*У статті сформульована задача кластеризації соціальних мереж, наведений її розв'язок за допомогою алгоритму кластеризації BSP*

*Ключові слова: соціальна мережа, кластеризація, BSP*

*В статье сформулирована задача кластеризации социальных сетей, приведено ее решение с помощью алгоритма кластеризации BSP*

*Ключевые слова: социальная сеть, кластеризация, BSP*

*In the article the problem of social network clustering is considered, its solution using BSP clustering algorithm is proposed*

*Keywords: social network, clustering, BSP*

УДК 004.032.26

# КЛАСТЕРИЗАЦИЯ СОЦИАЛЬНЫХ СЕТЕЙ С ПОМОЩЬЮ АЛГОРИТМА КЛАСТЕРИЗАЦИИ BSP

**Е.А. Бойко**
Кафедра искусственного интеллекта
Харьковский национальный университет радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166
Контактный тел.: (057) 702-13-37
E-mail: helenboyko@ukr.net

## 1. Введение

Социальная сеть представляет собой социальную структуру, состоящую из объектов (пользователей), также называемых узлами, которые соединены между собой связями, отображающими отношения между узлами и взаимодействие между ними [1]. Как правило, социальная сеть описывается графом или матрицей взаимоотношений.

За последние годы интерес к социальным сетям возрос во много раз. При этом тенденция роста сохраняется. Социальные сети являются богатым источником данных. Использование виртуальных социальных сетей, а также использование методов интеллектуального анализа данных повлияло на развитие исследования социальных сетей, внедрив множество новых методик и технологий [2].

В интеллектуальном анализе данных достаточно хорошо исследована задача кластеризации [3]. Традиционные алгоритмы кластеризации разделяют объекты на кластеры в зависимости от их сходства. Объекты из одного кластера похожи друг на друга и очень отличаются от объектов из других кластеров. Кластеризация в социальных сетях, в отличие от традиционной задачи кластеризации, делит объекты на кластеры не только по их атрибутам, но также и по связям между объектами. Основной проблемой при кластеризации в социальных сетях является поиск методов для решения вопроса о том, как объединить объекты в кластеры, основанные на связях этих объектов [4].

В данной статье проводится исследование одного из алгоритмов, который применяется для анализа социальных сетей.

## 2. Постановка задачи

Анализ социальных сетей (Social Network Analysis, SNA) – это направление, которое занимается описанием и анализом социальных сетей посредством различных методов, в том числе с помощью теории графов, теории информации и математической статистики [5].

Одной из основных задач анализа социальных сетей является определение групповых структур сетей (сообществ).

Сообщество в сети является подграфом графа связей, узлы которого плотно связаны между собой и слабо связаны с остальной частью сети. С этой точки зрения, поиск таких групп узлов в сети может рассматриваться как задача кластеризации.

Кластеризации в анализе социальных сетей отличается от традиционной кластеризации. Она требует группировки объектов не только в зависимости от значения их атрибутов, но также и в зависимости от связей между этими объектами.

Для анализа задачи кластеризации выбран алгоритм BSP. В данной работе ставится задача исследовать алгоритм кластеризации BSP и разработать программное приложение, реализующее данный алгоритм. Для программной реализации поставленной задачи выбран язык программирования Java.

## 3. Алгоритм кластеризации BSP

Алгоритм кластеризации business system planning (BSP) предложен компанией IBM. Этот алгоритм использует объекты (бизнес-процессы) и связи между объектами (классы данных) для проведения кластерного анализа. Социальные сети также включают в себя объекты и связи между этими объектами. Поэтому алгоритм BSP может быть использован и при анализе социальных сетей.

Социальную сеть можно представить в виде ориентированного графа, состоящего из объектов и связей между ними. На рис. 1 представлен пример фрагмента социальной сети. Окружность представляет объект, например, пользователя. Линии со стрелкой являются ребрами графа и представляют собой направленную связь между двумя объектами.

Пусть $O_i$ – объект в социальной сети $(i=1,...,m)$, $E_j$ – направленная связь между двумя объектами, направленное ребро графа $(j=1,...,n)$.

Существуют 2 типа отношения достижимости между объектами: длиной в один шаг и длиной в несколько шагов. Два объекта $O_i$ и $O_j$ находятся в отношении достижимости длиной в один шаг, если существует направленная связь от $O_i$ к $O_j$, проходящая через одно и только одно направленное ребро. Например, на рис. 1 существует направленная связь от объекта $O_1$ к объекту $O_2$, проходящая через направленное ребро $E_1$. Два объекта $O_i$ и $O_j$ находятся в отношении достижимости длиной в несколько шагов, если существует направленная связь от $O_i$ к $O_j$, проходящая через два или более направленных ребер.

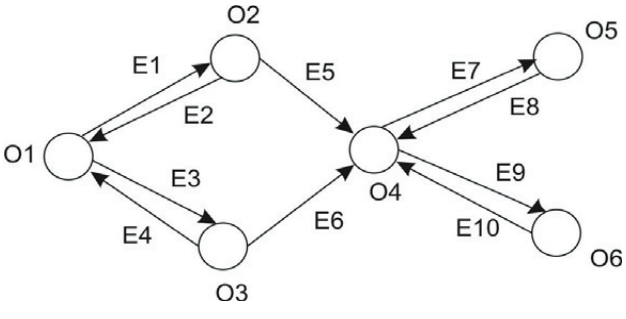



Рис. 1. Пример представления социальной сети в виде ориентированного графа

Для графа объектов социальной сети и связей между ними необходимо определить две матрицы $L_c$ и $L_p$. Пусть $L_c$ – матрица размерностью $m \times n$, которая определяет вершины и исходящие из них дуги. В данной матрице значение элемента $L_c(i,j)=1$, если объект $O_i$ является вершиной, из которой выходит дуга $E_j$. Значение элемента матрицы $L_c(i,j)=0$, если объект $O_i$ не является вершиной, из которой выходит дуга $E_j$. Пусть $L_p$ – матрица размерностью $m \times n$, которая определяет вершины и входящие в них дуги. В данной матрице значение элемента $L_p(i,j)=1$, если объект $O_i$ является вершиной, в которую входит дуга $E_j$. $L_p(i,j)=0$, если объект $O_i$ не является вершиной, в которую входит дуга $E_j$.

После определения матриц $L_c$ и $L_p$ вычисляется матрица достижимости длиной в один шаг по формуле

$$G = L_c \bullet L_p^T = \left( g_{i,j} = \bigvee_{k=1}^{n} \left( l_c(i,k) \wedge l_p^T(k,j) \right), i=1,...,m, j=1,...m \right), \quad (1)$$

где $\wedge$ – булево произведение, $\vee$ – булева сумма.

В полученной матрице значение элемента $G(i,j)=1$ обозначает, что объекты $O_i$ и $O_j$ находятся в отношении достижимости длиной в один шаг. Значение элемента матрицы $G(i,j)=0$ обозначает, что объекты $O_i$ не находятся в отношении достижимости длиной в один шаг $O_j$.

Кроме одношагового отношения достижимости, также между объектами существуют отношения достижимости длиной в несколько шагов. Расчет матрицы двухшаговой достижимости производится по следующей формуле

$$G^2 = G \bullet G = \left( g_{i,j}^2 = \bigvee_{k=1}^{m} \left( g(i,k) \wedge g(k,j) \right), i=1,...,m, j=1,...m \right) \quad (2)$$

Расчет матрицы $m-1$ -шаговой достижимости производится по следующей формуле

$$G^{m-1} = G^{m-2} \bullet G \quad (3)$$

Расчет обобщенной матрицы достижимости производится по следующей формуле

$$R = I \vee G \vee G^2 ... \vee G^{m-1}, \quad (4)$$

где $I$ – единичная матрица.

Значение элемента матрицы $R(i,j)=1$ означает, что существует отношение достижимости от объекта $O_i$ к объекту $O_j$.

Отношение достижимости не является симметричным. Значение элемента матрицы $R(i,j)=1$ означает, что существует отношение достижимости от объекта $O_i$ к объекту $O_j$, но это не означает, что также существует отношение достижимости от объекта $O_j$ к объекту $O_i$.

Поэтому требуется рассчитать матрицу взаимодостижимости объектов, основываясь на расчете матрицы $R$.

Расчет матрицы взаимодостижимости производится по следующей формуле

$$Q = R \wedge R^T, \quad (5)$$

где $R^T$ – транспонированная обобщенная матрица достижимости.

В рассчитанной матрице значение элемента $Q(i,j)=1$ означает, что существует отношение взаимодостижимости между объектом $O_i$ и объектом $O_j$.

Если в социальной сети два объекта находятся в отношении взаимодостижимости, они принадлежат к одному и тому же классу. Поэтому процесс кластеризации пользователей социальной сети основывается на процессе анализа рассчитанной матрицы взаимодостижимости объектов $Q$.

Таким образом, согласно матрице взаимодостижимости $Q$, следующим этапом алгоритма кластеризации BSP является разделение социальной сети на кластеры.

Этот процесс основывается на выделении сильно связанных подграфов и определении матриц сильной связности этих подграфов. Матрица является матрицей сильной связности, если все элементы в ней равны 1.

После кластеризации социальной сети также можно определить отношения между полученными кластерами. Этот процесс можно осуществить на основе кластеров и матрицы одношаговой достижимости $G$. Если существует отношение достижимости длиной в один шаг между двумя объектами из разных кластеров, то существует направленная связь между соответствующими кластерами. На рис. 2 представлено определение отношений между двумя полученными кластерами на основе связей, представленных на рис. 1.

После применения алгоритма образовались кластеры $C_1$ и $C_2$. Связи между кластерами отображаются в виде связей между узлами $O_2$ и $O_4$, а также между узлами $O_3$ и $O_4$.

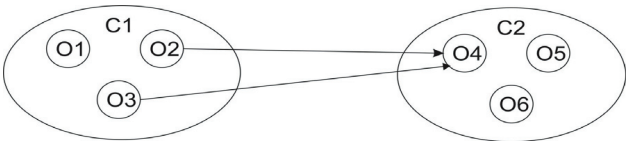



Рис. 2. Определение связей между кластерами

Основным недостатком данного подхода к реализации алгоритма кластеризации BSP является то, что он использует матрицы для хранения связей и существующих отношений достижимости. При анализе реально существующих социальных сетей эти матрицы будут иметь большую размерность, что затрудняет их загрузку в память и обработку. Матрица порядка n требует для своего хранения $n^2$ байт оперативной памяти, а время вычислений пропорционально $n^3$.

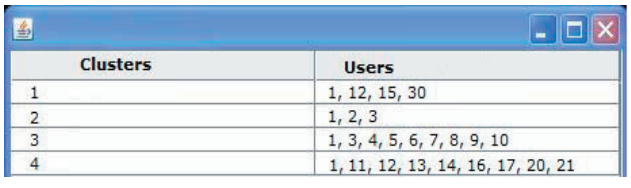
Одной из модификаций алгоритма кластеризации BSP является алгоритм, который для хранения данных об объектах и связях между ними использует не матрицы, а структуры данных в виде связного списка для устранения недостатков в работе, указанных выше. Элемент списка представляет собой структуру данных, состоящую из трех полей: i – позиция элемента объекта, являющегося элементом матрицы, по строкам; j – позиция объекта по столбцам; value – вес связи между двумя объектами. Эта структура данных обозначается триплетом $(i, j, value)$. Для ускорения обработки этих триплетов, они упорядочиваются в связном списке по возрастанию позиций по строкам, а затем по возрастанию позиций по столбцам. Если данные о весах связей не используются или веса связей имеют одинаковые значения, структура данных будет состоять из двух полей, указывающих позиции элемента объекта, являющегося элементом матрицы, по строкам и по столбцам. При этом данная структура данных обозначается парой $(i, j)$.

## 4. Экспериментальные исследования

В реализации алгоритма BSP были использованы выборки, содержащие данные о связях между узлами, которые могут храниться в формате txt или xml. Данные представляют собой пары чисел – индекс узла, от которого направлена связь и индекс узла, к которому эта связь направлена.

На рис. 3 представлены результаты работы приложения для трех итераций алгоритма в виде таблицы кластеров и пользователей.

С увеличением количества итераций алгоритма количество кластеров уменьшается. Это можно объяснить тем, что анализируется большее количество связей между объектами. При этом время работы алгоритма увеличивается.



| Clusters | Users |
|---|---|
| 1 | 1, 12, 15, 30 |
| 2 | 1, 2, 3 |
| 3 | 1, 3, 4, 5, 6, 7, 8, 9, 10 |
| 4 | 1, 11, 12, 13, 14, 16, 17, 20, 21 |

Рис. 3. Таблица кластеров и пользователей для трех итераций алгоритма

## 5. Выводы

В данной работе был исследован алгоритм кластеризации BSP для анализа социальных сетей. На основе полученных результатов можно сделать выводы о том, что при увеличении количества итераций алгоритма BSP, количество кластеров уменьшается, так как анализируется больше связей между объектами.

Время работы при этом увеличивается. Использование связных списков вместо матриц для хранения данных о связях между объектами улучшает работу алгоритма по объему использованной памяти и времени работы.

Для дальнейшего улучшения работы алгоритма и точности кластеризации возможно изменение структуры для хранения данных об узлах и связях между ними, а также использование алгоритма кластеризации BSP вместе с алгоритмами кластеризации на основе анализа профилей пользователей.

Литература

1. Xu, G. Web Mining and Social Networking: Techniques and Applications [Text] / Guandong Xu, Yanchun Zhang, Lin Li. – Springer, 2010. – 228 p.
2. Scott, J.P. Social Network Analysis: A Handbook [Text] / John P Scott. – Sage Publications Ltd, 2000. – 240 p.
3. Mitchell, T. Machine Learning [Text] / Tom Mitchell. – McGraw-Hill Science/Engineering/Math, 1997. – 432 p. – ISBN 978-0-070-42807-2.
4. Social Network Data Analytics [Text] / edited by Charu C. Aggarwal. – Springer, 2011. – 516 p. – ISBN 978-1-4419-8461-6.
5. Компьютерная визуализация социальных сетей [Электронный ресурс] / Журнал «КомпьютерПресс». Режим доступа: www/ URL: http://www.compress.ru/article.aspx?id=16593&iid=771/ – 19.04.12 г. – Загл. С экрана.