UDK 004.4
https://doi.org/10.31891/csit-2022-4-7

OKSANA KYRYCHENKO
Yuriy Fedkovich Chernivtsi National University (Ukraine)

INFORMATION TECHNOLOGY FOR STATISTICAL CLUSTER ANALYSIS OF INFORMATION IN COMPLEX NETWORKS

Information technology has been developed, which is used to collect, process and save large volumes of data from the web space. With the help of technology, the statistical characteristics of various segments of the web space and their cluster structure are studied. Two methods are used to find the optimal number of clusters and cluster centers: the well-known k-core decomposition algorithm and a new method developed by the authors. The new algorithm is based on the distribution of eigenvalues of the stochastic matrix, which describes the process of Markov transitions in the system. The clustering process is carried out using the Power iteration clustering algorithm.

With the help of written software (crawler), information is collected on a given segment of the web space. For the studied area, there are statistical characteristics, namely: node degree, clustering coefficient, node probability distributions by input and output connections. Oriented and unoriented graphs of web pages of the studied zones are constructed. By combining the calculated dependencies for the input and output subnets, we can obtain the statistical characteristics of the undirected graphs of the web pages of the web space zones that we are investigating.

For cluster analysis, the optimal number of clusters and cluster centers can be found in 2 ways: by the well-known k-core decomposition algorithm and by using a new method developed by the author. The new algorithm is based on the distribution of eigenvalues of the stochastic matrix, which describes the process of Markov transitions in the system. Using the Rower iteration clustering algorithm, the cluster structure of various segments of the web space is studied.

The advantage of the developed information technology is that with its help one can work with large sets of data collected on the Internet, study their structure and statistical characteristics, and perform the clustering process. To implement the clustering process and find the optimal number of clusters and centroids a new algorithm is suggested. The results of the algorithm indicate high accuracy in determining the optimal number of clusters.

Keywords: optimal number of clusters; cluster centers; k-core decomposition algorithm; eigenvalues; stochastic matrix; clustering process; statistical characteristics, process of Markov.

ОКСАНА КИРИЧЕНКО
Чернівецький національний університет ім. Ю. Федьковича

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПРОВЕДЕННЯ СТАТИСТИЧНО-КЛАСТЕРНОГО АНАЛІЗУ ІНФОРМАЦІЇ У СКЛАДНИХ МЕРЕЖАХ

Велика кількість інформації в Інтернеті та й загалом сам інформаційний простір являють собою складну мережу з усіма характерними для таких структур статистичними характеристиками, особливостями та зв'язками. Вивчення статистичних особливостей і кластерної структури таких мереж, а також найбільших доменів і зон цікавить сьогодні багатьох дослідників і вчених.

Розроблена інформаційна технологія, за допомогою якої проводиться збір, обробка та збереження даних великих об'ємів з веб-простору. За допомогою інформаційної технології досліджуються статистичні характеристики різних сегментів веб-простору та досліджується їх кластерна структура.

За допомогою написаного програмного забезпечення (кроулера) проводиться збір інформації по заданому сегменту веб-простору. Для досліджуваної зони знаходяться статистичні характеристики, а саме: ступінь вузла, коефіцієнт кластерності, розподіли ймовірностей вузлів за вхідними та вихідними зв'язками. Будуються орієнтовані та неорієнтовані графи веб-сторінок досліджених зон. Об'єднуючи розраховані залежності для вхідних та вихідних підмереж, можна отримати статистичні характеристики неорієнтованих графів веб-сторінок зон веб-простору, які досліджуємо.

Для проведення кластерного аналізу знаходиться оптимальне число кластерів та центри кластерів 2 способами: відомим алгоритмом k-core decomposition та за допомогою нового методу, розробленого автором. Новий алгоритм базується на розподілі власних значень стохастичної матриці, що описує процес Маркова переходів у системі. За допомогою алгоритму Power iteration clustering проводиться дослідження кластерної структури різних сегментів веб-простору.

Перевагою розробленої інформаційної технології є те, що з її допомогою можна працювати з великими масивами даних, зібраних в Інтернеті, вивчати їх структуру та статистичні характеристики, здійснювати процес кластеризації. Для реалізації процесу кластеризації та знаходження оптимальної кількості кластерів і центроїдів запропоновано новий алгоритм. Результати роботи алгоритму свідчать про високу точність визначення оптимальної кількості кластерів.

Ключові слова: оптимальне число кластерів, центри кластерів; алгоритм k-core decomposition; власні значення; стохастична матриця; процес кластеризації; статистичні характеристики, марковський процес.

Introduction

A large amount of information on the Internet, and in general the information space itself, represent a complex network with all the statistical characteristics, features and connections typical for such structures. The study of statistical features and cluster structure of such networks, as well as the largest domains and zones, is of interest to many researchers and scientists today.

In the theory of complex networks, three main directions of research are considered:

• studying the statistical characteristics that specify the behavior of networks;

• creating a network model;

• predicting the network behavior when its structure changes.

Active development of this field of research led to the study of network characteristics, taking into account not only its topology, but also statistical characteristics that characterize the behavior of the network when the structural properties change.

For this, researchers study the statistical characteristics of various networks: energy networks, transport networks, air transport networks, computer networks, co-authorship networks, social networks, the Internet and many others [1-5].

Many works are devoted to studying the structure of the WWW space and statistical characteristics of the web space [1-5]. The structure of such networks is presented by graphs. The nodes of the graph are the web pages, and the edges are the links between them. Both directed and undirected graphs are studied [2-4]. It was found that the World Wide Web obeys the statistical laws of complex networks, and it was established that the distribution of the nodes of the graph, which reflects the World Wide Web, obeys a power law with the indicator close to (-2,2) for input connections and (-2, 7) – for output ones [2-4]. This indicates the scalelessness of such a network, i.e. the high level of development of the network as a whole [1-4].

Description of information technology

The goal of our research consisted in the development of information technology for collecting, processing, saving and conducting statistical cluster analysis of information in complex networks. It is possible to study a complex network in several stages. The structure of the developed information technology is shown in Fig. 1.

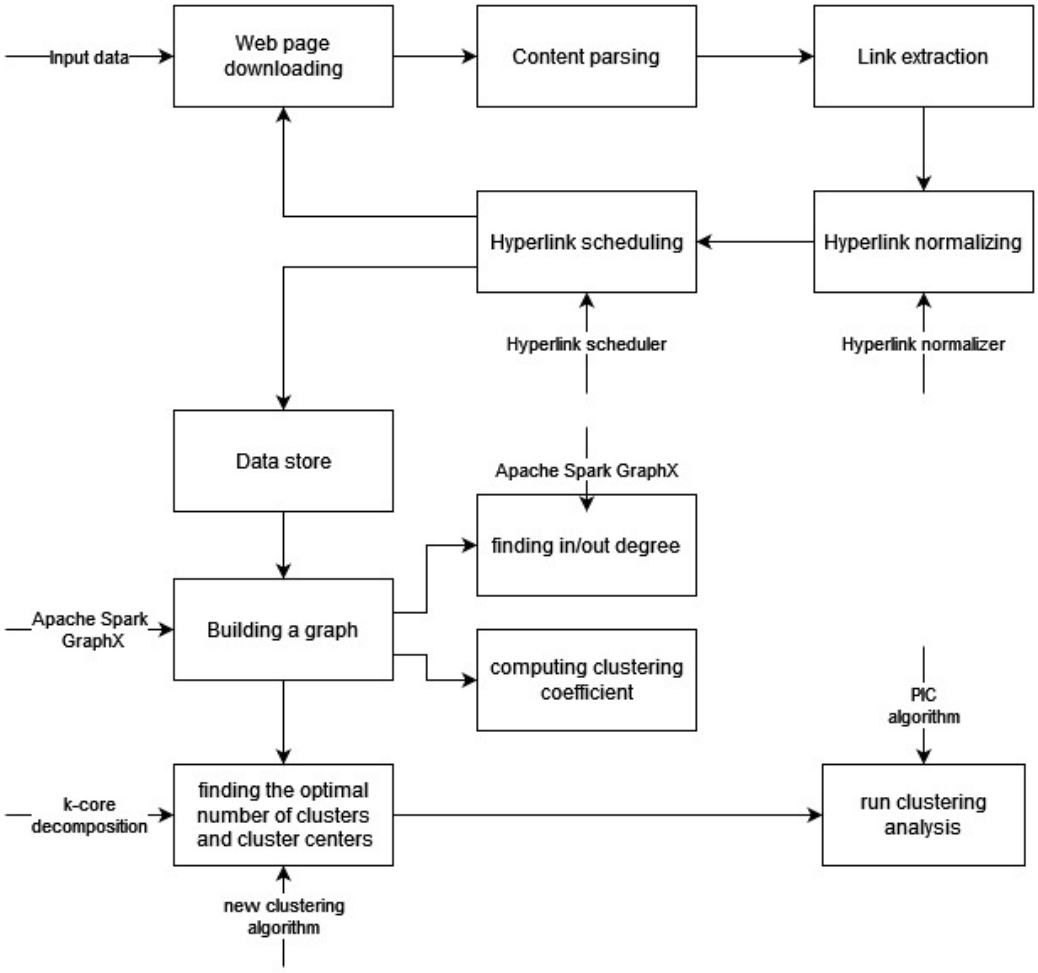

Fig.1. General scheme of information technology for conducting statistical cluster analysis of information in complex networks

We have developed the software (crawler) that is completely controlled by the settings before starting work. The user specifies a list of web page addresses – entry points, and if necessary, indexing depth, etc. This enables to fully control the process of searching and indexing pages as well as the calculating the main statistical parameters of the network under study [6]. The program was configured for a given number of "jumps" from the entry point on its links. The results of the research are recorded to the database, and the crawler moves on to the next entry point. If because of "wandering" the crawler comes across a page that is already in the database, no repeated study is required,

only a new connection is added. As a result, we obtain a graph of web pages, the statistical characteristics of which will be studied. For the area under study, the degree of each node is found, the clustering coefficient is determined, and probability distributions of nodes are constructed based on input and output connections. Combining the calculated dependencies for the input and output subnetworks, the statistical characteristics of the undirected graphs of the web pages of the investigated web space zones can be obtained.

Thus, a graph is built and the clustering process should be carried out. The process of clustering enables to consider large data sets and drastically reduce their dimensionality, make them compact and explore their structure. The task of clustering consists in dividing a set of objects into groups of similar objects, called clusters. The result of the clustering process is a set of clusters that contain similar elements of the input set of elements.

Concerning the clustering task, there arises a problem of choosing a clustering algorithm: it should be chosen in such a way that the division into clusters be the most correct.

One of the most common problems with clustering algorithms is that for most of them the number of clusters as an input parameter is required; however, the number of clusters is usually not known in advance. That is why the use of certain empirical rules to choose the optimal number of clusters is needed.

A large class of clustering algorithms is based on representing the sampling as a graph. The vertices of the graph correspond to the objects of the sampling while the edges correspond to the pairwise distances between the objects. An adjacency matrix is constructed, which is then examined. Spectral clustering is often used to divide a set of large dimensions into clusters. The data dimension is reduced using the spectral clustering algorithms. The next step consists in applying some clustering method (e.g., k-means). It should be noted that the main drawback of algorithms based on the k-means method is the requirement to determine the initial number of clusters and their centers.

Information about these parameters is usually not available at the initial stage of the information space research.

Another disadvantage of the k-means method is that it does not provide a solution to the problem of determining the optimal number of clusters in the data sampling.

After the stage of graph building, we determine the optimal number of clusters and find the cluster centers. Such process can be performed in two ways:

• by the k-Core decomposition method [7];

• by a new method developed by the authors, which is based on the asymptotic distribution of the eigenvalues of a stochastic random matrix without conditions of element independence, the spectrum of which can be decomposed into a regular part and outliers [8].

The next step is to perform the clustering process, i.e. the division into clusters, which is carried out using the Power Iteration Clustering (PIC) algorithm [9].

Unlike the spectral clustering algorithms, the PIC algorithm calculates only one eigenvector (which is actually a linear combination of several eigenvectors). In this way, a high speed of calculations is achieved if compared to the traditional spectral clustering algorithm [9]. The PIC algorithm pseudocode is given below:

Input: Normalized similarity matrix W, number of clusters k
Output: Clusters $C_1, C_2, ..., C_k$

1. Pick an initial vector $\mathbf{v}^0$.
2. $\mathbf{v}^{t+1} \leftarrow \frac{W\mathbf{v}^t}{\|W\mathbf{v}^t\|_1}$ and $\boldsymbol{\delta}^{t+1} \leftarrow |\mathbf{v}^{t+1} - \mathbf{v}^t|$.
3. Increment t and repeat above step until $|\boldsymbol{\delta}^t - \boldsymbol{\delta}^{t-1}| \simeq \mathbf{0}$.
4. Use k-means on $\mathbf{v}^t$ and return clusters $C_1, C_2, ..., C_k$.

Performing such tasks results in division of the collected data set into clusters and the possibility to compare the performance of different clustering algorithms.

k-core decomposition method

To determine the optimal number of clusters, we used the k-Core decomposition method.

A k-Core graph is a maximally connected subgraph in which every vertex is connected to at least k vertices in the subgraph. The k-core distribution is often used in largescale network analysis [7]. Its main aim is to find a strong subgroup, the members of which play the role of communicators on the graph. Each node in the subgraph must have at least k degree.

k-Core decomposition has the following properties:

$$\forall u \in V\colon k\text{-}core\,(u) = k \leftrightarrow$$

$$\begin{cases} \text{There is such maximal subgraph } V_k \text{ that } \forall v \in V_k\colon deg(v) \geq k, \\ \text{and} \\ \text{There is no such subgraph } V_{k+1} \text{ that } \forall v \in V_{k+1}\colon deg(v) \geq k+1. \end{cases}$$

This algorithm consists in finding the subgraph with the strongest connections *k*. This means that each member of this subgraph has at least *k* neighbors. In addition, there is no larger subgraph where each member has more than *k* neighbors. Therefore, if we find a vertex that has the highest degree in this subgraph, it will be a good candidate for its cluster center [7].

New clustering method

Along with the classical clustering method, for example k-means, in our approach we will use a new spectral clustering method, in which the selection of clusters is based on the transition matrix (eigenvalues) *P*. In [8, 10 – 12], the asymptotic distribution of the eigenvalues of the stochastic matrix *A* with random elements a_{ij} □ *dist*, where *dist* is the basic distribution of elements. It was determined that the normalized matrix with the elements (1)

$$p_{ij} = \frac{a_{ij}}{\sum_{j=1}^{N} a_{ij}}, \tag{1}$$

corresponds to the asymptotic distribution of the real parts of the eigenvalues of the matrix *P* with a distribution density

$$f_\lambda(x;N) = c\sqrt{(a_+ - x)(x - a_-)}, \quad x \in (a_-, a_+),$$

where a_-, a_+ – distribution parameters determined by the following relations (2)

$$a_{\pm} = \frac{\sigma^2}{\sqrt{N}}, \tag{2}$$

c – normalizing constant, N – the number of elements (elements in the network) in the matrix *P*. At the same time, it was established that the statement about the distribution of eigenvalues has a circular distribution.

Moreover, the number of clusters is suggested to be chosen according to the following rule

$$k_{opt} = \#\left\{\lambda_i : |\lambda_i(P) - 1| \le \frac{1}{\alpha\sqrt{N}}\right\}, \tag{3}$$

where the constant *α* depends on the intercluster connection in the matrix *A*.

Process of technology testing

This information technology was used to conduct a statistical cluster study of the following areas of the web space: the Polish segment of the web space (edu.pl), the Israeli segment (ac.il) and the Ukrainian (net.ua and edu.ua). For each segment of the web space, the probability distribution of nodes according to incoming connections (in degree) and the probability distribution of nodes according to degrees by outgoing connections (out degree) are constructed. The average values of a node degree for undirected graphs were constructed and determined [13]. The graphs of the investigated web space zones were built. The network clustering coefficients were calculated (Table 1). The data indicate a large number of the nearest-neighbors cross-references [13].

Table 1.

Clustering coefficients for subnetworks

Name of zone	edu.ua	net.ua	ac.il	edu.pl
Clustering coefficient	0.11	0.13	0.104	0.088

We can see that for all networks the clustering coefficients are within the value of 0.1, which also indicates similar statistical characteristics of all studied segments. It can be concluded from the research results that according to their statistical characteristics the studied segments of the web space (net.ua; edu.ua – Ukrainian, ac.il – Israeli and edu.pl - Polish) belong to scale-free networks. This fully corresponds to the modern trends in the development of the Internet.

The study of the cluster structure of the obtained data sets began with determining the optimal number of clusters using two described methods: the k-Core decomposition and a new spectral method developed by the authors

[8]. Based on the Monte Carlo methods it was found that the new spectral algorithm is more accurate in estimating the number of clusters. The networks under study were divided into clusters using the PIC algorithm.

Conclusion

The advantage of the developed information technology is that with its help one can work with large sets of data collected on the Internet, study their structure and statistical characteristics, and perform the clustering process. To implement the clustering process and find the optimal number of clusters and centroids a new algorithm is suggested. The results of the algorithm indicate high accuracy in determining the optimal number of clusters.

References

1. Broder A. Graph structure in the web / A. Broder, R. Kumar, F. Maghoul et al. // Proceedings of the 9th World Wide Web Conference, Computer networks, 2000. - 33 (1). – P. 309-320.
2. Newman, M.E.J. The Structure and Function of Complex Networks. *SIAM Review*. – 2003. – Vol. 45. – N. 2. – P. 167–256.
3. M. E. J. Newman. The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA 98, 404–409 (2001).
4. S. H. Strogatz. Exploring complex networks. *Nature* 410, 268–276 (2001)
5. Newman, M. E. J. Models of the small world: a review. J. Stat. Phys. 101, 819–841 (2000).
6. Kyrychenko O.L., Kanovsky I., Ostapov S.E. Software for the statistical characteristics of the global WWW network research. *Information Processing Systems*, 2013, Vol. 2(3), pp. 99-104.
7. Sheng-Tzong Cheng, Yin-Chun Chen, and Meng-Shuan Tsai, 'Using k-Core Decomposition to Find Cluster Centers for k-Means Algorithm in GraphX on Spark', CLOUD COMPUTING 2017: The Eighth International Conference on Cloud Computing, GRIDs, and Virtualization.
8. Kyrychenko O.L., Malyk I.V., Ostapov S.E. Cluster structure analysis of Internet networks based on random matrixes. *International Scientific Technical Journal "Problems of Control and Informatics",* 2022, 1, pp. 37-46.
9. Frank Lin and William W. Cohen, 'Power iteration clustering', in ICML(to appear), (2010).
10. Robert C. Qiu, Paul Antonik (2017). Smart Grid using Big Data Analytics. A Random Matrix Theory Approach. Wiley Online Library, 2017.
11. T. Tao and V. Vu. Random matrices: Universality of the local eigenvalue statistics. *Acta Math*. 206 (2011), no. 1, 127–204.
12. T. Tao and V. Vu. Random matrices: universality of local eigenvalue statistics up to the edge. *Comm. Math. Phys.* 298 (2010), no. 2, 549–572.
13. Oksana Kyrychenko, Sergey Ostapov, Igor Kanovsky. Comparison of Statistical Characteristics of Certain Internet Subdomains. Monograph. Scientific Publishing of the Academy of Business in Dabrowa Gornicza: Wydawnictwo Naukowe, 2014. – 138 p. (ISBN: 978-83-62897-91-9)

Oksana Kyrychenko **Оксана Кириченко**	Assistant Professor, Department of Mathematical Problems of Control and Cybernetics, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine, e-mail: o.kyrychenko@chnu.edu.ua https://orcid.org/0000-0003-0282-9958	асистент кафедри математичних проблем управління і кібернетики, Чернівецький національний університет імені Юрія Федьковича, Чернівці, Україна.