Research Article

Directed correlation network of concepts determined by the dynamics of publications

Dmytro Lande [a], *Leonard Strashnoy* [b] (✉) and *Irina Balagura* [a]

[a] *Institute for Information Recording of NAS of Ukraine*
[b] *UCLA, Institute Infectious disease department, USA*

Abstract:

A technique for forming, clustering and visualizing so-called directed correlation networks is herein proposed. The links between nodes of such networks correspond to the values of cross-correlations between vectors - sets of parameters corresponding to these nodes modified in a certain way. To build network structures for each node (topic), vectors are formed - arrays of numbers corresponding to a certain time series. As an example, the article considers a time series generated by the Google Books Ngram Viewer service.

This approach, unlike the existing one, has advantages such as intuitive and realistic rules, the definition of the weight of nodes and links, a reliable mathematical basis for correlation analysis, an accounting of previously unused parameters of time series of publications corresponding to entities, allowing one to the group said entities according to their trends in time, and objectivity and relative simplicity. This technique can be based on data obtained, for example, from content monitoring systems, and can be used in analytical systems for various purposes in order to generalize a set of variables without explicit links between them.

Keywords:

Cross-correlation network, publication dynamics, Google Books Ngram Viewer, visualization of network structures, cluster analysis.

✉ Corresponding Author E-mail: lstrashnoy@gmail.com

Introduction

Modern information technologies cannot be imagined without methods and tools for processing network structures, but the structural features are not always clearly expressed. There is always a question of how to build a network in order to apply a wide range of methods and tools for processing it, to obtain and interpret the results if the researcher has only certain entities – nodes -- at his disposal, but the connections between them are not clearly defined. If a single entity can be represented as a homogeneous multidimensional vector of parameters, it is possible to establish similarity relationships, and apply classification or cluster analysis methods to identify groups of similar documents.

In this paper, a method is proposed that puts the dynamics vector corresponding to the distribution of documents by dates (years) in accordance with the essence (a concept from the subject area). More specifically, each year is assigned a number-the number of times an entity appears in publications covered by the Google Books system. The dimension of this vector corresponds to the number of years and the length of the time interval during which the array of publications was analyzed.

Goals

The purpose of this paper is to present a methodology for forming, clustering, and ranking nodes and visualizing so-called directed correlation networks, graph structures, and relationships between nodes (concepts, entities) that correspond to the values of correlations between sets of parameters corresponding to these entities.

At the same time, it should be noted that correlation does not directly mean the presence of causal relationships, so correlation networks cannot be considered as causal, semantic maps. At the same time, correlation, along with other criteria, can be considered as the basis for probabilistic estimates of similarity. In other words, correlation networks can be considered as the basis for applying fuzzy semantic network technologies, for example, for further scenario analysis.

Method

To build network structures, vectors corresponding to entities are formed for each concept/entity. For this purpose, the use of the Google Books Ngram Viewer service is provided (https://books.google.com/ngrams). This system allows you, in part, to get arrays of numbers corresponding to the relative frequency of appearance of an entity in publications by year. To get these

arrays, the system can be accessed via the user interface by entering the name of the entity.

After forming vectors corresponding to individual entities, a correlation network is formed, which can be considered as a way to store and visualize entities that are objectively related to each other [1]. Indeed, it is possible to form vectors of dynamics for various entities, the relationship between which is not always explicit.

Figure 1 shows a fragment of the interface for obtaining dynamics corresponding to the topic "Artificial Intelligence".

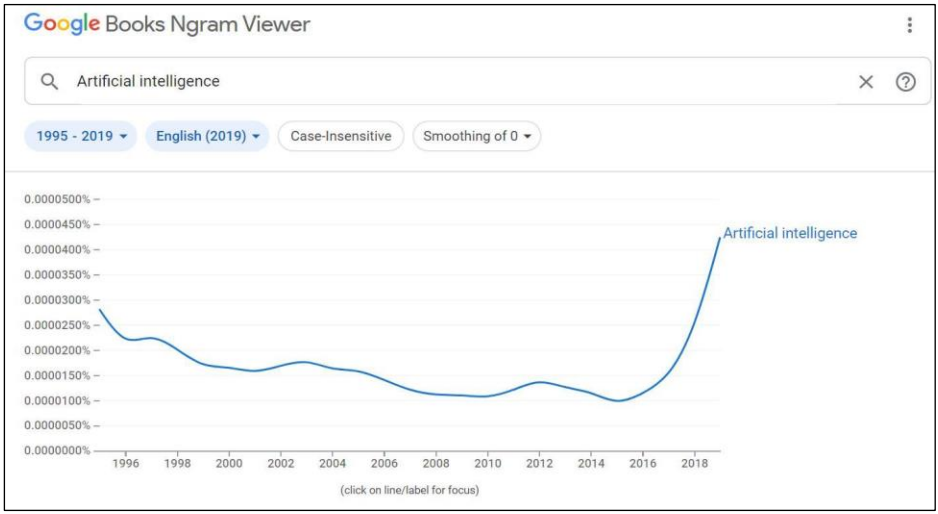

Figure 1 – A fragment of the interface of the Google Books Ngram Viewer server, where the vector of dynamics of the emergence of the concept of Artificial Intelligence is represented as a graph

Below, a method is proposed for constructing a network of interconnections of entities (concepts), consisting of the following stages [2, 3]:

1) A request to the Google Books Ngram Viewer service is generated for each entity. The analysis period is also defined - the dimension of the corresponding dynamics vectors.
2) As a result of performing queries, a set of dynamics vectors corresponding to the given concept is determined, similar to those shown in Fig. 1.
3) The set of maximal cross-correlations between the obtained vectors is calculated, and the corresponding correlation matrix with elements is formed:

$$a_{ij} = \max_{m} \frac{\sum_{k=1}^{n-m} w_{k+m}^{i} w_{k}^{j}}{\sqrt{\sum_{k=m+1}^{n} \left(w_{k}^{i}\right)^{2}} \sqrt{\sum_{k=1}^{n-m} \left(w_{k}^{j}\right)^{2}}}. \quad (1)$$

Here each entity s_k from the multitudes $S = \{s_k\}_{k=1}^{|S|}$, a vector of parameter values is assigned to each entity in the set $\overline{w^k} = \left(w_1^k, w_2^k, ..., w_n^k\right)$, where n is the number of elements in the set of parameters.

The max function is used for the reasons that processes are similar in nature and may have similar dynamic behavior, but it is possible with a time shift.

4) The adjacency matrix is formed in accordance with formula (1) and stored in a CSV file. Due to the fact that the adjacency table contains links between all nodes, according to [1], links whose value is less than some selected threshold are ignored. The choice of this threshold completely depends on the experience of analysts. In the information technology described, a correlation matrix is formed and transmitted for processing and visualization to the network structure analysis system Gephi (https://gephi.org/) [4]. Gephi is a fast and simple program for visualization and analysis of network structures, provides the fast layout, efficient data research, as well as visualization of large-scale networks. At the same time, the CSV adjacency matrix for the Gephi system has some features that need to be taken into account (zeros on the diagonal, the location of the characters ";" and others).
5) The values of this matrix are sent in CSV format to the Gephi system. This system has a number of modes, among which the "Data Lab" mode is used for monitoring network characteristics. In this mode, in addition to the usual degrees of matrix nodes, you can calculate their values by PageRank, Hits, modularity, and so on. In addition, there are options for ranking matrix nodes (entities) by these parameters.
6) Object group modularity classes are defined and the loaded network structure is then clusterized [1,3]. Modularity is calculated as the difference between the fraction of edges within a cluster in the

network under consideration and the expected fraction of edges within a cluster in a network where vertices have the same degree as in the original one, but the edges are randomly distributed. The modularity of the network can be expressed by the formula:

$$Q = \frac{1}{2m}\sum_{i,j}\left[a_{ij} - \frac{k_i k_j}{2m}\right]\ \delta\left(c_i, c_j\right), \qquad (2)$$

where a_{ij} is the element of the adjacency matrix A, m is the number of edges in the graph, k_i, k_j are the steps of the corresponding nodes, and δ is the Kronecker Delta function (shows whether the nodes are located *i and j* in the same modularity class).

7) Network visualization is performed in the Gephi system.
8) At the last stage, an expert interpretation of the results is performed.

Improving the method

It is proposed to take into account two points when constructing a correlation network, namely:

1) which process started first;
2) absolute values of time series for mutual correlation, i.e. to determine the value of the directed connection between nodes *A* and *B* in proportion to the sum of the values of the numerical series corresponding to node *A*.

Let s_k vector of parameter values be matched to each element from a set of objects $S = \{s_k\}_{k=1}^{|S|}$ be matched with a vector of parameter values $\overline{w^k} = \left(w_1^k, w_2^k, ..., w_n^k\right)$, where n is the number of elements in this set.

To implement the point 1, the formula determining the relationship between objects i and j (1), is used:

$$a_{ij} = \max_{0<m\le K} \frac{\sum_{k=1}^{n-m} w_{k+m}^i w_k^j}{\sqrt{\sum_{k=m+1}^{n}\left(w_k^i\right)^2}\sqrt{\sum_{k=1}^{n-m}\left(w_k^j\right)^2}}, \qquad (1)$$

where *K* is the width of the window of possible time offsets.

The max function is used for the following reasons: processes that are similar in nature may have similar dynamic behavior, but it is possible with a time shift. In contrast to the method described above, accounting m is performed not according to the range of $[-K, K]$, but in an interval $[1, K]$.

To implement the second point, each of the matrix elements a_{ij} is multiplied by the value of the sum of the values of the corresponding vector $v_i = C\sum_{k=1}^{n} w_k^j$, where C is the normalizing constant.

When further using the Gephi visualization tools, the network was defined as directional, the node sizes corresponded to the node degrees of the weighted directed network, clusterization, if necessary, is calculated using the OpenOrd or Fruchterman Reingold algorithms, and node modularity is calculated with Resolution = .5 .

Examples

As a demonstration example, let's consider three entities (Node1, Node2, Node3), each of which corresponds to a time series:

Node1: (0, 1, 2, 3, 4, 5, 4, 3, 2, 1, 0, 0, 0)
Node2: (0, 0, 0, 0, 1, 2, 3, 4, 3, 2, 1, 0, 0)
Node3: (0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 1, 0)

The processes corresponding to these three nodes are shown in Figure 2.

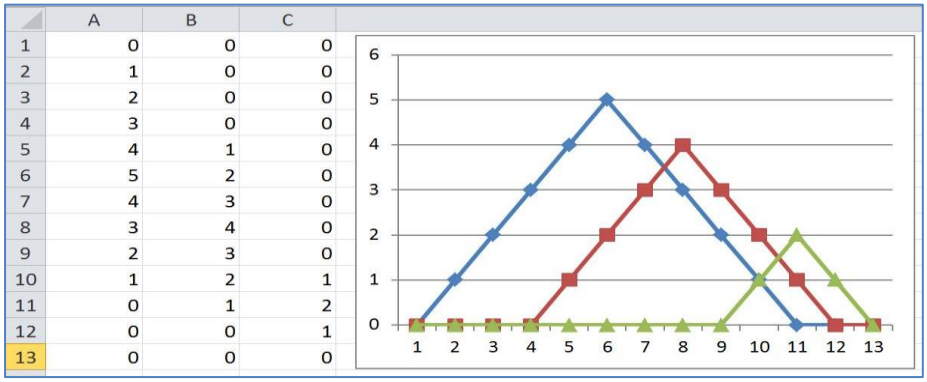

Figure 2 – Processes that match the test example.

Visualization of a table corresponding to the correlation matrix calculated using the above algorithm:

;Node1;Node2;Node3
Node1;0.000;0.818;0.623
Node2;0.818;0.000;0.766
Node3;0.623;0.766;0.000

it has the form shown in Fig. 3.

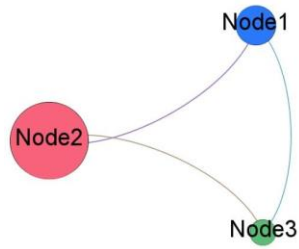

Figure 3 – The correlation network of the example calculated by the algorithm from [2].

In this matrix, node 2 is represented by the largest circle, although it is obvious that the process corresponding to node 1 started earlier and has a larger amplitude.

To correct this discrepancy, the presented improved algorithm allows us to obtain the following matrix of node relationships, the visualization of which is shown in Figure 4:

;Node1;Node2;Node3
Node1;0.000;1.022;0.779
Node2;0.611;0.000;0.613
Node3;0.002;0.050;0.000

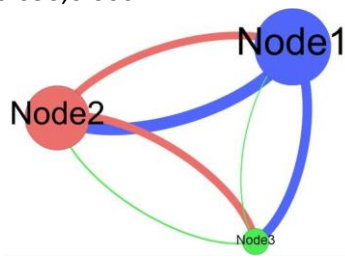

Figure 4 – Directed weighted correlation network of the example calculated using an improved algorithm.

Concept network based on Google Books Ngram Viewer

To build a network of concepts related to modern trends in Computer Science, data obtained by accessing the Google Books Ngram Viewer service was considered as an information source. As an example, we consider 20 concepts listed in Table 1. It also defines the period of analysis (1995-2019 years).

Table 1. Entities-requests to the Google Books Ngram Viewer service

N Entities	Entity	Abbreviation
1	Big data	BDT
2	Complex networks	CNT
3	Machine learning	MNL
4	Deep learning	DPL
5	Neural networks	NNT
6	Data mining	DTM
7	Semantic web	SWB
8	Pattern Recognition	PTR
9	Complex systems	CST
10	Artificial intelligence	ARI
11	Smart grids	SMG
12	Social computing	SCC
13	Natural language processing	NLP
14	Informetrics	INM
15	Social network analysis	SNA
16	Information retrieval	INR
17	Information extraction	INE
18	Computer vision	CMV
19	Digital libraries	DLB
20	Recommender Systems	RSS

Based on the Google Books Ngram Viewer service, dynamic vectors corresponding to specified concepts are defined, represented in JSON format in the source code of the output form, for example:

```
ngrams.data = [{"timeseries": [9.575330750521971e-09, 6.9888743681190135e-09,
1.2791656622823666e-08, 1.1356319440380958e-08, 1.2087312484254653e-08,
1.0711201703372808e-08, 1.2456910170044466e-08, 3.029423822908939e-08, 1.1776428721077536e-
08, 7.028031934197543e-09, 8.192412970231544e-09, 6.390932227873236e-09,
8.478197699446355e-09, 7.651633993077667e-09, 6.700836774342633e-09, 5.973650463175725e-
09, 4.666732333902246e-09, 1.3329707115872225e-08, 8.425238284814895e-09,
1.3226189032877755e-08, 1.0391362437189855e-08, 4.0528647105020355e-08, 2.276777522070006e-
08, 1.8817276625782142e-08, 1.8356137942987516e-08], "parent": "", "ngram": "Informetrics",
"type": "NGRAM"}, {"timeseries": [6.626123649766669e-08, 4.4549572919549973e-08,
4.416813581542556e-08, 3.641325463377143e-08, 3.4242845003973343e-08, 3.139854243272566e-
08, 2.8454564926505554e-08, 2.8115911376858094e-08, 3.27441078695756e-08,
3.3102121932415685e-08, 3.2795835380738936e-08, 3.006990922926889e-08, 2.393094433728038e-
08, 2.0940479572573167e-08, 2.411564814508438e-08, 2.3531855575242844e-08,
3.441038387563822e-08, 5.1720220994866395e-08, 3.282066529664007e-08, 2.867479231838388e-
08, 3.1084773866041366-08, 3.3960155576551196-08, 4.183678115055045e-08,
4.8998536783528834e-08, 7.350192987587434e-08], "parent": "", "ngram": "Computer vision",
"type": "NGRAM"}, {"timeseries": [2.1733773891696728e-08, 2.3040740870783338e-08,
2.6175097644909336-08, 2.3740966526020202e-08, 1.9674967433047646-08, 1.9771428938497593e-
08, 1.961558915297701e-08, 1.8487837039060656-08, 1.9918157789788893e-08,
2.0516173648843505e-08, 1.9428796349529875e-08, 2.2530743493121008e-08,
1.8192798378890984e-08, 1.2451295106075122e-08, 1.1564347701664701e-08,
1.6192725382779827e-08, 1.9444717835881426-08, 1.9834835995879985e-08, 1.9200374623551397e-
08, 2.3690152062272318e-08, 2.5081311250119143e-08, 2.884201322174585e-08,
3.207855669984383e-08, 4.345080029111159e-08, 6.216553316562567e-08], "parent": "",
"ngram": "Natural language processing", "type": "NGRAM"}];
```


As a result of the analysis of 20 concepts of existence, a corresponding weighted correlation matrix was obtained, a network was formed and its clustering was carried out (Fig. 5).

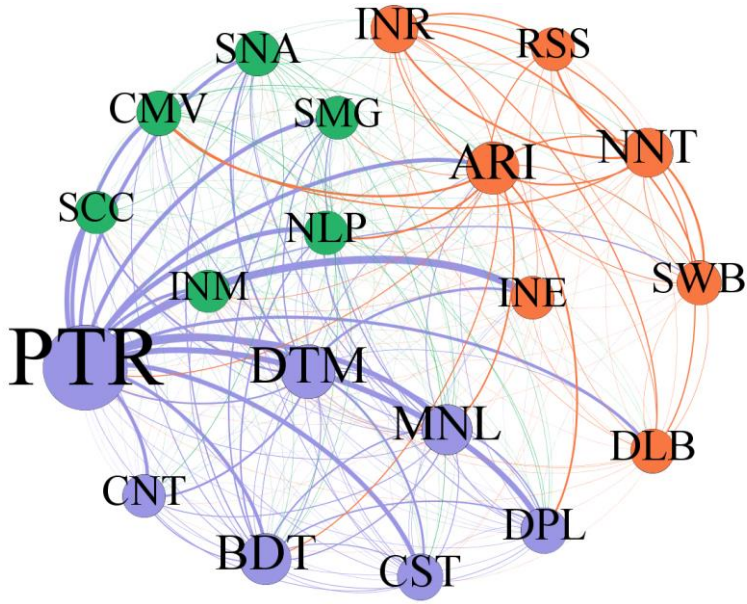

Figure 5 – Network of entities (concepts) in the Gephi system environment.

Figure 6-8 shows typical dynamics corresponding to the concepts included in the various clusters shown in Figure 5.

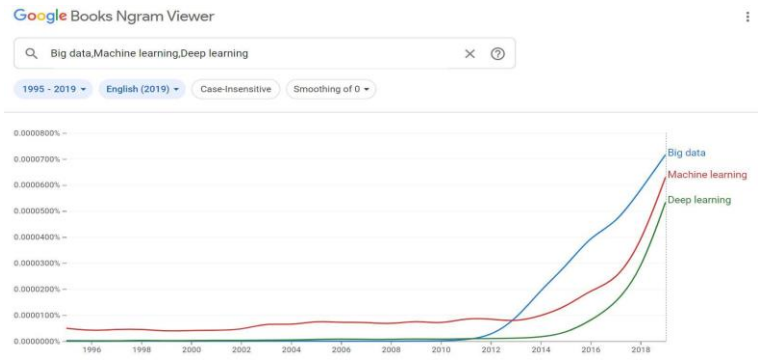

Figure 6 – Entity dynamics (the cluster Big data, Machine learning, Deep learning).

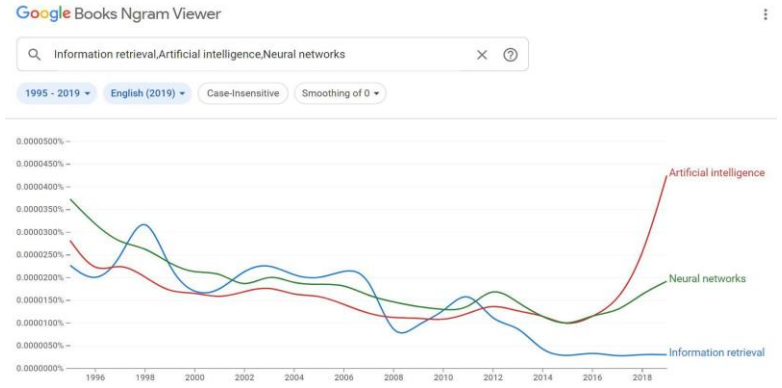

Figure 7 – Entity dynamics (the cluster Artificial intelligence, Neural networks, Information retrieval).

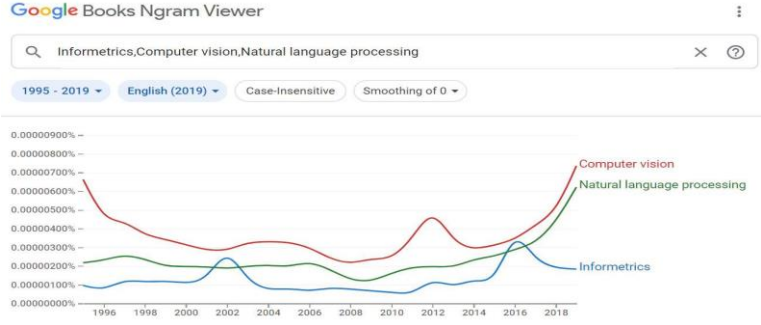

Figure 8 - Entity dynamics (the cluster Computer Vision, Natural language processing, Informetrics).

Examples of entities of other types that you can use the developed method for:

1. Political leaders are characterized by their attitude to various spheres of public life.

2. Consumers of products - here are options sellers, the sources of products [1].

3. Mass media as content entities, in this case, the parameters can be words in the headings of articles that are printed in these publications.

Conclusions

The article proposes the concept of a directed correlation network determined by the dynamics of its appearance in publications and describes the methodology for its formation, clustering and visualization.

The presented approach, in contrast to the existing ones, has the following advantages:

- intuitive, close-to-reality rules for determining the weight of nodes and links;
- the reliable mathematical basis for correlation analysis;
- accounting for previously unused parameters, time series of publication dynamics that correspond to specific features (topics) and allows you to group entities by their development trends over time;
- objectivity – the dataset is responsible for the "purity" of data;
- the relative ease of implementation (ready-made software systems such as Gephi, the R language, etc. can be used).

The method is demonstrated using data obtained from the Google Books Ngram Viewer system. At the same time, it can be used on other data, for example, obtained from a content monitoring system [2], used in analytical systems for various purposes (for example, medical [3, 5]) in order to generalize a set of variables without explicit links between them.

References

1. Using Data Science to Transform Information into Insight *Data Smart.* / *John W. Foreman.* – Wiley, 2013.
2. *Lande D. V., Snarskii A. O.* Networks determined by the dynamics of thematic information streams // Data Recording, Storage Processing, 2020. Vol. 22. ISS. 1. pp. 56-61.

3. *Dmitry Lande, Leonard Strashnoy.* Cross-Correlation of Publications Dynamics and Pandemic Statistics. Available at SSRN: https://ssrn.com/abstract=3625725 or DOI: https://dx.doi.org/10.2139/ssrn.3625725 (June 12, 2020). - 9 p.
4. *Ken Cherven.* Mastering Gephi Network Visualization. – Packt Publishing, 2015.
5. *Dmitry Lande, Leonard Strashnoy*. Directed Correlation Networks, Determined by the Dynamics of COVID-19 Distribution in Various Countries. Available at SSRN: http://ssrn.com/abstract=3674041, DOI: https://dx.doi.org/10.2139/ssrn.3674041 (28 Aug 2020). - 7 p.