ordered by increasing the level of dependence (line 26). After sorting, the search for FCLs is done. Found FCLs are added to FL^{t} list and then by removing sub FCLs links located in FL_{Vmax}, are added to FL_{Vmax} as MFCL (lines 44–27).

More exactly, this search is done so that for every item collection $m_i \in LI_{cand}$ and $m_j \in RI_{cand}$, with the proviso that $|m_i| = t$ or $|m_j| = t$ is checked whether the link (m_i, m_j) is frequent or not. In the proposed algorithm, before this review at this stage, the dependent items m_i and m_j are checked. If none of the sets of dependent items are added in FL^{t}, reviewing the frequency of this pair is ignored (line 33). It is recalled that the set of items in LI_{cand}^{t} and RI_{cand}^{t} are arranged in ascending order of dependency, so when reviewing a set of items, all of the items related to it, has already been investigated at this stage. After this step, similar to H–MFCL Min algorithm, checking the frequency of the link is done (line 34). If the link is frequent, (m_i, m_j), m_i are added to LI_{sel}^{t} and m_j is added to RI_{sel}^{t} after the review of items in LI_{cand}^{t} and RI_{cand}^{t}, set of items LI_{cand}^{t+1} and RI_{cand}^{t+1} are modified to extract FCLs at stage t. At this point, any item sets ($m^{t+1}\left(m^{t+1} \in LI_{cand}^{t+1}\left(RI_{cand}^{t+1}\right)\right)$) whose both sets of parent item (Definition 8) are not in $LI_{sel}^{t}\left(RI_{sel}^{t}\right)$ are removed from the list (49–45). This elimination is carried out by the lower cylinder head features (Feature 2).

Algorithm 1: D–MFCL Min Algorithm

Require: G = (V;E): Network, $\beta \in [0..1]$: Link support threshold and $\alpha \in [0..1]$:Item set filtering threshold

1. FL_{Vmax} :*Set of MFCLs* $\leftarrow \phi$

2. LI_{cand} : Stack of left–hand item set candidates $\leftarrow \phi$

3. RI_{cand} : Stack of right–hand item set candidates $\leftarrow \phi$

4. FL^t : List of frequent conceptual links $\leftarrow \phi$

5. t: Iteration $\leftarrow 1$ {Generation of the 1–itemsets}

6. LI^1_{cand} Generate 1–itemsets m from V such as $|V_m| > \alpha \; and \; |LE_m| > \beta \times |E|$

7. RI^1_{cand} Generate 1–itemsets m from V such as $|V_m| > \alpha \; and \; |LE_m| > \beta \times |E|$

8. Sort LI^1_{cand} , RI^1_{cand} item sets by their Supports

9. $t \leftarrow 1$

10. do {Determining Dependencies between LI^t_{cand} $\left(RI^t_{cand}\right)$ do item sets}

11. for all item set $m^t_i \in LI^t_{cand}\left(RI^t_{cand}\right)$ do

12. for all item set $m^t_i \in LI^t_{cand}\left(RI^t_{cand}\right)$ do

13. if (m^t_i and m^t_j share t–1 item)

14. m^{t+1}_k join m^t_i and m^t_j

15. if (sup(m^{t+1}_k) = sup(m^t_i))

16. add m^t_j to D(m^t_i)

17. else

18. if $|V_{m^{t+1}_k}| > \alpha \; and \; |LE_{m^{t+1}_k}| > \beta \times |E| \left(|RE_{m^{t+1}_k}| > \beta \times |E|\right)$

19. add m^{t+1}_k to $LI^{t+1}_{cand}\left(RI^{t+1}_{cand}\right)$

20. parent1 $\left(m^{t+1}_k\right) \leftarrow m^t_i$

21. parent 2 $\left(m^{t+1}_k\right) \leftarrow m^t_j$

22. end if

23. end if

24. end for

25. end for

26. Sort $LI^t_{cand}\left(RI^t_{cand}\right)$ item sets by their calculated dependency level

{Generation of frequent conceptual links}

27. $FL^t \leftarrow \phi$

28. $LI^t_{sel} \leftarrow \phi$

29. $RI^t_{sel} \leftarrow \phi$

30. for all item set $m_i \in LI_{cand}$ do

31. for all item set $m_i \in RI_{cand}$ do

32. if ($\left(|m_i|=t \text{ or } |m_j|=t\right)$)

33. if $((\exists (m_k, m_j) \in FL^t, \forall m_k \in D(m_i)$ and $(\exists (m_i, m_k) \in FL^t, \forall m_k \in D(m_j))$

34. if ($\left(\exists l \in FL^t \; such \; as \; (m_i, m_j) \subset l \text{ and} |(m_i, m_j)| > \beta \times |E|\right)$)

35. add (m_i, m_j) to FL^t

36. remove all $q \in FL_{Vmax}$ such as $q \subset (m_i, m_j)$

37. add (m_i, m_j) to FL_{Vmax}

38. add m_i to L^t_{sel}

39. add m_j to R^t_{sel}

40. end if

41 end if

42. end if

43. end for

44. end for

45. for all item set $m_i \in LI^{t+1}_{cand}\left(RI^{t+1}_{cand}\right)$ do

46. if $\left(parent1(m_i) \notin L^t_{sel}\left(R^t_{sel}\right) \text{ and } parent2(m_i) \notin L^t_{sel}\left(R^t_{sel}\right)\right)$

47. remove m_i from $LI^{t+1}_{cand}\left(RI^{t+1}_{cand}\right)$

48. end if

49. end for

50. $t \leftarrow t+1$

51. while $FL^t \neq \phi$ and all Combinations() = false

52. return FL_{Vmax}

Citation: Tabatabaee H. Offering a new approach for extracting recurring conceptual links from social networks. *Int Rob Auto J.* 2018;4(1):4–12. DOI: 10.15406/iratj.2018.04.00084

Analysis of the proposed method

First, the cost of H–MFCL Min algorithm is discussed. Suppose that we want check the existence of conceptual link between the two sets of items m_1^i and m_2^j (i = t or j = t) at step t($m_1^i \in LI_{cand}^t$, $m_2^j \in RI_{cand}^t$) To this end, the edges of the network whose source node belong to m_1^i and their destination node belongs to m_2^j will be counted, the cost of this study can be obtained as follows:

$$C\left(m_1^i, m_2^j\right) = 2.N.|E| \tag{10}$$

In the above equation, N is the number of features of each item set. To search for a node belonging to a set item, it is enough to compare attribute values of nodes with the item set that will have cost of N and because this action should be done for source and destination group of each of the edges, double of these costs will be imposed. In D–MFCL Min algorithm, by taking into account the dependencies, the above costs will change as follows:

$$C\left(m_1^i, m_2^j\right) = C_d + (1-p)\left(2.N.|E|\right) \tag{11}$$

In the above relation, C_d is the cost of studying the dependencies of two sets of items m_1^i and m_2^j ، and p is the possibility that dependencies on these two item sets would stop counting the edges of social networks to check for conceptual link between them. C_d value depends on the number of dependencies of the item sets being checked and the number of conceptual links found in the intended stage. In the algorithm D–MFCL Min, for every pair of items being checked, their dependency of participation in the conceptual links that have been found so far in the current phase is evaluated, so this cost is as follows.

$$C_d = \left(\left|D\left(m_1^i\right)\right| + \left|D\left(m_2^j\right)\right|\right)\left|FL^t\right| \tag{12}$$

Therefore, in the following the value of two factors of the dependency factor of set of items and conceptual links are examined.

The number of dependencies of an item set

There is no possibility to determine the exact number of dependencies of a set of items, so we will consider their maximum number. For simplicity, we assume that the number of items in stage t, in LI_{cand}^t are RI_{cand}^t equal. According to this assumption, in continuation, this text assumes no difference between the two sets and therefore to be concise we will use the abbreviations I^t instead of these two sets. As already mentioned, the set of items in each stage are ordered based on the support arranged in ascending. Based on the assumption of the existence of maximum possible dependencies in the set I^t, the first set of items will not be dependent on any item other, the second set of items only may be dependent on the first set of items, the third dependency will maximally be dependent on two previous items, and the same way, so the maximum number of dependencies between all series of items in the set I^t is equal to:

$$\frac{\left|I^t\right|\left(\left|I^t\right|-1\right)}{2} \tag{13}$$

By considering the smooth distribution of this dependency between sets of items of this collection, the maximum number of dependencies for each set of item is obtained as:

$$\left|D\left(m^t\right)\right| = \frac{\left|I^t\right|-1}{2} \tag{14}$$

It should be noted that the maximum number of items set in a stage can be obtained from the following recurrence relation:

$$|I^M| = T(N,M) = \begin{cases} \sum_{i=1}^{N} K_i & M=1 \\ \prod_{i=1}^{N} K_i & N=M \\ \sum_{i=M}^{N} K_i.T(i-1,M-1) & else \end{cases} \tag{15}$$

In the above relation K_i shows the number of possible values for i–th feature. For example, about the characteristics of gender, the number of possible values is equal to 2.

The number of conceptual links found

The second factor affecting the cost of checking dependencies is the number of conceptual links found in a stage $\left(\left|FL^t\right|\right)$. Given the steady growth of the number of conceptual link, the maximum number of conceptual links assessed per pair set of items is equal to:

$$\frac{\left(2\left|\cup I^t\right|\left|I^t\right|-\left|I^t\right|^2\right)^2}{2} \tag{16}$$

According to the above values, the number of conceptual links that are checked for every pair of set items on average is equal to

$$\frac{2\left|\cup I^t\right|\left|I^t\right|-\left|I^t\right|^2}{2} \tag{17}$$

According to relations (14) and (17), the overall amount of C_d is obtained as follows:

$$C_d \frac{2\left|\cup I^t\right|\left|I^t\right|^2-\left|I^t\right|^3}{2} \tag{18}$$

Now, with regard to determining the amount of the dependencies cost, we will analyze the behavior of the proposed algorithm. The worst situation in the proposed algorithm occurs, when despite the large amount of dependencies, there is no pruning. The amount of pruning depends on the number of conceptual links found, as the number of conceptual link found is low, an increase in dependency, will be more likely in pruning the set of items. On the other hand, the number of FCLs depends on the amount of β, as the value of this parameter is less more FCLs will be found. Therefore, we expect that the proposed algorithm when β is a small amount show a weaker performance.

Citation: Tabatabaee H. Offering a new approach for extracting recurring conceptual links from social networks. *Int Rob Auto J.* 2018;4(1):4–12.
DOI: 10.15406/iratj.2018.04.00084

Tests and results

In this section, the results of the assessment of the proposed method (D–MFCL Min) are provided. H–MFCL Min method is considered as the method used for comparison. First, in the sub section, data set used is introduced, and then we will examine the results.

A the dataset used

Most existing dataset in the field of social networking only have network connections information. Since the approach of FCL connections are of the methods that by simultaneous use of connections data and the attribute of nodes traits attempts to extract knowledge, we need a dataset that in addition to the links information has node attributes as well. In addition, since in this method, the attributes of the nodes must be homogeneous, using a dataset that relates to several entities is not possible. Due to these limitations, despite the existence of large number of dataset related to social networks, the range of usable dataset is very limited. In this study, dataset of a social network called Pokec was used.[23] Pokec is the most popular online social network in Slovakia. This dataset includes altered profiles of the users of this social network with links of friendship between them. It should be noted that in social network Pokec friendship relationship are directed. User's profile includes 59 fields that only eight fields are mandatory. In the Table 1 below, the features of these eight fields are shown. * Frequently areas in Slovakia but some areas included in the Czech and German as well. After reviewing and doing the refinement necessary, of the eight fields, five fields were considered as user features that include public or private profile, gender, region, year of registration and age. Age group means the result of dividing the age declared over 10 that 10 different categories will be achieved $\left(\frac{age}{10}\right)$. Members until 2012 (the time of data extraction) are 1,632,803 and the number of connections is equal to 30,622,564. Due to the large number of nodes and connections, and the incompleteness of much of other fields in this study, only the nodes that have value more than 80 percent of profile were considered, the number of which is 31211 nodes and the number of connections between these nodes is equal to 261,945.

A tests and results

As mentioned, in order to evaluate the performance of the proposed method, its results were compared with the results of H–MFCL Min algorithm. It should be noted that the output of both methods is similar in the sense that, there are no differences in the extracted FCL in the two methods. In the following figure, conceptual view extracted from Pokec social networking is shown. β Value for this extraction is selected as 0.25. An interesting feature visible in Figure 2 is the two–way communications between sets of items. In fact, if there are conceptual links between the sets of items A to B there is a conceptual link between B to A set of items. As already mentioned, the mentioned social network is directional, which means that friendship is one–sided. However, with the resulting output, it is revealed that the users of this social network have bilateral friendship relations. By reducing the value of the parameter β, greater number of items set and FCLs are obtained and the conceptual view obtained becomes larger, so we will stick only to this amount.

Although the proposed algorithm (D–MFCL Min) and H–MFCL Min algorithm extract similar conceptual views from the social network, the time taken to do this is in two algorithms is slightly different. In Figure 3, the period of implementation of each of these two algorithms to extract MFCL from Pokec social network is shown at different values of parameter β. It should be noted, parameter α value is considered as equal to zero. Both algorithms have been implemented 10 times and the achieved average execution time is considered as the time of their implementation. As can be seen, at high levels of β of both algorithms, there is almost the same performance but with a lower value of this parameter, the difference in the time of two algorithms becomes greater. This difference reviews the proposed methodology to determine the dependencies between items of the sets. It is noteworthy that, unfortunately, the dependency between items dataset are set to zero, so in fact no pruning is done due to the dependency in this trial. However, if there is dependency between sets of items, the possibility of pruning the search space and thus accelerating the extraction of FCLs will be possible link exists no superset concept will result, and thus the difference in performance of the two algorithms will be a greater increase.

Table 1 Mandatory fields feature in Social Network

Field title	Type of field	Domain	Description
user_id	Integer	The number of users-1	An integer that maps the user name of choice
Public	Boolean	True.. False	Profile's Being Public
Completion percentage	Integer	[1-100]	The value of the fields
Gender	Boolean	True.. False	Sex
Region	Textual	[1-183]	User living area*
last_login	Date time	1999 to 2012	Last logon of the user
Registration	Date time	1999 to 2012	Time of User's Registration to the System
Age	Integer	[1-100]	User age

Figure 2 The conceptual view extracted from Pokec social network (Beta=0.25).

Citation: Tabatabaee H. Offering a new approach for extracting recurring conceptual links from social networks. *Int Rob Auto J.* 2018;4(1):4–12. DOI: 10.15406/iratj.2018.04.00084

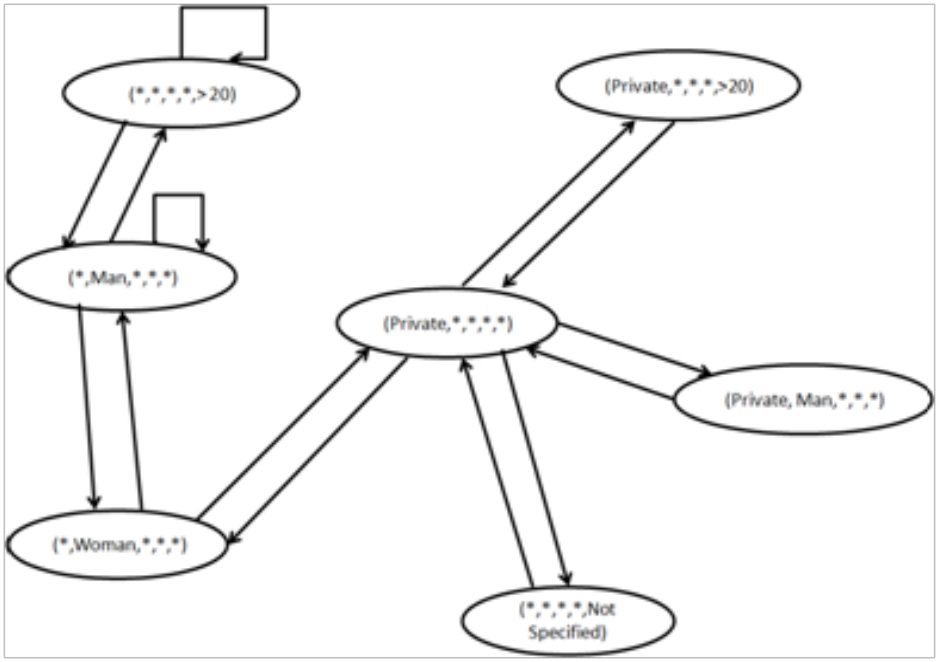


Figure 2 The conceptual view extracted from Pokec social network (Beta=0.25).

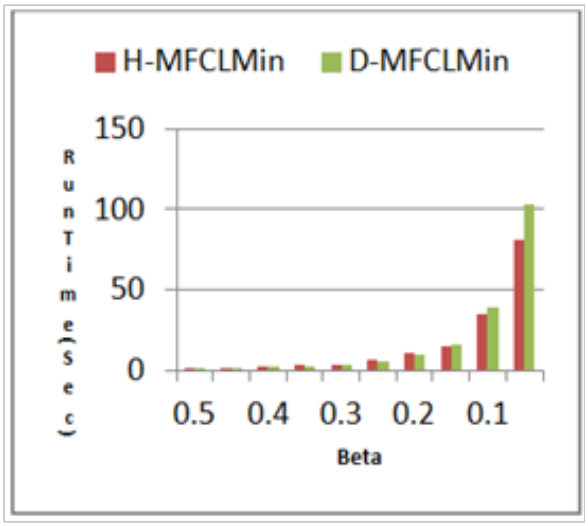


Figure 3 The implementation time of the two algorithms, D-MFCL Min and H-MFCL Min in different amounts of Beta.

Conclusions and future works

Widespread use of social networks has caused very high volume of information and knowledge extraction has become one of the areas of interest for researchers. FCLs are one of the approaches to extract knowledge from these networks that in addition to the data related to communications emphasizes the data related to the existence of these networks. In this paper, by introducing and using the concept of dependence, a new algorithm is presented to accelerate the extraction of FCLs. The existence of dependencies between data causes a pruning of portion of the search space and thus accelerates the process of extracting conceptual links. Due to the lack of dependency in the dataset used, this acceleration was not observed, but the test results showed that despite the lack of dependencies, the proposed algorithm compared with H–MFCL Min algorithm has almost the same performance. In this paper, the concept of dependency was used as definitive, while by extending this concept as approximate dependencies, further pruning of the search space was done that it will be done in future work.

Acknowledgments

None.

Citation: Tabatabaee H. Offering a new approach for extracting recurring conceptual links from social networks. *Int Rob Auto J.* 2018;4(1):4–12.
DOI: 10.15406/iratj.2018.04.00084

Conflicts of interest

Author declares that there is none of the conflicts.

References

1. Aggarwal CC. An introduction to social network data analytics. In: *Social Network Data Analytics*. USA: Springer; 2011. p. 1–15.
2. West DB. *Introduction to graph theory*. 2nd edn. USA: Prentice Hall; 2000.
3. Tian Y, Hankins RA, Patel JM. Efficient aggregation for graph summarization. *Proceedings of the ACM SIGMOD international conference on management of data*. 2008. p. 567–580.
4. Zhou Y, Cheng H, Yu JX. Graph clustering based on structural/attribute similarities. *VLDB Endow*. 2009;2(1):718–729.
5. Stattner E, Collard M. Towards a hybrid algorithm for extracting maximal frequent conceptual links in social networks. In: *IEEE international conference on research challenges in information science*. 2013. p. 1–8.
6. Stattner E, Collard M. Social–based conceptual links: Conceptual analysis applied to social networks. *International Conference on Advances in Social Networks Analysis and Mining*. 2012.
7. Yang G. The complexity of mining maximal frequent itemsets and maximal frequent patterns. In: *KDD 04: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*. USA: ACM Press; 2004. p. 344–353.
8. Stattner E, Collard M. Descriptive Modeling of Social Networks. *Procedia Computer Science*. 2015;52:226–233.
9. Xie J, Szymanski BK. Towards linear time overlapping community detection in social networks. *PAKDD*. 2012;7302:25–36.
10. Macropol K, Singh AK. Scalable discovery of best clusters on large graphs. *PVLDB*. 2010;3(1):693–702.
11. Gopalan PK, Blei DM. Efficient discovery of overlapping communities in massive networks. *Proc Natl Acad Sci U S A*. 2013;110(36):14534–14539.
12. Riedy J, Bader DA, Meyerhenke H. Scalable multithreaded community detection in social networks. In: Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW). *IEEE 26th International*. 2012. p. 1619–1628.
13. Ovelgonne M, Geyer–Schulz A. An ensemble learning strategy for graph clustering. *Graph Partitioning and Graph Clustering*. 2012. p. 187–206.
14. De Meo P, Ferrara E, Fiumara G, et al. Mixing local and global information for community detection in large networks. *J Comput Syst Sci*. 2014;80(1):72–87.
15. Chen J, Saad Y. Dense sub graph extraction with application to community detection. Knowledge and Data Engineering. *IEEE Transactions on*. 2012;24(7):1216–1230.
16. Jiang P, Singh M. Spici: a fast clustering algorithm for large biological networks. *Bioinformatics*. 2010;26(8):1105–1111.
17. Yang J, Mcauley J, Leskovec J. Community Detection in Networks with Node Attributes: In Data Mining (ICDM). *IEEE 13th International Conference*. 2013. p. 1151–1156.
18. Getoor L, Diehl CP. Link mining: a survey. *SIGKDD Explor Newsl*. 2005;7(2):3–12.
19. Agrawal R, Srikant R. *Fast Algorithms for Mining Association Rules in Large Databases*. VLDB Conference. 1994. p. 487–499.
20. Han J, Pei J, Yin Y, et al. Mining frequent patterns without candidate generation: A frequent–pattern tree approach. *Data Mining and Knowledge Discovery*. 2003;8(1):53–87.
21. Stattner E, Collard M. *MAX–FL Min: An Approach for Mining Maximal Frequent Links and Generating Semantical Structures from Social Networks*. 23rd International Conference, DEXA. 2012. p. 468–483.
22. Stattner E, Collard M. *FLMin: An Approach for Mining Frequent Links in Social Networks*. 4th International Conference. 2012. p. 449–463.
23. Takac L, Zabovsky M. Data Analysis in Public Social Networks. *International Scientific Conference & International Workshop Present Day Trends of Innovations*. 2012. p. 1–6.

Citation: Tabatabaee H. Offering a new approach for extracting recurring conceptual links from social networks. *Int Rob Auto J.* 2018;4(1):4–12.
DOI: 10.15406/iratj.2018.04.00084