ЗАСТОСУВАННЯ МЕТОДУ K-CORE DECOMPOSITION ДЛЯ ПРОВЕДЕННЯ ОПТИМАЛЬНОЇ КЛАСТЕРИЗАЦІЇ

Кириченко О.Л., Остапов С.Е.
Чернівецький національний університет імені Юрія Федьковича, м. Чернівці

У роботі розглянуто результати пошуку оптимальної кількості кластерів методом k-Core decomposition та роботу алгоритму PIC (Power iteration clustering) для проведення дослідження кластерної структури деяких зон веб-простору (українського – net.ua; edu.ua, польського сегменту – edu.pl та ізраїльського – ac.il) на основі раніше зібраних статистичних даних.

Задача кластеризації полягає в розподіленні деякої множини об’єктів на групи схожих об’єктів, що називаються кластерами. За допомогою розробленого нами кроулера було просканировано більше двох мільйонів веб-сторінок та зібрана інформація для дослідження деяких сегментів веб-простору: українського (net.ua; edu.ua), ізраїльського (ac.il) та польського (edu.pl). Для кожної зони побудовані матриці суміжності, встановлено ступінь кожного вузла, визначено коефіцієнт кластеризації, побудовано розподіл ймовірностей вузлів за вхідними та вихідними зв’язками, побудовані графи різних зон веб-простору [1]. Проведено розбиття на кластери за допомогою алгоритму PIC (Power iteration clustering). Для визначення оптимальної кількості кластерів ми застосували метод k-Core decomposition, за допомогою якого ми знайшли число k (мінімальний ступінь кожного вузла підграфа) та виділили k-ядро, кожна вершина, якого може розглядатись як центр кластера. k-Core графа – це максимально зв'язний підграф, в якому кожна вершина пов'язана принаймні з k вершинами в підграфі [2]. Головна мета розподілу k-core – знайти сильну підгрупу, члени якої грають роль комунікаторів на графі. Кожен вузол в підграфі повинен мати хоча б ступінь k.

Таким чином, за допомогою методу k-Core decomposition нами визначено оптимальну кількість кластерів для досліджуваних підмереж (таблиця) та за допомогою алгоритму PIC проведено дослідження кластерної структури різних сегментів веб-простору.

Таблиця. Результати методу k-Core decomposition

Назва сегменту веб-простору	Значення k	Кількість кластерів (k-ядро)
польський сегмент веб-простору (edu.pl)	337	213
ізраїльський сегмент веб-простору (ac.il)	365	196
український сегмент веб-простору (net.ua)	121	264
український сегмент веб-простору (edu.ua)	95	229

Література:

1. Kyrychenko Oksana Comparison of Statistical Characteristics of Certain Internet Subdomains / Oksana Kyrychenko, Sergey Ostapov, Igor Kanovsky // Monograph. Scientific Publishing of the Academy of Business in Dabrowa Gornicza: Wydawnictwo Naukowe. – 2014. – 138 p. **2.** Sheng-Tzong Cheng. Using k-Core Decomposition to Find Cluster Centers for k-Means Algorithm in GraphX on Spark / Cheng Sheng-Tzong, Yin-Chun Chen, and Meng-Shuan Tsai. – CLOUD COMPUTING 2017: The Eighth International Conference on Cloud Computing, GRIDs, and Virtualization.