АНОТАЦІЯ

Кириченко О.Л. Дослідження статистичних характеристик складних мереж методами інтелектуального аналізу даних. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 121 – «Інженерія програмного забезпечення» – Чернівецький національний університет імені Юрія Федьковича, Чернівці, 2023.

Дисертаційна робота присвячена дослідженню статистичних характеристик складних мереж та кластерної структури веб-простору з використанням методів інтелектуального аналізу даних, зокрема розробці інформаційної технології для кластеризації даних великого розміру, які були зібрані й оброблені спеціально створеним програмним забезпеченням. Також вивчено стохастичні матриці, які завдяки особливостям своїх спектральних властивостей є основним математичним об’єктом при дослідженні кластерної структури веб-простору.

Результати роботи є підґрунттям для подальших теоретичних і практичних наукових розробок із досліджень проблематики теорії складних мереж.

Дисертація складається зі вступу, чотирьох розділів, висновків, переліку використаних джерел та чотирьох додатків. У **вступі** обґрунтовано актуальність теми дослідження, сформульовано мету, завдання, предмет, об’єкт та методи дослідження, вказано наукову новизну, теоретичне та практичне значення отриманих результатів, подано та проаналізовано зв’язок роботи з науковими темами. Зазначено особистий внесок здобувача, а також наведено відомості про апробацію та публікації основних результатів дисертації. Описано структуру та обсяг дисертаційної роботи.

Перший розділ містить ключові відомості з теорії складних мереж, опис основних напрямів досліджень та завдання, якими займається теорія складних мереж. Тут проведено огляд та опис основних моделей (Ердоша–Рені, Уаттса–Строґаца, Барабаші–Альберт), які призвели до сьогоднішнього

розуміння цього напряму. Розглянуто та проаналізовано приклади реальних складних мереж (онлайнові, наукової співпраці, WWW, цитування наукових праць, Інтернет, транспортна, різні біологічні мережі тощо) та їх особливості Здійснено класифікацію та огляд методів однієї з важливих технік інтелектуального аналізу складних мереж – кластерного аналізу.

У **другому розділі** дисертаційного дослідження описано концепцію кроулінгу як одного із засобів збирання інформації, проведено огляд існуючих програмних засобів для збирання інформації у веб-просторі. Описано розроблене власне програмне забезпечення (кроулер), який сканує веб-простір, завантажує та зберігає знайдені гіперпосилання з веб-стоінок у базу даних. Перевагою даної розробки над існуючими аналогами є наявність аналітичного модуля, який надає можливість проводити статистичний та кластерний аналіз отриманого веб-графу.

Другий розділ має прикладне значення, основним його результатом є розроблене спеціалізоване програмне забезпечення – кроулер з вбудованим аналітичним модулем для інтелектуальної обробки інформації.

Третій розділ присвячений дослідженню освітніх сегментів веб-простору (українського (edu.ua), ізраїльського (ac.il) та польського (edu.pl), інформація про які була зібрана та оброблена за допомогою самостійно розробленої інформаційної технології, детальний опис якої проведено в пункті 3.3. Застосування даної розробки дозволило отримати статистичні характеристики та кластерну структуру вказаних вище сегментів веб-простору та здійснити порівняльний аналіз.

Для проведення кластеризації важливо знати оптимальну кількість кластерів, у розділі описано два класичних методи знаходження оптимальної кількості кластерів (метод «ліктя» та k-core decomposition), проведено порівняльний аналіз, який показав їх узгодженість щодо оптимальної кількості кластерів для кожного досліджуваного сегменту веб-простору.

Основні результати даного розділу можна підсумувати наступним чином:

- для збирання та проведення статистичного і кластерного аналізу даних у складних мережах розроблено інформаційну технологію;
- для підмереж edu.ua, edu.pl та ac.il проведено порівняльний аналіз статистичних характеристик та їх кластерної структури ;
- встановлено, що всі три підмережі відповідають сучасним тенденціям розвитку глобальної мережі інтернет, володіють властивостями безмасштабних графів, причому виявилось, що український сегмент edu.ua є найменш розвиненою структурою з найменшою кількістю вузлів у кластерах.

У **четвертому розділі** розглянуто питання кластеризації в графі на основі матриці суміжності. Основним об'єктом дослідження даного розділу є стохастична матриця P, що задає ймовірності переходу на графі та визначається із матриці суміжності. У даному розділі детально проаналізовано спектральні властивості стохастичної матриці P із врахуванням кластерної структури графу.

Основні теоретичні результати цього розділу можна описати наступним чином:

- доведено факт збіжності власних значень матриці P за умов, накладених на елементи матриці суміжності A (теорема 4.3.1). Причому, накладені умови послаблені порівняно із класичними результатами, де вимагається існування скінченного другого моменту для елементів матриці суміжності;
- встановлений факт про асимптотичну еквівалентність спектрів матриць P та $\tilde{P}$ дозволяє використовувати стохастичну матрицю із незалежними елементами замість відповідної стохастичної матриці P, елементи якої не є незалежними (лема 4.4.1). Даний результат дозволяє користуватися класичними результатами щодо розподілу власних значень випадкових матриць та переносити дані твердження на матриці із слабко корельованими елементами;

– у твердженнях пункту 4.5. розглянуто частинний підхід до оцінки розподілу елементів матриці P за умови показникового розподілу елементів матриці A (лемах 4.5.1 та 4.5.2). Такий підхід дозволив розробити новий алгоритм перевірки належності елементів (вершин графу) до одного кластеру.

На основі отриманих теоретичних результатів проведено порівняльний аналіз з класичними методами кластеризації, а саме: методом «ліктя», k-core decomposition та методом силуету. У результаті проведених досліджень, побудовано критерій оцінки оптимальної кількості кластерів k_{opt}, обчислення якого ґрунтується на власних значеннях стохастичної матриці P, а саме

$$\hat{k}_{opt} = \#\{\lambda_i(P): Re(\lambda_i(P)) > \max|Im(\lambda_i(P))|\}.$$

Використовуючи метод Монте – Карло, вдалося встановити, що у ряді випадків, запропонований спектральний метод визначення кількості кластерів дає більш точні оціночні значення кількості кластерів у графі в порівнянні з відомими методами («ліктя», k-core decomposition та методом силуету), що задається стохастичною матрицею P чи матрицею суміжності A. Крім того, запропонований новий метод є менш чутливим до наявності кластерів різної розмірності.

У **висновках** підсумовано основні результати дисертаційного дослідження.

У **додатках** подано наукові публікації, в яких відображено основні наукові результати роботи, відомості про апробацію результатів дисертації – акти та довідки про впровадження результатів роботи, діаграма основних класів кроулера та їх опис, лістинг частини коду програми.

Теоретичне значення. Результати теоретичних досліджень, а саме розвитку теорії графових досліджень, сформульовані та доведені леми і теореми, можуть використовуватися для подальших досліджень у цій галузі, а також у навчальних курсах кафедр математичних проблем управління та кібернетики та програмного забезпечення комп’ютерних систем

Чернівецького національного університету імені Юрія Федьковича (та інших ЗВО), пов'язаних з інтелектуальним аналізом даних, методичних розробках, навчальних посібниках для освітнього процесу та науково-дослідної роботи студентів аспірантів.

Практичне значення. Розроблені у дисертаційній роботі кроулер, інформаційна технологія та метод визначення оптимальної кількості кластерів можуть в подальшому використовуватися для практичного дослідження складних мереж. Запропоновані підходи до архітектури аналітичного модуля використовуються у компанії «Квант Азимут» для розробки власного програмного забезпечення та компанії «Qlicks B.V.» – для проведення сегментації клієнтів на різні категорії, які потім ефективно використовуються для персоналізованих маркетингових кампаній і стратегій та передбачення поведінки клієнтів на основі аналізу покупок, історії пошуку або профілей в соціальних мережах.

Ключові слова: модель (математична, економічна), моделювання, динаміка, інтелектуальний аналіз даних, кластеризація, k-means, інформаційна система, інформаційна технологія, інтелектуальна система, програмне забезпечення, тестування програмного забезпечення, рівні тестування програмного забезпечення, специфікація вимог до програмного забезпечення, функціональні та нефункціональні вимоги до програмного забезпечення, статистичні методи.

ABSTRACT

Kyrychenko O.L. The Study of Statistical Characteristics of Complex Networks by Methods of Intelligent Data Analysis. – Qualification research work published in the manuscript.

Dissertation for the degree of Doctor of Philosophy, speciality 121 – "Software Engineering" – Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 2023.

The dissertation deals with the study of statistical characteristics of complex networks and the cluster structure of the web space using methods of intelligent data analysis, in particular, the development of information technology for the clustering of large data, which were collected and processed by specially created software. In addition, stochastic matrices have been studied, which, due to their specific spectral properties, are the main mathematical object in the study of the cluster structure of the web space.

The results of the research are the basis for further theoretical and practical scientific development in the research of the problems of the theory of complex networks.

The dissertation contains an introduction, four chapters, conclusions, a list of literature, and four appendices. The **introduction** substantiates the relevance of the research topic, formulates the goal, task, subject, object and research methods; indicates the scientific novelty, theoretical and practical significance of the obtained results; presents and analyzes the link between the current research and scientific topics. The personal contribution of the candidate, as well as information about the approval and publication of the main results of the research are shown. The structure and scope of the dissertation are outlined.

Chapter 1 contains key information on the theory of complex networks, a description of the main areas of research and tasks that the theory of complex networks deals with. An overview and description of the main models (the Erdős–Rényi, Watts-Strogatz, Barabási–Albert models) having led to current insight into this trend are provided here. Examples of real complex networks (online, scientific collaboration, WWW, citation of scientific works, Internet, transport, various biological networks, etc.) and their features are considered and analyzed. The methods of cluster analysis, an important technique of intellectual analysis of complex networks, are classified and reviewed.

Chapter 2 of the dissertation describes the concept of crawling as one of the means of gathering information, and provides an overview of existing software tools for collecting information in the web space. The own development of the

software (crawler) is described, which scans the web space, downloads and stores the found hyperlinks from the web sites in the database. The advantage of this development over existing analogues is the presence of an analytical module, which enables to conduct statistical and cluster analysis of the resulting web graph.

Chapter 2 has a practical value; its main result is the developed specialized software – a crawler with a built-in analytical module for intelligent information processing.

Chapter 3 deals with the study of following educational segments of the web space: Ukrainian (edu.ua), Israeli (ac.il) and Polish (edu.pl). The information on these segments was collected and processed using a personally developed information technology, a detailed description of which is provided in point 3.3. The application of this development enables to obtain the statistical characteristics and cluster structure of the above-mentioned segments of the web space and to carry out a comparative analysis.

As far as it is important to know the optimal number of clusters to carry out clustering, two classical methods of finding the optimal number of clusters (the "elbow" method and *k-core* decomposition) are described in the chapter. A comparative analysis was carried out, which showed their agreement regarding the optimal number of clusters for each studied segment of the web space .

The main results of this section can be summarized as follows:

- information technology has been developed for collecting information and conducting statistical and cluster analysis of data in complex networks;
- a comparative analysis of statistical characteristics and their cluster structure was carried out for the *edu.ua, edu.pl* and *ac.il* subnets;
- it was established that all three subnets correspond to modern trends in the development of the global Internet network, have the properties of scale-free graphs; in addition, it turned out that the Ukrainian segment of *edu.ua* is the least developed structure with the least number of nodes in clusters.

Chapter 4 deals with the issue of clustering in a graph based on the adjacency matrix. The main object of research in this section is the stochastic matrix P, which

specifies the transition probabilities on the graph and is determined from the adjacency matrix. In this chapter, the spectral properties of the stochastic matrix P are analyzed in detail, taking into account the cluster structure of the graph.

The main theoretical results of this chapter can be described as follows:

- the fact of the convergence of the eigenvalues of the matrix P under the conditions imposed on the elements of the adjacency matrix A is proved (theorem 4.3.1). Moreover, the imposed conditions are weaker compared to the classical results, which require the existence of a finite second moment for the elements of the adjacency matrix;
- the established fact about the asymptotic equivalence of the spectra of the matrices P and $\tilde{P}$ allows using a stochastic matrix with independent elements instead of the corresponding stochastic matrix *P*, the elements of which are not independent (lemma 4.4.1). This result enables to use classical results on the distribution of eigenvalues of random matrices and transfer these statements to matrices with weakly correlated elements;
- in the statements of point 4.5 a partial approach to the estimation of the distribution of the elements of matrix P is considered under the condition of the indicative distribution of matrix A elements (lemmas 4.5.1 and 4.5.2). This approach made it possible to develop a new algorithm for checking whether elements (graph vertices) belong to one cluster.

Based on the obtained theoretical results, a comparative analysis was carried out with classical clustering methods, namely: the "elbow" method, k-core decomposition, and the silhouette method. The research resulted in building a criterion for estimating the optimal number of clusters k_{opt}, the calculation of which is based on the eigenvalues of the stochastic matrix P, namely

$$\hat{k}_{opt} = \#\{\lambda_i(P): Re(\lambda_i(P)) > \max|Im(\lambda_i(P))|\}.$$

Using the Monte-Carlo method, it was possible to determine that in a number of cases, the proposed spectral method for finding the number of clusters gives more accurate estimates of the number of clusters in the graph in comparison with

known methods ("elbow", k-core decomposition and the silhouette method), given by the stochastic matrix P or adjacency matrix A. In addition, the proposed new method is less sensitive to the presence of clusters of different dimensions.

The main results of the dissertation research are summarized in the **conclusions**.

The **appendices** present scientific publications reflecting the main scientific results of the work, information on the approval of the results of the dissertation: acts and certificates on the implementation of the results of the research, a diagram of the main classes of the crawler and their description, and listing of part of the software code.

Theoretical significance. The results of theoretical research, namely the development of the theory of graph research, formulated and proven lemmas and theorems, can be used for further research in this field. They can also be applied in the educational courses of the Departments of Mathematical Problems of Management and Cybernetics and Software of Computer Systems of Yuriy Fedkovych Chernivtsi National University (and other higher educational institutions), related to the intellectual analysis of data, methodological developments, teaching aids for the educational process and research work of graduate and postgraduate students.

Practical significance. The crawler, information technology, and method for determining the optimal number of clusters developed in the dissertation can be used for further practical research of complex networks. The proposed approaches to the architecture of the analytical module are used by the company "Kvant Azimuth" for the development of its own software and by the company "Qlicks B.V." for segmenting customers into different categories, which are then effectively used for personalized marketing campaigns and strategies and for predicting customers' behavior based on purchase analysis, search history or social media profiles.

Keywords*:* model (mathematical, economic), simulation, dynamics, intelligent data analysis, clustering, k-means, information system, information

technology, intelligent system, software, software testing, software testing levels, specification of software requirements, functional and non-functional software requirements, statistical methods.