6С

УДК 316.77:[004.82+004.89]

О.О. Савельев, А.И. Шевченко
Донецкий национальный технический университет, Украина
Украина, 83050, г. Донецк, пр. Богдана Хмельницкого, 84

Определение классов состояний динамической социальной сети по трафикам ее мониторинга

O.O. Saveliev, A.I. Shevchenko
Donetsk National Technical University, Ukraine
Ukraine, 83050, c. Donetsk, Bogdana Khmelnitskogo av.

The Mining the Classes of States from Monitoring Traffic of Dynamic Social Network

О.О. Савельєв, А.І. Шевченко
Донецький національний технічний університет, Україна
Україна, 83050, м. Донецьк, пр. Богдана Хмельницького, 84

Визначення класів станів динамічної соціальної мережі по трафіках її моніторингу

Рассматривается задача определения классов состояний динамической социальной сети. Предложен автоматический подход, основанный на использовании модели динамического графа и метода иерархической агломеративной кластеризации его состояний. Ряд экспериментов на наборе данных MIT Reality Mining показал корректность подхода и достаточное качество решения задачи.
Ключевые слова: динамический граф, иерархическая агломеративная кластеризация.

The article examines the task of the mining the classes of states from monitoring traffic of dynamic social network. We propose the automatic approach based on usage of dynamic graph model and hierarchical agglomerative clustering of its states. A series of experiments on an MIT Reality Mining dataset showed the correctness of the approach and sufficient quality of problem solution.
Key words: dynamic graph, hierarchical agglomerative clustering.

Розглядається задача визначення класів станів динамічної соціальної мережі. Запропонований автоматичний підхід, заснований на використанні моделі динамічного графа і методу ієрархічної агломеративної кластеризації його станів. Низка експериментів на наборі даних MIT Reality Mining показала коректність підходу і достатню якість розв'язку задачі.
Ключові слова: динамічний граф, ієрархічна агломеративна кластеризація.

Введение

Динамические социальные сети относятся к классу естественных природных систем, качественное изучение которых возможно только благодаря мониторингу – наблюдению и протоколированию событий между акторами сети, т.е. записи соответствующих трафиков. Подобная информация описывает множество конкретных мгновенных состояний отдельных компонент сети (диад, триад и т.п.) в отдельные моменты времени, но не позволяет наглядно видеть текущие продолжительные состояния всей сети. С другой стороны данный способ описания наиболее удобен для безмасштабных (scale-free) сетей, когда множества акторов, отношений, а соответственно, и состояний бесконечны и динамически изменяются. Следовательно, наблюдается некий конфликт

между состояниями сети, представленными ее трафиками, и желанием аналитика оценить состояния сети для некоторого ее ограниченного сегмента в некотором ограниченном промежутке времени. Поэтому определение классов состояний динамической социальной сети является ***актуальной*** задачей.

Данной и родственными задачами занимается ряд ***научных школ*** профессоров: A.S. Pentland из MIT [1], [2], J. Kleinberg из Корнуоллского университета [3], D.K.J. Lin из университета Пенсильвании [4], В.В. Геппенера из СПбГЭТУ [5] и другие исследователи [6], [7].

В работе [8] был предложен подход к решению ***задачи прогнозирования временных связей*** [3], [4], основанный на методах машинного обучения. Подход состоит из трех стадий: построение динамического графа, заполнение базы знаний, логический вывод. Вопросы предварительной подготовки исходных данных и стадия построения динамического графа подробно рассмотрена работах [9-11]. Стадия заполнения базы знаний уже в свою очередь состоит из трех стадий: определение классов состояний, определение паттернов следования классов состояний, генерация правил динамики сети.

Объектом исследования данной работы является процесс определения классов состояний динамической социальной сети. Данный процесс рассматривается как профилирование событий и временных сегментов. Поэтому ***предметом исследования*** являются модели и методы профилирования состояний динамической социальной сети.

Цель данной работы – увеличение степени автоматизации и улучшение качества получаемого решения в процессе определения классов состояний динамической социальной сети по трафикам ее мониторинга. Для достижения цели необходимо выполнить следующие ***задачи***: провести обзор и анализ существующих работ, выполнить постановку задачи исследования и выбрать подход к ее решению; выбрать конкретный метод, формализовать и расширить его на модель динамической социальной сети; спланировать и провести вычислительный эксперимент; проанализировать результаты.

Модель динамической социальной сети. Социальная сеть – структура, образованная множеством акторов, и отношений между ними. Динамическая социальная сеть – такая социальная сеть, у которой множества акторов и отношений есть функции времени. Для моделирования такой сети вводится понятие динамического графа. Динамический граф (ДГ) представим как множество

$$DG = \{G_{t_1}, \dots, G_{t_i}, \dots, G_{t_n}\}, \quad (1)$$

где n – количество разбиений времени существования соцсети на равные промежутки времени, длиной $t_i - t_{i-1}$; $t_1,..,t_i,..,t_n$ – конечные значения промежутков времени $t_i - t_{i-1}$; $G_{t_1},..,G_{t_i},..,G_{t_n}$ – статические графы, представляющие состояния сети в моменты времени $t_1,...,t_i,...,t_n$. Величину $\Delta t = t_i - t_{i-1}$ назовем периодом квантования трафиков соцсети по времени.

Каждый статический граф определяется как

$$G_{t_i} = \left(V, E^{G_{t_i}}\right), \quad (2)$$

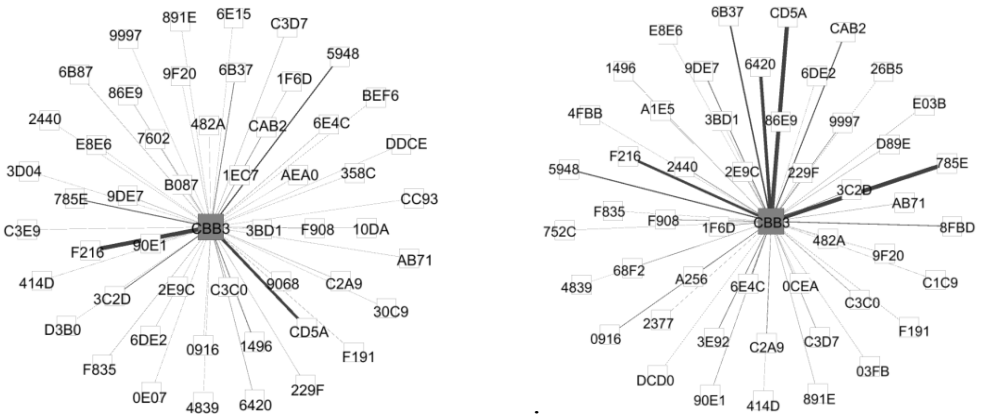
где V – постоянное множество вершин – акторы v_i, общее для всех G_{t_i}: $V = \{v_1,..., v_i,... v_{n^V}\}$ $E^{G_{t_i}}$ – множество ребер – связи между акторами.

Каждое ребро e определяется как

$$e = \{v_i, v_j, w\}, \quad (3)$$

где v_i, v_j – вершины, образующие ребро; w – вес, который определяется по количеству событий коммуникации между v_i, v_j во время $t_i - t_{i-1}$.

Алгоритм построения ДГ из трафиков [10] позволяет получить модель как для ограниченного замкнутого сегмента сети из нескольких акторов, так и для ограниченного, но открытого сегмента из одного актора. Так как, в реальном мире социальная сеть свободно растущая, то будем рассматривать частный случай ДГ как модели социального графа одного актора, т.е. эгоцентрической социальной сети, где анализируемый актор – это центральная вершина, а остальные акторы представляются висячими, либо изолированными вершинами. Примеры визуализации такой модели, построенной на основе данных [1], приведены на рис. 1, висячие вершины не показаны.



а) октябрь 2004 г. б) ноябрь 2004 г

Рисунок 1 – Визуализация динамического социального графа для выборок

Обзор и анализ литературы. Среди исследований, посвященных определению классов состояний социальных сетей, рассмотрим следующие. Работа [2] посвящена моделированию поведения акторов по динамическим геолокационным данным. В качестве состояния отдельного актора вводится понятие *eigenvector* – вектор значений признаков в многомерном пространстве. Пространства признаков представлены временем и географическими координатами. Ввиду большого количества реальных экспериментальных данных классы определяются с помощью методов кластеризации. В работе [5] для определения моментов перехода объекта из одного состояния в другое использовались методы сегментации сигналов. Полученные сегменты подвергались кластеризации для определения всего пространства актуальных состояний. В работе [6] осуществляется майнинг паттернов перемещения через определение траекторий и кластеризацию их состояний. В работе [7] осуществляется построение пространственно-временных профилей пользователей. Решается сопутствующая задача осцилляции базовых станций через их топологическую кластеризацию. Резюмируя выполненный обзор, можно сделать вывод, что во всех рассмотренных работах использовались методы кластеризации, поэтому выберем их в качестве подхода для определения классов состояний динамической социальной сети.

Постановка задачи. Задачу определения классов состояний динамической социальной сети сформулируем как формирование множества классов (кластеров)

$$C = \{C_1, \ldots, C_x, \ldots, C_{n^C}\}\ , \tag{4}$$

к каждому из которых отнесено множество близких состояний из ДГ (1)

$$C_x = \{G_1^{C_x}, \ldots, G_{n^{C_x}}^{C_x}\}, \tag{5}$$

и для которых определены эталоны

$$CE = \{G_1, \ldots, G_x, \ldots, G_{n^C}\}, \forall C_x : \exists! G_x \in CE . \quad (6)$$

Материалы и методы

Кластеризация состояний динамического графа. Исходными данными для кластеризации является ДГ (1). Выходными данными являются классы (кластеры) (4), эталоны (центры) (6). Так как заранее количество кластеров не известно, то от метода кластеризации требуется автоматическое определение данного параметра. Среди множества методов кластеризации [12] этому требованию удовлетворяют несколько методов, однако выберем метод иерархической агломеративной кластеризации, поскольку требуемый критерий сечения дендрограммы хорошо формализуется для предметной области, а известные правила разделения/слияния кластеров хорошо исследованы.

Вводится функция меры расстояния, позволяющая оценивать расстояние между отдельными состояниями. Используем евклидово расстояние

$$DM_E(G_x, G_y) = \sqrt{\sum_{i=1}^{|DMS|} \left(DM_i(G_x, G_y)\right)^2} , \quad (7)$$

где *DMS* – множество функций нормированных расстояний по признакам. Чтобы используемый подход оставался достаточно универсальным, будем использовать всего два признака:

1) наличие ребра между одинаковыми парами вершин

$$DM_{ep}(G_x, G_y) = \sqrt{\sum_{i=1}^{|EU|} \left(\begin{cases} 1, e_i \in E^{G_x} \\ 0, иначе \end{cases} - \begin{cases} 1, e_i \in E^{G_y} \\ 0, иначе \end{cases} \right)^2}, EU = E^{G_x} \bigcup E^{G_y} ; \quad (8)$$

2) вес соответствующих ребер

$$DM_{ew}(G_x, G_y) = \sqrt{\sum_{i=1}^{|EU|} \left(\begin{cases} w^{e_i}, e_i \in E^{G_x} \\ 0, иначе \end{cases} - \begin{cases} w^{e_i}, e_i \in E^{G_y} \\ 0, иначе \end{cases} \right)^2}, \; EU = E^{G_x} \bigcup E^{G_y} . \quad (9)$$

Выбор последнего признака обусловлен самым очевидным параметром графа, а выбор первого признака должен препятствовать попаданию в один кластер графов с разными ребрами, но близкими весами.

В качестве правила слияния кластеров будем использовать average linkage

$$LR_A(C_x, C_y) = \frac{\sum_{i=1}^{|C_x|} \sum_{j=1}^{|C_y|} DM_E(G_i, G_j)}{|C_x||C_y|} . \quad (10)$$

Для реализации критерия сечения дендрограммы (останова) возможно использование различных требований: например, для каждого кластера должен существовать связный граф, состоящей только из ребер графов этого кластера (11). Алгоритм кластеризации представим в виде (12).

$$CC_{GLE}(C_x) = \begin{cases} true, \exists G: \\ G = \{V^G, E^G\}: \\ \begin{cases} V^G = V \\ E^G = \left| \begin{array}{l} E \leftarrow E^{G_1^{C_x}} \\ C \leftarrow C_x \setminus G_1^{C_x} \\ пока\ |C| \neq 0 \\ \left| \begin{array}{l} E \leftarrow \begin{cases} E \cup E^{G^C}, \\ если\ E \cap E^{G^C} \neq \varnothing \\ \varnothing, иначе \end{cases} \\ C \leftarrow \begin{cases} \varnothing, E = \varnothing \\ C \setminus G^C, иначе \end{cases} \end{array} \right. \\ E \end{array} \right. \\ E^G \neq \varnothing \\ false, иначе \end{cases} \end{cases} \quad (11)$$

$$C_{HA}(DG) = \left| \begin{array}{l} C \leftarrow \varnothing \\ для\ i = 1 до |\ DG\ | \\ \left| C \leftarrow C \cup \{DG_i\} \right. \\ пока\ CC_{GLE}(\forall C_x \in C) = true \\ \qquad и |\ C\ | \neq 1 \\ \left| \begin{array}{l} \{C_x, C_y\} : C_x \in C, C_y \in C, \\ LR_A(C_x, C_y) \to \min \\ C_n \leftarrow C_x \cup C_y \\ C \leftarrow C \setminus C_x \\ C \leftarrow C \setminus C_y \\ C \leftarrow C \cup C_n \end{array} \right. \\ если |\ C\ | \neq 1 \\ \left| \begin{array}{l} C \leftarrow C \setminus C_n \\ C \leftarrow C \cup C_x \\ C \leftarrow C \cup C_y \end{array} \right. \\ C \end{array} \right. \quad (12)$$

Определим операцию отображения классов (4) на эталоны (6) как функцию вычисления центров кластеров. Центр можно вычислить различными способами, будем использовать объединение графов кластера и усреднение весов ребер.

Оценка качества. Для оценки качества результатов кластеризации [12] будем использовать формальные относительные критерии компактности и отделимости, а также кросс-проверку как неформальный критерий.

Критерий компактности определяет среднее внутрикластерное расстояние

$$CQAC_D(C, CE) = \sum_{i=1}^{|C|} \left(\frac{1}{|\ C_i\ |} \sum_{j=1}^{|C_i|} DM(CE_i, G_j^{C_i}) \right). \quad (13)$$

Критерий отделимости определяет среднее межкластерное расстояние.

$$CQAC_S(CE) = \sum_{i=1}^{|CE|} DM(CE_i, DGE), DGE = CC_A(DG). \quad (14)$$

При этом DM в (13), (14) – это квадрат евклидова расстояния для попарно нормированных расстояний по признакам (8), (9) между графами (1) и центрами кластеров (6), и центрами кластеров и центром всего динамического графа (1).

Для кросс-проверки введем критерий близости результатов двух кластеризаций как расстояние между ближайшими центрами полученных кластеров (15). При оценке качества кластеризации необходимо рассматривать критерии как функционалы оптимизации (16), (17).

$$CQAC_{CS}(CE_x, CE_y) =$$

$$\left|\begin{array}{l} \\ CEU \leftarrow CE_x \cup CE_y \\ d \leftarrow 0 \\ для\ i = 1\ до\ |CEU| \\ \left|\begin{array}{l} если\ CEU_i \in CE_x \\ \left|CE_s = CE_y\right. \\ иначе \\ \left|CE_s = CE_x\right. \end{array}\right. \\ G \leftarrow G : DM(CEU_i, G) \rightarrow \min,\ G \in CE_s \\ если\ CE_s = CE_y \\ \left|CEU \leftarrow CEU \setminus G\right. \\ d \leftarrow d + DM(CEU_i, G) \\ d \end{array}\right. . \quad (15)$$

$$\begin{array}{l} CQAC_D \rightarrow \min \\ CQAC_S \rightarrow \max \end{array}, \quad (16)$$

$$\frac{CQAC_D}{CQAC_S} \rightarrow \min .$$

$$CQAC_{CS} \rightarrow \min . \quad (17)$$

Реализация. Программная реализация метода иерархической агломеративной кластеризации состояний динамического графа и критериев его оценки выполнена на языке программирования C# 4.0, благодаря чему полученный исходный код соответствует стандартам современного промышленного ПО и может использоваться в ИСППР при анализе ТТС [9].

Вычислительный эксперимент. Вычислительный эксперимент проводился для предметной области телефонии. В качестве исходных данных были взяты свободно распространяемые данные эксперимента MIT Reality Mining (майнинг реальности) [1]. В качестве исследуемого актора был выбран абонент с идентификатором 95, так как он является одним из участников с наибольшим количеством событий. В качестве выборки были взяты события голосовых звонков и текстовых сообщений за период с 1-о октября 2004 года по 31 января 2005 года включительно (123 дня), всего 3385 событий. Сформируем следующие тестовые выборки на основе трафиков (табл. 1).

Таблица 1 – Тестовые выборки трафиков

№	Временной интервал $t_1 \ldots t_n$	Количество событий	Количество ребер в G_{t_1} при $\Delta t = t_n - t_1$	Визуализация G_{t_1} при $\Delta t = t_n - t_1$, рис.
1	Октябрь 2004 г.	735	51	1, а
2	Ноябрь 2004 г.	608	49	1, б
3	Октябрь – ноябрь 2004 г.	1343	66	
4	Октябрь – декабрь 2004 г.	2469	86	
5	Январь 2005 г.	916	56	

Одним из исходных параметров алгоритма построения ДГ из выборки трафиков [10] есть величина Δt. Для генерации тестовых динамических графов для выборок (табл. 1) будем использовать значения

$$\Delta t \in \left\{\begin{array}{l} \{00:15:00\}, \{00:20:00\}, \{00:30:00\}, \{00:40:00\}, \\ \{01:00:00\}, \{02:00:00\}, \{03:00:00\}, \{04:00:00\}, \\ \{06:00:00\}, \{08:00:00\}, \{12:00:00\}, \{24:00:00\} \end{array}\right\}. \quad (18)$$

Выбор такого множества обусловлен кратностью 24 часов суток его элементам. Кроме того, это позволит сравнить результаты работы [11] с новыми результатами. Итого получим 106 тестовых динамических графов.

Проведем следующий ряд экспериментов.

1. Для тестовых ДГ выборки 1 выполним кластеризацию по методу (12), измерим критерии (13), (14). Оптимизируем функционалы (16). Сравним результаты с результатами работы [11]. Для лучшего решения приведем полную дендрограмму и характеристики полученных кластеров.

2. Для тестовых ДГ выборок 2, 3 выполним кластеризацию по методу (12), измерим критерии (13), (14). Сравним результаты с экспериментом № 1.

3. Для тестовых ДГ выборок 1 и 2, 4 и 5 выполним кластеризацию по методу (12), измерим критерий (17).

Результаты

Для первого эксперимента наилучшие результаты получены при $\Delta t = 6 часов$ (табл. 2). Приведем результаты работы [11], полученные для такой же выборке. В той работе ДГ оценивался методами корреляционного и спектрального анализа метрик максимальной степени вершины и инверсии коэффициента смежной корреляции состояний. Автокорреляционные функции обеих метрик устанавливаются в ноль при времени разделения $\approx 6 часов$. Спектры мощности обеих метрик имеют сильные пики для частот 1, 2, 3, 4 *раз в день* и менее сильные – для 6, 7 *раз в день*. Выбирая большую частоту при максимальной амплитуде, имеем время, равное 6 *часам*. Таким образом оба метода определяют наилучшее значение $\Delta t_{nat} = 6 часов$.

На рис. 2 показано, в каком порядке и на каких шагах алгоритма происходило слияние меньших кластеров в большие. Рамкой выделен шаг, на котором сработал критерий сечения дендрограммы.

Таблица 2 – Результаты кластеризации для ДГ выборки 1

Δt	$\lvert DG \rvert$	$\lvert DG^* \rvert$	$\lvert C \rvert$	$CQAC_D$	$CQAC_S$	$\frac{CQAC_D}{CQAC_S}$
{00:15:00}	2976	406	51	0.818667209	52.32268949	0.015646505
{00:20:00}	2232	383	51	0.841354489	51.7260158	0.016265596
{00:30:00}	1488	339	59	0.782622514	60.24045466	0.012991644
{00:40:00}	1116	314	53	0.81062516	55.78239775	0.014531917
{01:00:00}	744	279	58	0.769827382	60.14805388	0.012798874
{02:00:00}	372	207	53	0.811414242	55.66662872	0.014576314
{03:00:00}	248	170	50	0.625496404	53.47375102	0.011697261
{04:00:00}	186	139	53	0.473717346	56.27347074	0.008418129
{06:00:00}	124	102	51	0.381850645	54.69002854	0.006982089
{08:00:00}	93	90	37	0.413514208	40.64751736	0.010173173
{12:00:00}	62	62	2	NaN	1.046628736	NaN
{1.00:00:00}	31	31	2	NaN	1.153711165	NaN

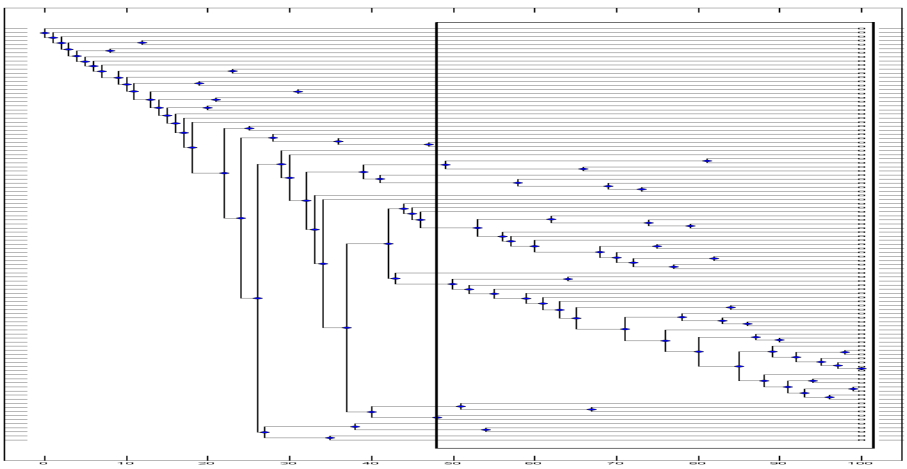


Рисунок 2 – Дендрограмма кластеризации для ДГ выборки 1 при $\Delta t = 6 часов$

В табл. 3 показаны характеристики кластеров из одного графа (первая строка), кластеров из 2 до 31 графов (остальные строки кроме последней), кластеров из одного графа, на которых сработал критерий сечения (последняя строка).

Таблица 3 – Характеристика кластеров для ДГ выборки 1 при $\Delta t = 6$ *часов*

Число графов в кластере	Число кластеров	Число ребер в графе минимальное	Число ребер в графе максимальное	Инцидентная вершина частого ребра
1	42	3	11	
4	1	1	2	6420
2	1	3	4	F216, C3C0, 6B37
14	1	2	4	F216, CD5A
3	1	1	2	6B37
31	1	1	3	CD5A
4	1	1	3	F216
1	2	1	1	30C9, C2A9

Во втором эксперименте для выборки 2 получены результаты (табл. 4), аналогичные первому эксперименту. Однако для выборки 3 лучшие результаты получены при $\Delta t = 8$ *часов* (табл. 5).

Таблица 4 – Результаты кластеризации для ДГ выборки 2

Δt	$\lvert DG\rvert$	$\lvert DG^*\rvert$	$\lvert C\rvert$	$CQAC_D$	$CQAC_S$	$\frac{CQAC_D}{CQAC_S}$
{00:15:00}	2880	365	46	0.89404941	48.11450031	0.018581704
{00:20:00}	2160	350	45	1.040051914	46.71392312	0.022264281
{00:30:00}	1440	307	46	1.022723273	48.44991349	0.021108877
{00:40:00}	1080	292	48	0.976264073	51.37536155	0.019002573
{01:00:00}	720	246	55	0.850494584	58.34618912	0.014576695
{02:00:00}	360	186	63	0.748929286	67.67831661	0.011066015
{03:00:00}	240	154	62	0.665593802	66.2070565	0.010053215
{04:00:00}	180	128	63	0.424993129	67.28141512	0.00631665
{06:00:00}	120	98	56	0.382922938	61.24570567	0.006252241
{08:00:00}	90	82	46	0.386817119	51.33712698	0.007534842
{12:00:00}	60	60	41	NaN	45.49026335	NaN
{1.00:00:00}	30	30	2	NaN	1.106184911	NaN

Таблица 5 – Результаты кластеризации для ДГ выборки 3

Δt	$\lvert DG\rvert$	$\lvert DG^*\rvert$	$\lvert C\rvert$	$CQAC_D$	$CQAC_S$	$\frac{CQAC_D}{CQAC_S}$
{00:15:00}	5856	771	73	1.263429	76.14864093	0.016591611
{00:20:00}	4392	733	74	1.347157	76.77669411	0.017546431
{00:30:00}	2928	646	84	1.297135	88.10144601	0.014723199
{00:40:00}	2196	606	79	1.264802	84.81571712	0.014912352
{01:00:00}	1464	525	92	1.146887	98.7953691	0.011608715
{02:00:00}	732	393	100	1.11603	108.6604372	0.010270804
{03:00:00}	488	324	97	0.910949	105.3574829	0.008646269
{04:00:00}	366	267	99	0.742463	106.8491368	0.006948707
{06:00:00}	244	200	94	0.602368	103.4027359	0.005825454
{08:00:00}	183	172	88	0.493901	97.50671992	0.005065302
{12:00:00}	122	122	70	NaN	78.0441743	NaN
{1.00:00:00}	61	61	2	NaN	1.099584775	NaN

Третий эксперимент показывает, что для выборок 1 и 2 лучший результат достигается при $\Delta t = 20\,минут$, для 4 и 5 при $\Delta t = 15 минут$, однако для 4 и 5 второй по качеству результат достижим при $\Delta t = 6\,часов$ (табл. 6).

Таблица 6 – Кросс-проверка кластеризаций для ДГ выборок 1 и 2, 4 и 5

Δt	$\lvert C_{Oct}\rvert$	$\lvert C_{Nov}\rvert$	$CQAC_{CS}$	$\lvert C_{Oct-Dex}\rvert$	$\lvert C_{Jan}\rvert$	$CQAC_{CS}$
{00:15:00}	51	46	27.40420647	90	46	13.07359485
{00:20:00}	51	45	25.49378214	90	48	13.34739791
{00:30:00}	59	46	31.43118865	111	50	15.68619717
{00:40:00}	53	48	29.10286169	106	49	13.79891127
{01:00:00}	58	55	35.92591047	124	52	14.94304977
{02:00:00}	53	63	37.73554584	131	52	16.15832546
{03:00:00}	50	62	36.67299703	131	53	18.66219759
{04:00:00}	53	63	41.90328759	127	50	18.41737447
{06:00:00}	51	56	41.06629134	127	37	14.79958916
{08:00:00}	37	46	33.88342885	124	45	18.91433202
{12:00:00}	2	41	21.77996787	100	2	0.917539185
{1.00:00:00}	2	2	1.082880811	2	2	1.017353096

Выводы

Анализируя результаты экспериментов можно сделать следующие выводы. Сравнение результатов работы [11] с результатами данной работы показывает, что выбор оптимального значения параметра Δt можно осуществить до кластеризации методами регрессионного и спектрального анализа, однако данную предобработку необходимо проводить для всей исходной выборки. Такой комплексный подход позволяет исключить оптимизацию функционалов (16).

Метод иерархической агломеративной кластеризации хорошо подходит для автоматического определения классов состояний динамической социальной сети. Получаемая дендрограмма допускает экспертное толкование. Так, дендрограмма на рис. 2 содержит 42 кластера из одного графа. Некоторые из них имеют до 11 ребер, что трактуется как редкое для него событие. Таким образом возможно использование этого подхода в ИСППР, когда система выдает аналитику рекомендацию на тщательную проверку поведения актора в определенные интервалы времени. 6 кластеров из 2 до 31 графа трактуются как частые штатные состояния сети. Для них возможно определение связей (ребер), вокруг которых кластеры были образованы. Данные кластеры могут использоваться в дальнейшем для поиска последовательных паттернов смены классов состояний. Особого внимания заслуживают последние два кластера из одного графа с одним ребром. При попытке объединить данные графы в один кластер алгоритм остановился, поскольку акторы 30C9, C2A9 разные и одиночные связи с ними не могут трактоваться как единый класс состояний. Это свидетельствует об адекватности метода и высоком качестве получаемого решения.

Выбор параметра Δt следует проводить на основе всей доступной выборки, если результаты кластеризации предполагается использовать в дальнейшем для этой же выборки, либо выборки, смежной с исходной. Видно, что при лучшем значении критериев и одинаковом размере исходной выборки, количество получаемых кластеров примерно одинаково. Однако при увеличении размера выборки растет и количество кластеров, что может говорить о неоднородности исходных данных.

Результат кросс-проверки подтверждает, что исходные выборки неоднородны. При увеличении размеров исторически первой выборки увеличивается Δt при лучшем

значении критерия. То есть, если имеется задача с обучающей и тестовой выборками и необходимо использовать результаты кластеризации обучающей выборки, то: при выборках одинакового размера необходимо выбирать низкое значение Δt ; но при обучающей выборке большого размера, чем тестовая, можно довериться формальным критериям.

Кроме задачи прогнозирования связей предложенный подход может использоваться в задачах поиска аномалий, анализа связей, ролей и позиций акторов. Несмотря на ограничение данной работы классом эгоцентрических сетей, предложенный подход должен работать и для других видов сетей при использовании соответствующих признаков и критериев останова в методе кластеризации.

Литература

1. Eagle N. Reality mining: sensing complex social systems / Nathan Eagle, Alex (Sandy) Pentland // Journal Personal and Ubiquitous Computing. – 2006. – Volume 10, Issue 4. – P. 255-268.
2. Eagle N. Eigenbehaviours: identifying structure in routine / Nathan Eagle, Alex Sandy Pentland // Behavioral Ecology and Sociobiology – 2009. Volume 63, Issue 7. – P. 1057-1066.
3. Liben-Nowell D. The Link Prediction Problem for Social Networks / D. Liben-Nowell, J. Kleinberg // Proceedings of the twelfth international conference on Information and knowledge management. – New York : ACM. – 2003. – P. 556-559.
4. Huang Z. The Time Series Link Prediction Problem with Applications in Communication Surveillance / Zan Huang, Dennis K.J. Lin // INFORMS Journal on Computing – 2009. – Volume 21, Issue 2. – P. 286-303.
5. Васильев А.В. Применение алгоритмов кластеризации и классификации в задачах обработки и интерпретации телеметрической информации / [Васильев В.А., Геппенер В.В, Жукова Н.А., Клионский Д.М., Тристанов А.Б.] // Доклады 9-й международной конференции «Цифровая обработка сигналов и ее применение», 28 – 30 марта 2007 г. – М. : ИПУ РАН. – 2007. – С. 389-392.
6. Marketos G. Mobility Data Warehousing and Mining [Электронный ресурс] / Gerasimos Marketos, Yannis Theodoridis // Proceedings of 35th International Conference on Very Large Data Bases PhD Workshop (VLDB'09). – Lyon, 2009. – 6 pp. – Режим доступа : http://infolab.cs.unipi.gr/pubs/confs/VLDB09PhDWorkshop.pdf
7. Bayir M.A. Discovering Spatiotemporal Mobility Profiles of Cellphone Users [Электронный ресурс] / Murat Ali Bayir, Murat Demirbas, Nathan Eagle // Proceedings of 10th IEEE International Symposium on a «World of Wireless, Mobile and Multimedia Networks», 15 – 19 June, 2009 – Kos, Greece. – 2009. – 9 p. – Режим доступа : http://reality.media.mit.edu/pdfs/bayir.pdf
8. Савельев О.О. Постановка задачи исследования прогнозирования связей в трафиках телефонных сетей / О.О. Савельев, А.И. Шевченко // Восточно-Европейский журнал передовых технологий. – 2012. – № 6/3 (60). – С. 51-60.
9. Савельев О.О. Интеллектуальная система поддержки принятия решений при анализе трафиков телефонных сетей / О.О. Савельев // Материалы 14-й Международной научно-технической конференции «Системный анализ и информационные технологии» (SAIT 2012), Киев, 24 апреля 2012 г. – К. : УНК «ИПСА» НТУУ «КПИ», 2012. – С. 224-225.
10. Савельев О.О. Построение динамического социального графа по транзакционным данным трафиков телефонных сетей / О.О. Савельев // Матеріали доповідей VI Міжнародної науково-практичної конференції молодих учених, аспірантів, студентів «Сучасна інформаційна Україна: інформатика, економіка, філософія», 26 квітня 2012 р. – Донецьк : Наука і освіта. – 2012. – С. 79-83.
11. Савельев О.О. Определение естественного периода активности абонента телефонной сети / О.О. Савельев, А.И. Шевченко // Сборник докладов IV Всеукраинской научно-технической конференции студентов, аспирантов и молодых ученых «Информационные управляющие системы и компьютерный мониторинг», Донецк, 24 – 25 апреля 2013 г. – Донецк : ДонНТУ, 2013. - С. 755-759.
12. Воронцов К.В. Лекции по алгоритмам кластеризации и многомерного шкалирования [Электронный ресурс] / К.В. Воронцов – М. : МГУ. – 2007. – 18 с. – Режим доступа : http://www.ccas.ru/voron/download/Clustering.pdf

Literatura

1. Eagle N. Reality mining: sensing complex social systems / Nathan Eagle, Alex (Sandy) Pentland // Journal Personal and Ubiquitous Computing. – 2006. – Volume 10, Issue 4. – P. 255-268.

2. Eagle N. Eigenbehaviours: identifying structure in routine / Nathan Eagle, Alex Sandy Pentland // Behavioral Ecology and Sociobiology – 2009. Volume 63, Issue 7. – P. 1057-1066.
3. Liben-Nowell D. The Link Prediction Problem for Social Networks / D. Liben-Nowell, J. Kleinberg // Proceedings of the twelfth international conference on Information and knowledge management. – New York : ACM. – 2003. – P. 556-559.
4. Huang Z. The Time Series Link Prediction Problem with Applications in Communication Surveillance / Zan Huang, Dennis K.J. Lin // INFORMS Journal on Computing – 2009. – Volume 21, Issue 2. – P. 286-303.
5. Vasiliev A.V., Geppener V.V, Zhukova N.A., Klionskij D.M., Tristanov A.B. Using clustering and classification algorithms in problems of processing and interpretation of the telemetry data [Primenenie algoritmov klasterizacii i klassifikacii v zadachah obrabotki i interpretacii telemetricheskoj informacii]. Doklady 9 Mezhdunarodnoj Konferencii "Cifrovaja obrabotka signalov i ee primenenie" (Proc. of 9th Int. Conf. "Digital Signal Processing and its Application"). Moscow, 2007. – P. 389-392.
6. Marketos G. Mobility Data Warehousing and Mining [Электронный ресурс] / Gerasimos Marketos, Yannis Theodoridis // Proceedings of 35th International Conference on Very Large Data Bases PhD Workshop (VLDB'09). – Lyon, 2009. – 6 p. – Режим доступа : http://infolab.cs.unipi.gr/pubs/confs/VLDB09PhDWorkshop.pdf
7. Bayir M.A. Discovering Spatiotemporal Mobility Profiles of Cellphone Users [Электронный ресурс] / Murat Ali Bayir, Murat Demirbas, Nathan Eagle // Proceedings of 10th IEEE International Symposium on a "World of Wireless, Mobile and Multimedia Networks", 15-19 June, 2009 – Kos, Greece. – 2009. – 9 p. – Режим доступа : http://reality.media.mit.edu/pdfs/bayir.pdf
8. Saveliev O.O., Shevchenko A.I. Research Problem Statement of Links Prediction in Phone Networks Traffics [Postanovka zadachi issledovanija prognozirovanija svjazej v trafikah telefonnyh setej]. Vostochno-Evropejskij Zhurnal Peredovyh Tehnologij – Eastern-European Journal of Enterprise Technologies. – 2012. – № 6/3 (60). – P. 51-60.
9. Saveliev O.O. Intelligence Decision Support System for Phone Network Traffic Analysis [Intellektual'naja sistema podderzhki prinjatija reshenij pri analize trafikov telefonnyh setej]. Materialy 14-j Mezhdunarodnoj nauchno-tehnicheskoj konferencii "Sistemnyj analiz i informacionnye tehnologii SAIT 2012" (Proc. of 14th Int. scientific conf. «System Analysis and Information Technologies SAIT 2012»). – Kyiv, 24 April 2012. – P. 224-225.
10. Saveliev O.O. Building a dynamic social graph from transactional data of phone networks traffics [Postroenie dinamicheskogo social'nogo grafa po tranzakcionnym dannym trafikov telefonnyh setej]. Materiali dopovidej VI mizhnarodnoї naukovo-praktichnoї konferenciї molodih uchenih, aspirantiv, studentiv «Suchasna informacijna Ukraïna: informatika, ekonomika, filosofija» (Proc. of 6th Int. scientific conf. of young scientists «Modern Information Ukraine: computer science, economics, philosophy»). – Donetsk, 26 April 2012. – P. 79-83.
11. Saveliev O.O., Shevchenko A.I. Determination of the natural period of telephone network subscriber activity [Opredelenie estestvennogo perioda aktivnosti abonenta telefonnoj seti]. Sbornik dokladov IV Vseukrainskoj nauchno-tehnicheskoj konferencii studentov, aspirantov i molodyh uchenyh «Informacionnye upravljajushhie sistemy i komp'juternyj monitoring» (Proc. of 4th Int. scientific conf. of young scientists «Information Control Systems and Computer Monitoring»). – Donetsk, 24 – 25 April 2013. – P. 755-759.
12. Vorontsov K.V. Lekcii po algoritmam klasterizacii i mnogomernogo shkalirovanija (Lectures on Clustering and Multidimensional Scaling Algorithms) (2007). Available at: http://www.ccas.ru/voron/download/Clustering.pdf (accessed 1 June 2013).

RESUME

O.O. Saveliev, A.I. Shevchenko

The Mining the Classes of States from Monitoring Traffic of Dynamic Social Network

The article examines the task of the mining the classes of states from monitoring traffic of dynamic social network. This task is a subtask of previously proposed approach for solving time series link prediction problem. The dynamic graph model is proposed for modeling of dynamic social network. Research was limited for egocentric social network. Literature review showed that most obvious approach for task solving is a clustering. The

hierarchical agglomerative clustering was chosen from all of clustering methods thanking its good fit on domain. The task was formalized in the mix of pattern recognition and clustering terms.

The preprocessing of source states was proposed, like empty graphs filtration and determination the best sampling period of monitoring traffic by regression and spectral analysis. The edge presence and edge weight were selected as static graph attributes. The euclidean distance measurement and average linkage rule were used. The requirement that all graphs in every cluster should produce connected graph was used as stop criteria. The graph union and edge weights averaging was used for building cluster center. Density and separability criterias and cross-checking was used for quality assessment of clustering.

The series of experiments on an MIT Reality Mining dataset took place. The results showed the correctness of the approach and sufficient quality of task solution. Hierarchical agglomerative clustering fits the approach requirements. The obtained dendrograms allow good interpretation by expert. Conclusion about rules of sampling period of monitoring traffic selection was done. Centers of resulting clusters could be used for mining of states change patterns and anomalous states detection.

Proposed automatic approach could be used in intelligence decision support system for phone traffic analysis.

Статья поступила в редакцию 09.07.2013.