ITHEA

International Journal INFORMATION THEORIES & APPLICATIONS

2022 Volume 29 Number 4

International Journal

INFORMATION THEORIES & APPLICATIONS
Volume 29 / 2022, Number 4

Special issue dedicated to memory of
Vitalii Yurievich Velychko,
1962 - 2022

**International Journal “INFORMATION THEORIES & APPLICATIONS” (IJ ITA)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society**

IJ ITA welcomes scientific papers connected with any information theory or its application. IJ ITA rules for preparing the manuscripts are compulsory. The **rules for the papers** for IJ ITA are given on *www.ithea.org*.

Responsibility for papers *published in* IJ ITA belongs to authors.

**International Journal “INFORMATION THEORIES & APPLICATIONS”
Vol. 29, Number 4, 2022**

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with: University of Telecommunications and Posts, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politécnica de Madrid, Spain,
Hasselt University, Belgium, University of Perugia, Italy,
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia

**Printed in Bulgaria
Publisher ITHEA®**

Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: office@ithea.org
Technical editors: **Ina Markova, Vera Markova**

ISSN 1310-0513 (printed) **ISSN 1313-0463 (online)**

IN MEMORIAM

Vitalii Yurievich Velychko,

1962 - 2022

Vitaly Yurievich Velichko has left us - a scientist with extremely high qualifications and the ability to apply his achievements in practice.

In 2004, Vitaly Velichko received a doctorate in Automated Management Systems and Modern Information Technologies, and in 2021 he became a Doctor of Technical Sciences with a specialty in Information Technologies.

In 2007, he became an associate professor of economic-mathematical methods, statistics and economic informatics.

He worked as an associate professor at the V. M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine. He was also a visiting lecturer at the Department of Software Engineering at the National Aviation University, Kyiv, Ukraine.

Vitaly Yurievch was among the most active members of the editorial boards of ITHEA's international journals: "Information theories and applications"; "Information models and analyses"; "Information technologies and knowledge"; "Content and processing of information".

He is the author and co-author of more than 120 scientific articles, reflected in:

https://www.scopus.com/authid/detail.uri?uthorId=56110673500

https://orcid.org/0000-0002-7155-9202

https://scholar.google.com.ua/citations?user=oUHKC8cAAAAJ&hl=uk

He is a co-founder of the ITHEA International Scientific Society (ITHEA ISS) and the Association of Developers and Users of Intelligent Systems, Ukraine, of which he was a director.

His main research interests were in the areas of data mining and knowledge discovery, knowledge representation and management, ontology engineering, logical inference, natural language text processing, neural networks and growth networks.

Professional Java, Delpi, c++ developer.

He participated in a number of TEMPUS PROJECTS.

He was an invited lecturer at ISSI (International Summer School of Informatics) Varna (Bulgaria), as well as at ITHEA conferences.

He has made a significant contribution as a member of steering and program committees of a number of international conferences, among which: KDS - Knowledge - Dialogue - Decision; III – Information – Interaction – Intelligence; GIT - General Information Theory and many other ITHEA ISS committees.

We will remember Vitaly as a wonderful friend and dedicated scientist.

This Special issue of IJ ITA is dedicated to memory of Vitalii Yurievich Velychko and contains three papers coauthored by him.

ITHEA

SELFSTRUCTURIZED SYSTEMS[1]

Victor Gladun, Vitalii Velychko, Yurii Ivaskiv

Abstract: *The problems of constructing the selfsrtucturized systems of memory of intelligence information processing tools, allowing formation of associative links in the memory, hierarchical organization and classification, generating concepts in the process of the information input, are discussed. The principles and methods for realization of selfstructurized systems on basis of hierarchic network structures of some special class – growing pyramidal network are studied. The algorithms for building, learning and recognition on basis of such type network structures are proposed. The examples of practical application are demonstrated.*

Keywords: *knowledge discovery, classification, prediction, growing pyramidal networks, concept formation.*

ACM Classification Keywords: *I.2.4 Knowledge Representation Formalisms and Methods - Semantic networks, F.1.1 Models of Computation - Self-modifying machines (e.g., neural networks)*

Introduction

The task of constructing the self structurized systems is considered in context with intellectualization of the information processing tools. Selfstructurization provides a possibility of changing the structure of data, stored in memory, in the process of the tools functioning as a result of interaction between the received and already stored information. Systems in which the perception of new

[1] Reprinted from: International Journal "INFORMATION THEORIES & APPLICATIONS" Vol.15, Number 1, ITHEA, 2008. ISSN 1310-0513 (printed), ISSN 1313-0463 (online). pp. 05-13

information is accompanied by simultaneous structurization of the information stored in memory, we shall name hereinafter selfstructurized.

Development of principles and methods of constructing the selfstructurized systems in many respects defines a possibility of intellectualization of the information processing tools. Adaptability to the task, being solved, has to do with changing the structure of data. As a result, the possibility of searching in the memory focused on storage of complex data of the large volume occurs, which allows increasing productivity of the used tools, raise accuracy and reliability of received results.

Main processes of structurization of the perceived information consist in formation of semantic and syntactic links among objects by separation of crossings of their attributive representations, as well as the generalized logic attributive models of classes of objects - concepts. As a result of these processes realization, the semantic and syntactic similarity of the perceived information with the stored information is established. Detected associations are fixed as structural changes in memory.

Following basic requirements to data structures in the intellectual systems are set for the decision of such tasks as regularity discovery, classification, forecasting, diagnostics [1]:

The structure of the data should be the multiple-parameter model, reflecting significant properties of researched object. It should provide a possibility to account for the simultaneous influence on researched factor of various combinations of known properties of the researched object.

The model of the researched object should minimize scanning of large-scale data: along with growth of data size, the time of performing of the choice operations grows. It interferes with application of some analysis methods. The model also should be applicable for verification and interpretation.

It should be noted, that in solving tasks of diagnostics and forecasting the models characterized by higher level of generalization of models of classes of objects have advantage. The logic expressions describing such models turn out easier if the complexity is evaluated by number of variables. Simplification of logic expressions results in simple structure of memory and, therefore, simplifies the process of structurization.

In knowledge representation in intelligent systems, those network structures have advantages, which have some information units in vertices, and arches describing links among them. In similar systems, the elements of knowledge representation are combined in the hierarchical structure, realizing such functions, as formation of links among attributive presentations of researched object by allocation of their crossings, hierarchical ordering, classification, concepts formation. In selfstructurized systems, such functions should be performed in the process of the information perceiving.

Condition of an element formation of network structure, for example, unit or link between units, is some relation between determined structural elements of a network. The relations determining formation of structure elements of selfstrtucturized systems we call structurized.

There are two basic ways of objects representation in the information processing systems: by name (condensed) or by sets of attribute values (expanded). The memory structures in selfstrtucturized systems and the appropriate network structures should provide bidirectional conversion between such representations.

Building of selfstrtucturized systems is proposed to be realized on basis of network with hierarchical structures, named as growing pyramidal networks (GPN) [5].

The theory as well as practical application of GPN is expounded in a number of publications [3-6]. GPN realization has following stages:

- to construct the structure of a network for some initial set of objects, assigned by attributive descriptions,
- to train the structure, with a purpose to allocate its elements, allowing to classify all objects of the initial set,
- to recognize belonging to some class of objects of certain object, which is not belonging to initial set of objects.

The mechanisms, providing conversion between converged representation of objects and representation as a set of attributes values in human neurosystem, are discussed in the article [2]. The present work illustrates recent versions of algorithms for building and training GPN, as well as examples of their application.

Building of GPN

A *growing pyramidal network* is an acyclic oriented graph having no vertexes with a single incoming arc. Examples of the pyramidal networks are shown in Figs.1,2,3. Vertices having no incoming arcs are referred to as *receptors*. Other vertices are called *conceptors*. The subgraph of the pyramidal network that contains vertex *a* and all the vertices from which there are paths to vertex *a* is called the *pyramid* of vertex *a*. The set of vertices contained in the pyramid of vertex *a* is referred to as the *subset* of vertex *a*. The set of vertices reachable by paths from vertex *a* is called the *superset* of vertex *a*. The set of vertex, having paths from vertex *a*, is referred to its *superset.*

In *subset* and *superset* of the vertex, 0-*subset* and 0-*superset* are allocated, consisting of those vertices, which are connected to it directly. When the network is building, the input information is represented by sets of attributes values describing some objects (materials, states of the equipment, a situation, illness etc.). Receptors correspond to values of attributes. In various tasks, they can be represented by names of properties, relations, states, actions, objects or classes of objects. Conceptors correspond to descriptions of objects in general and to crossings of descriptions and represent GPN vertices.

Initially the network consists only of receptors. Conceptors are formed as a result of algorithm of construction of a network. After input of object attribute description, corresponding receptors switch to a *state of excitation*. The process of excitation propagates through the network. A conceptor switches into the state of excitation if all vertices of its 0-subset are excited. Receptors and conceptors retain their state of excitation during all operations of network building.

Let F_a be the subset of excited vertices of the 0-subset of vertex *a*; *G* be the set of excited vertices in the network that do not have other excited vertices in their supersets. New vertices are added to the network by the following two rules:

Rule A1. If vertex *a*, that is a conceptor, is not excited and the power of set F_a exceeds 1, then the arcs joining vertices of set F_a with the vertex *a* are liquidated and a new conceptor is added to the network which is joined with vertices of set F_a by incoming arcs and with the vertex *a* by an outgoing arc.

The new vertex is in the state of excitation. Rule A1 is illustrated in Fig.1 (a,b). According to the Rule A1, the condition for adding a new vertex to the network is a situation, when certain network vertex is not completely excited (at least two vertices of 0-subset are excited). Fig. 1.a shows a fragment of network in some initial state. Receptors 4,5 switch to a state of excitation, the network switches to state Fig. 1.b, a new vertex appears – a new conceptor. Receptors 2,3 switch to a state of excitation additionally. The network switches to state Fig. 1.c.

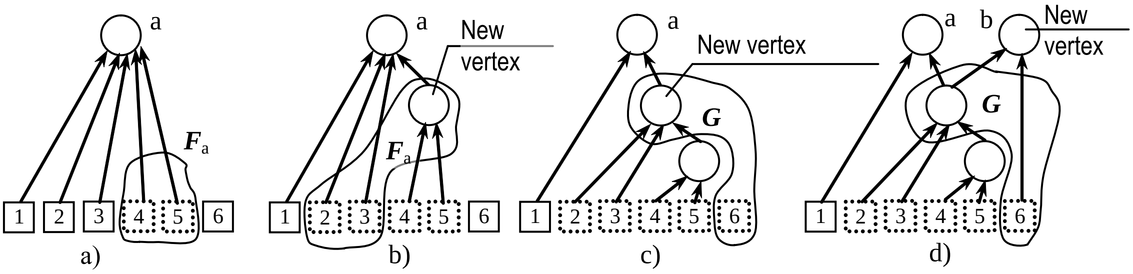


Fig. 1.

New vertices are inserted in 0-subset of vertices, which are not completely excited. New vertices correspond to intersection of object descriptions, represented by incoming arches. Once new vertices have been introduced into all network sections where the condition of rule A1 is satisfied, rule A2 is applied to the obtained network fragment, concluding the object pyramid building.

Rule A2. If the power of set *G* exceeds 1 element, a new conceptor is added to the network, which is joined with all vertices of set *G* by incoming arcs.

The new vertex is in the state of excitation. Rule A2 is illustrated in Fig.1c,d. Network Fig1.d was obtained after the excitation of receptors 2-6.

In applying the Rule A1 the main cross-linking relation is a relation of intersection of receptor set, excited by input of the object description and other sets of receptors included into pyramid, recently formed by conceptors. Rule A2 concludes the building of pyramid, which represents complete description of the introduced object.

Pyramidal networks are convenient for execution of various operations of associative search.

For example, it is possible to select all the objects that contain a given combination of attribute values by tracing the paths that outgo from the network vertex corresponding to this combination. To select all the objects whose descriptions intersect with the description of a given object it is necessary to trace the paths that outgo from vertices of its pyramid. Rules A1, A2 establishes associative proximity between objects having common combinations of attribute values.

Hierarchical organization is an important property of pyramidal networks. This provides a natural way for reflecting the structure of complex objects and generic-species interconnections.

Conceptors of the network correspond to combinations of attribute values that define separate objects and conjunctive classes of objects. By introducing the excited vertices into the object pyramid, the object is referred to classes, which descriptions are represented by these vertices. Thus, during the network building the conjunctive classes of objects are formed, the classification of objects is performed without a teacher. Classifying properties of pyramid network are vital for modeling environments and situations.

Conversion from converged representation of objects (conceptors) to expanded (sets of receptors) is performed by scanning pyramids in top-down and down-top directions.

Training GPN

Training GPN consists in formation of the structures representing concepts, on a basis of attributive descriptions of the objects incorporated into classes with known properties.

Concept is an element of knowledge system, representing generalized logic attributive model of a class of objects, by which processes of recognition of objects are realized. The set of objects generalized in concept is its *volume*.

Consider a task of inductive formation of concepts for not intersected sets of objects $V_1, V_2, ..., V_n$, each set represents some class of objects with known properties. Let L - be a set of objects used as training set. All the objects of set L are represented by sets of attribute values. Relations $L \cap V_i \neq \varnothing$ and $V_i \not\subset L(i = 1,2,...,n)$ are set. Each object from set L corresponds to one set V_i. It is necessary to generate *n* concepts by analysis *L*. The amount of these concepts must be sufficient for correct recognition of belongings of anyone $l \in L$ to one of sets V_i.

Each concept, generated on the basis of training set, is approximation to real concept, the proximity of concepts depends on representativeness of training set, i.e. on the detalization of peculiarities of the concept volume.

In forming the concept corresponding to set V_i , the objects of training set included in V_i**,** are considered as examples of set V_i , and the objects, not included in V_i, - as counterexamples of set V_i**.**

The combinations of attributes allocated in building of a pyramidal network, representing descriptions of objects of training set, are used as "a building material", a basis of further logic structure of concept.

Let *L* be the pyramidal network representing all of training set objects. For formation of concepts $A_1, A_2, ..., A_n$, corresponding to sets $V_1, V_2, ..., V_n$, pyramids of all objects of training set are scanned in order. The vertices of scanned pyramid during its scanning are considered excited. Special vertices in network are identified in order to recognize objects from the concept volume. They are referred to as *check vertices* of a certain concept. In selecting the check vertexes, two characteristics of network vertexes are used: $\{m_1, m_2, ..., m_n\}$, where $m_i \ (i = 1,2,...,n)$ is a number of objects of volume of concept A_i , which pyramids include the given vertex; and *k*, which is the number of receptors in the pyramid of this vertex. For receptors *k*=1. While scanning, the pyramid is transformed by the following rules.

Rule B1. If in the pyramid of an object from concept volume A_i, the vertex, having the largest *k* among all the vertices with the largest m_i, is not a check vertex of concept A_i, then it is marked as a check vertex of the concept A_i.

The rule allows existence several vertexes among the excited vertexes with identical m_i, exceeding m_i of other excited vertexes. If in group of the vertexes having largest m_i, values *k* of all vertexes are equal, any of vertexes can be marked as check vertex of concept A_i.

The rule B1 is illustrated in Fig. 2. In a situation demonstrated by Fig. 2, in excitation in pyramid of vertex 2 vertex 6 is selected as check vertex as having the largest *k* among vertices with the largest m_i (6, 13, 14). Values m_i are shown inside symbols of vertices.

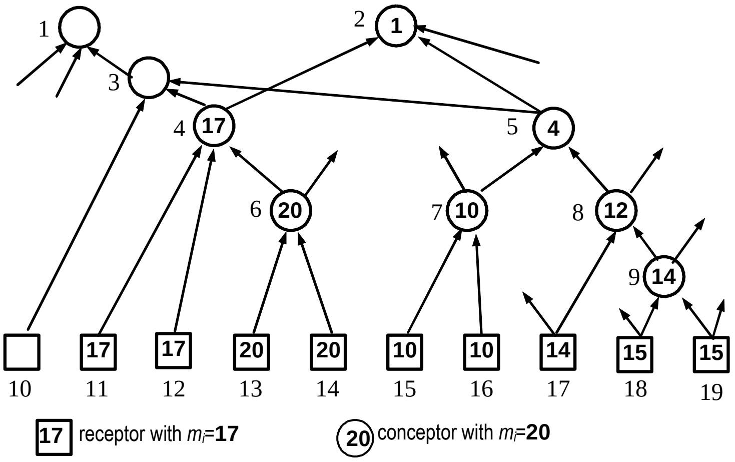


Fig. 2.

Rule B2. If the pyramid of an object from concept volume A_i contains check vertices of other concepts whose supersets do not contain excited check vertices of concept A_i, then in each of these supersets the vertex, having the largest *k* among all excited vertices with the largest m_i, is marked as a check vertex of concept $A_{i.}$

According to this rule the excitation of the pyramid of vertex 2 (Fig.3.a) on the condition, that it represents an objects from concept volume A_i, results in choosing vertex 5 as the check vertex of concept A_i (Fig. 3.b).

By check vertexes we select the most typical (having the largest m_i) combinations of attribute values, belonging to objects from concept volume. For example, selecting the vertex 8 (Fig 3a.) as a check vertex means selection of combination of value attributes, corresponding to receptors 17,18,19.

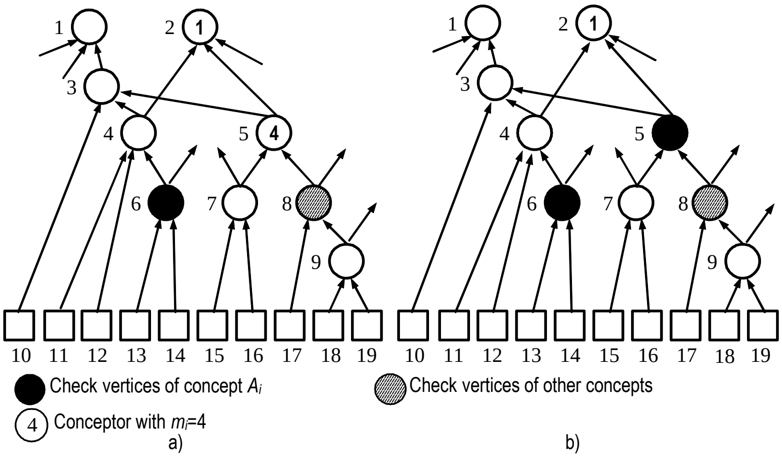


Fig. 3.

If at least one new check vertex appears while scanning objects of the training set, i.e. conditions of Rules B1 or B2 have been performed once at least, the training set is rescanned. The algorithm stops if during the scanning of the training set no new check vertex appears.

Recognition on basis of GPN

The task of recognition is based on the following rule.

Certain object belongs to the concept volume A_i if its pyramid has check vertexes A_i and does not contain check vertices of any other concept not having excited check vertices of concept A_i concept in their supersets. If this condition does not hold for any of the concepts, the object is referred to as unrecognized.

The execution time of the above algorithm is always finite. If the volumes of the formed concepts $V_1,V_2...,V_i,...,V_n$ do not intersect, than after execution the algorithm the recognition rule completely divides the training set into subsets $L_i = V_i \cap L(i=1..n)$.

The formed concepts are represented in the network as ensembles of check vertexes.

There is an algorithm of composing the logic descriptions of concepts, formed in the network as a result of the training process, described above. The formed logical expression contains logical relations, represented by allocation of check vertexes, describing the concepts in the network, defining different classes of objects.

The analytical tasks, such as diagnostics or prognosis, can be reduced to the task of classification, i.e. to belongings the research object to a class of objects, with a property characteristic or a set of properties significant for diagnostics of prognosis.

GPN Application

The following example illustrates the result of concepts formation on the basis of the analysis of a fragment of training set shown in the table 1.

The table has descriptions of ceramic materials of two classes with the following attributes: M - material, T - fineness of powder, C - mix proportion, PP – powder manufacturing method, GP - conditions of obtaining the sample at hot pressing, NoGP - conditions of obtaining the sample without hot pressing, DU - special conditions of manufacturing of a sample, Por - porosity, Z - granularity.

Letters and figures in sections specify values of the appropriate attributes.

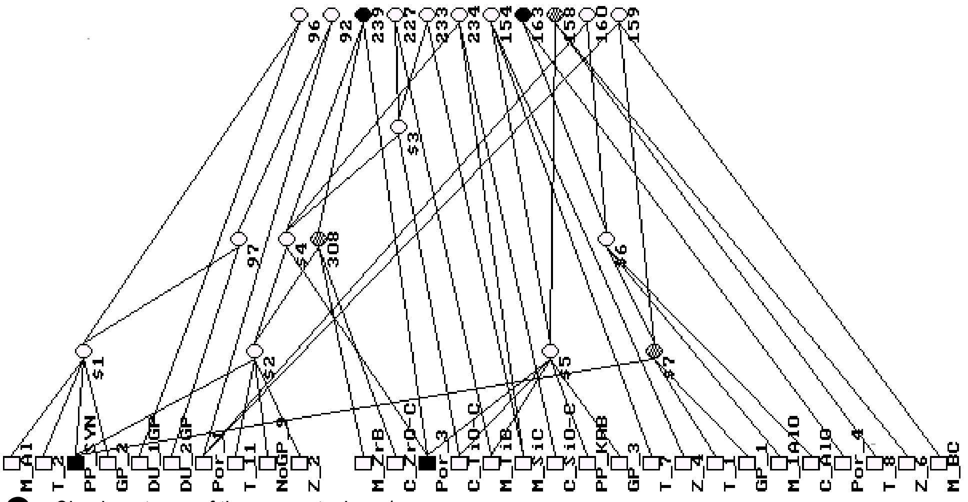
Fig. 4 demonstrates the appropriate pyramidal network with the formed concepts. Check vertices PP_SYN, Por_3, 239, 163 characterize class 1, check vertexes 158, 308 and $7 characterize class 2.

The class 1 is described by the following logical expression, where ∨, ∧, ¬ - logical operations of a disjunction, conjunction and negation:

PP_SYN ∧ ¬{T_1 ∧ GP_1} ∧ ¬{M_ZrB ∧ C_ZrO-C ∧ T_11 ∧ NoGP_9 ∧ Z_2} ∨
Por_3 ∧ ¬{T_8 ∧ Z_6 ∧ M_TiB ∧ C_TiO-C ∧ PP_KRB ∧ GP_3} ∨
M_ZrB ∧ C_ZrO-C ∧ T_11 ∧ PP_SYN ∧ NoGP_9 ∧ Por_3 ∧ Z_2 ∨
M_1AlO ∧ T_1 ∧ C_AlO ∧ PP_SYN ∧ GP_1 ∧ Por_4.

Table 1. Training set.

Object	Class	M	T	C	PP	GP	NoGP	DU	Por	Z
97	1	Al	2		SYN	2		2GP		
96	1	Al	2		SYN	2		1GP		
92	1	Al	2		SYN	2		2GP	1	
227	1	TiB	11	TiO-C	SYN		9		3	2
228	1	TiB	11	TiO-C	SYN		9		3	2
229	1	TiB	11	TiO-C	SYN		9		3	2
233	1	SiC	11	TiO-C	SYN		9		3	2
234	1	SiC	11	SiO-C	SYN		9		3	2
235	1	SiC	11	SiO-C	SYN		9		3	2
237	1	SiC	11	SiO-C	SYN		9		3	2
239	1	ZrB	11	ZrO-C	SYN		9		3	2
240	1	ZrB	11	ZrO-C	SYN		9		3	2
241	1	ZrB	11	ZrO-C	SYN		9		3	2
242	1	ZrB	11	ZrO-C	SYN		9		3	2
154	1	TiB	7	TiO-C	KRB	3			3	4
156	1	TiB	7	TiO-C	KRB	3			3	4
163	1	1AlO	1	AlO	SYN	1			4	
158	2	TiB	8	TiO-C	KRB	3			3	6
160	2	1AlO	1	AlO	SYN	1			1	
159	2	BC	1		SYN	1			1	
308	2	ZrB	11	ZrO-C	SYN		9			2



● Check vertexes of the concept, class 1

▨ Check vertexes of the concept, class 2

○ Conceptor, which is not check vertex

Fig 4.

The logic expressions, defining various classes of objects, are united in *cluster databases* (CDB). CDB contain the information on groups of objects (clusters), specific to the area of study. On basis CDB problems of classification, diagnostics and forecasting are solved. After the concept for some class of objects has been formed, problems of forecasting and diagnostics are reduced to a problem of classification. Classification of new objects is performed by comparing the attribute descriptions with the concept, defining a class of predictable or diagnosing objects. Objects can be classified by evaluating the value of the logical expressions that represent corresponding concepts. The variables, corresponding to the attribute values of the recognized object, set 1, other variable set 0. If the entire expression takes the value 1, that means that the object is included into volume of concept.

The next geometric interpretation of the concept formation algorithm can be proposed.

Every network vertex, having *k* receptors in its subset, corresponds to $(s-k)$-dimensional plane in *s*-dimensional attribute space. The plane contains all the points corresponding to objects whose perceiving results in exiting of this vertex. $(s-k)$-dimensional planes corresponding to check vertices of concept A_i are referred to as *zones* of concept A_i.

The following statements are true for growing pyramidal networks.

Statement 1. The zone of any network vertex is totally included in zones of its subset vertices and totally includes all zones of its superset vertices.

Statement 2. The point corresponding to an object in the attribute space is located inside an intersection of zones of those check vertices, which are exited when the object is perceived.

Point *a* corresponding to the object in the attribute space is directly included in the zone *Z* of concept A_i if there is no other zones of this concept which include point *a* and totally are included in zone *Z*.

The geometric interpretation of the above-described rules for concept formation algorithm is as follows.

Rule B1. For every object of concept volume A_i $(s-k)$-dimensional plane of the exited vertex having the highest *k* among all the vertices with the highest m_i becomes the zone of concept A_i.

Rule B2. If the point, corresponding to an object of concept volume A_i in the attribute space, is directly included in zones of the other concepts, then a zone of concept A_i is created inside each of those zones.

The algorithm of concept formation stops, when during regular examination of the training set, points corresponding to objects from any class are not directly included in zones of the other concepts. When learning is finished, an object corresponds to concept volume A_i if the appropriate point in the attribute space is directly included in at least one zone of concept A_i and is not included in any zone of the other concepts.

Zones of concept A_i , directly inclusive points of objects, corresponding to objects from its volume, as well as points, corresponding to objects from different concepts, are referred to as *boundary zones* of concept A_i *.*

Statement 3. According to Rule B2 new zones can be created only directly inside boundary zones.

Formation of new zones inside boundary zones results in division of boundary zones.

Construction of approximating region for concept A_i consists of two processes: rough covering with concept A_i *i* zones the distribution domain of training set objects corresponding to concept A_i (Rule B1); and division of arising boundary zones (Rule B2).

On the basis of geometrical interpretation, algorithm convergence can have the following explanation.

For each concept the total covering by zones of allocation area of the training set objects, which are included in its volume, results in scanning of all objects, i.e. during single scanning of training set. The boundary zones include points of objects of training set, for which conditions of Rule B2 work. Therefore in every scanning of training set the division of all boundary zones, formed by previous scanning, occurs.

Process of division of boundary zones proceeds, as long as boundary zones exist, and can result in allocation of separate points of attribute space as zones. As number of the points corresponding to objects of training set is finite in each boundary zone, the process of division of boundary zones is finite too.

Absence of boundary zones after the termination of process of division means, that each of concepts in attribute space has area containing all points, corresponding objects of training set which are included in concept volume, and not including any point corresponding to other objects of training set. Thus, after the termination of division of boundary zones total division of training set into subsets $H_i = V_i \cap H(i = 1,2,...,n)$ occurs. As a result algorithm operation for each of the formed concepts, the area is composed of zones of attribute space. This area contains all points of objects of the appropriate class and does not contain any point corresponding to objects of other classes. This area approximates allocation area of objects of the corresponding class. As the approximating area consists of linear elementary areas (hyperplanes), its limiting surface is piecewise-linear. Therefore, the algorithm performs the piecewise-linear division of objects, which correspond to different concepts.

The described method provides decisions of analytical problems of classification, diagnostics and forecasting on the basis of logic models of objects classes. The model displays dependences of an investigated class on combinations of values of attributes, i.e. allows taking account for combined influence of several attributes.

An important distinction of a method of concepts formation in growing pyramidal networks is the possibility to introduce in concepts the so-called excluding attributes which do not correspond to objects of a researched class. As a result, the formed concepts have more compact logic structure, which allows increasing the accuracy of diagnosis or forecasting. In logic expression the excluding attributes are presented by variables with negation.

All search operations in growing pyramidal network are limited to rather small fragment of a network, which includes an object pyramid and vertices directly linked to it. As a result, we have a possibility solve practical analytical problems based on large-scale data.

In a pyramidal network the information is stored by its representation in structure of network. Rules A1-A2, B1-B2 define the rules of memory organization while new information perception. The information of objects and classes of objects is presented by ensembles of vertices (pyramids), allocated in all network. Incoming of the new information causes redistribution of links among vertices of network, i.e. modifying of its structure.

The advantages of growing pyramidal networks become obvious in implementation, which allows parallel distribution of signals in network. The important property of a network as means of information storage is that the possibility of parallel distribution of signals is combined with parallel reception of signals to receptors.

Despite of the certain similarity of the processes proceeding in GPN and neural networks there some distinctions in operating. Main distinction of GPN is that its structure is formed depending on the input data automatically. The adaptation of network structure to the structure of data results in optimization of the information representation. In addition, in contrast to neural networks, the adaptation does not require the introduction of aprioristic redundancy of a network, and training process does not depend on the predetermined configuration of a network. The weakness of neural networks comparing with GPN is that the allocated generalized knowledge cannot be explicitly represented as rules or logic expression. It complicates their understanding by person.

Various set-theoretic descriptions of GPN are given in [4,5]. The [5] considers the so-called β–pyramidal networks (β-PN) modification of GPN for the ranked data. β-PN are useful for data presentation in problems of management, taking and planning decisions (for example, in planning the actions of robots), and also in semantic analysis and synthesis of natural-language texts. In [3-5] the algorithm of formation of concepts in GPN for nondetermined learning process, i.e. for a case when crossing volumes of different concepts occurs, is considered.

The program complex used for experimentation and solving the applied tasks using GPN [8], includes systems CONFOR, realizing processes of building and training GPN, and DISCRET by which the attributes given in numerical scales, are transformed in nominal scales. Discretization of attributes is performed on numerical scales by analysis of distributions of training set objects belonging to different classes.

Typical application fields for GPN are as follows: forecasting of new chemical compounds and materials with the predefined properties[7-9], forecasting in genetics, geology, the solar activity forecasting, medical and technical diagnostics, the robot planning, forecasting of failures of complex units etc. As an example we offer tasks of inorganic compounds forecasting with predefined properties. The tables containing attribute descriptions of binary, ternary and quaternary systems of chemical elements, forming or not forming the chemical compounds were used as training set. Training sets for binary, ternary and quaternary systems included 1333, 4278 and 4963 descriptions, and test set - 692, 2156 and 2536 descriptions. Each chemical element was described by the set of 87 attribute values. Descriptions of binary, ternary and quaternary systems had 174, 261 and 348 attributes. The recognition furnished the 99% accuracy result.

Conclusion

The growing pyramidal network is the network memory self-adapting to the structure of incoming information. In selfstructurized systems the structure of data adapts to the task (classes of objects are allocated and defined) which results in optimization of the solution. In contrast to neural networks, the adaptation does not require the introduction of aprioristic redundancy of a network. In GPN various combinations of the assigned initial properties are formed, which increase the accuracy of analytical tasks solving. Selfstructurized systems allow not only to locate the dependences providing the diagnosis or the forecasting but also to create their logic descriptions.

The researches, operating on complex large-scale data, have shown high efficiency in applying the growing pyramidal networks for solving the analytical tasks. Such properties as simplicity of modification, combining the input of information with classification, generalization and allocation of essential attributes, high associativity, all make the growing pyramidal networks an indispensable component of intellectual systems.

Bibliography

1. Pospelov D.A. Logic-linguistic models in control systems. -Moscow: Energoizdat.-1981.
2. Voronkov G.V., Rabinovich Z.L. Natural environment of memory and thinking: modelling representation. Proceedings of international conference. "Knowledge - Dialogue-Solution"-2001.-SPb.-2001.
3. Gladun V.P. Partnership with computer. Man-Computer Purposeful Systems.-Kiev: Port-Royal. - 2000.
4. V.P.Gladun. Processes of New Knowledge Formation. Sofia: SD Pedagog, 1994, 192 p.
5. V.P.Gladun. Planning of Solutions. Kiev: Naukova Dumka. 1987. 168 p.
6. Gladun V.P. and Vashchenko N.D. Analytical processes in pyramidal networks // Intern. Journal on Information Theories and Applications. FOI-COMMERCE, Sofia.-2000.-Vol.7, - №3.
7. Kiselyova N., Gladun V., Vashchenko N. Computational Materials Design Using Artificial Intelligence Methods. Journal of Alloys and Compounds. 279 (1998), pp. 8-13.
8. www.aduis.com.ua <http: // www.aduis.com.ua>
9. Kiseleva N.N. [editor V.S.Zemskov] Computer designing of inorganic compounds: use of databases and methods of artificial intelligence; Institute metallurgy and sience of material named for A.A.Bajkov.-of M.: Nauka.-2005.

Authors' Information

Victor Gladun, Vitalii Velychko *– V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail:* *glad@aduis.kiev.ua*

Yurii Ivaskiv *– National Aviation University, Prospekt Kosmonavta Komarova 1, 03058, Kiev, Ukraine.*

USEFULNESS OF SCIENTIFIC CONTRIBUTIONS[1]

Krassimir Markov, Krassimira Ivanova, Vitalii Velychko

Abstract: *The prevailing role of counting citations over the added scientific value evaluating distorts the scientific society. As result, the scientific work becomes a kind of business, for instance, to obtain as more citations as possible. It is important to counterbalance the role of counting citations by using additional qualitative criteria. The aim of this survey is to discuss an approach based on measure of “usefulness of scientific contribution” called “usc-index” and published in [Markov et al, 2013]. It is grounded on theory of Knowledge Market. In accordance with this, we remember main elements of this theory. After that we recall some information about Bibliometrics, Scientometrics, Informetrics and Webometrics as well as some critical analyses of journals’ metrics and quantity measures. Finally, we outline the approach for evaluation usefulness of the scientific contributions.*

Keywords*: Information Market, Knowledge Market, Usefulness of the Scientific Contributions*

ACM Classification Keywords*: A.1 Introductory and Survey*

Introduction

The main goal of this paper is to continue the investigation of Knowledge Markets started in [Ivanova et al, 2001; Markov et al, 2002; Markov et al., 2006; Ivanova et al, 2006].

[1] Reprinted from: International Journal “INFORMATION THEORIES & APPLICATIONS” Vol. 20, Number 1, ITHEA, 2013. ISSN 1310-0513 (printed), ISSN 1313-0463 (online). pp. 04-38

Now, our attention will be paid to the *Usefulness of the Scientific Contributions (USC)*.

What is “scientific contribution”? May be the most popular understanding is:

(1) The *added scientific value* of the published researcher’s results;

(2) Its impact on obtaining new scientific results registered by corresponded *citations*.

It is very difficult to measure the added scientific value.

Because of this, in recent years, it became very popular to measure the second part – the citations.

There are a number of ways to analyze the impact of publications of a particular researcher. A longtime favorite has been ISI’s (Social) Science Citation Index, which has come to the web as Web of Science. The web has introduced a number of other tools for assessing the impact of a specific researcher or publication. Some of these are Google Scholar, Scopus, SciFinder Scholar, and MathSciNet among many others. In addition, Publish or Perish uses data from Google Scholar, but it automatically does analysis on the citation patterns for specific authors. After searching for an author one can select the papers to analyze and to get metrics such as total citations, cites per year, h-index, g-index, etc. [Peper, 2009]. In the same time, a negative tendency appears.

The prevailing role of counting citations over the added value evaluating distorts the scientific society.

As result, the scientific work becomes *a kind of business*, for instance, to obtain as more citations as possible.

For examples see [Harzing, 2012].

It is important to counterbalance the role of counting citations by using additional qualitative criteria [DORA, 2012; ISE, 2012].

In an early work (1964) Garfield suggested 15 different reasons for why authors cite other publications (reprinted in [Garfield, 1977]). Among these were**:** paying homage to pioneers**;** giving credit for related work**;** identifying methodology**;**

providing background reading**;** correcting a work**;** criticizing previous work**;** substantiating claims**;** alerts to a forthcoming work**;** providing leads to poorly disseminated work**;** authenticating data and classes of fact – physical constants, etc.; identifying original publications in which an idea or concept was discussed**;** identifying original publication or other work describing an eponymic concept**;** disclaiming works of others and disputing priority claims.

Similarly, the textual function of citations may be very different. In a scientific article some of the references will represent works that are crucial or significant antecedents to the present work; others may represent more general background literature. For example, reviewing the literature published on this topic during 1965–1980, Henry Small identified *five distinctions*: a cited work may be

1) *Refuted;*
2) *Noted only;*
3) *Reviewed;*
4) *Applied;*
5) *Supported by the citing work*.

These categories were respectively characterized as [Small, 1982]:

1) ***Negative***;
2) ***Perfunctory***;
3) ***Compared***;
4) ***Used***;
5) ***Substantiated***.

Thus, the different functions that citations may have in a text are much more complex than merely providing documentation and support for particular statements [Aksnes, 2005].

The aim of this survey is to discuss an approach for evaluating the “usefulness of scientific contribution” called “***usc-methodology***” [Markov et al, 2013]. It is grounded on theory of Knowledge Market. In accordance with this, the next chapter remembers main elements of this theory. After that we recall some information about Bibliometrics, Scientometrics, Informetrics and Webometrics as well as some critical analyses of journals’ metrics and quantity measures.

Finally, we outline the approach for evaluation usefulness of scientific contributions. In more details, the chapters of the paper concern:

- Basic concepts of Knowledge Markets’ Theory;
- Structure of the Knowledge Market;
- Science, Publishing, and Knowledge Market;
- National and International Knowledge Markets;
- Bibliometrics, Scientometrics, Informetrics and Webometrics;
- Citation tracking and Evaluation of Research;
- Journal metrics;
- Quantity measures;
- Disadvantages of journal metrics and quantitative measures;
- Evaluation of Scientific Contributions;

Basic concepts of Knowledge Markets’ Theory

Information society

At the stage of social growth, called “information society”, the information and information activities get decisive value for existence of the separate individuals or social teams. Certainly, at earlier stages of development of mankind, the information had important value too. But never, in all known history, other means for existence have been so dominated by the information means as it is in the information society [Markov et al., 2006].

From the origin, human society has been "information" one, but levels of information service differ in different periods of existence of societies. It is possible to allocate following levels of information society:

- *Primitive* (people having knowledge, letters on stones etc.);
- *Paper based* (books, libraries, post pigeons, usual mail etc.);
- *Technological* (telephone, telegraph, radio, TV, audio- and video-libraries etc.);
- *High-Technological* (computer systems of information service, local information networks etc.);
- *Global* (global systems for information service, opportunity for everybody to use the information service with help of some global network etc.).

The information society does not assume compulsory usage of the information services by a part or all inhabitants of given territory. One very important feature thus is emphasized: for everyone will be necessary diverse and qualitative (from his point of view) information, but also everyone cannot receive all necessary information. *The enterprising experts will accumulate certain kinds of information* and will provide existence through favorable to them information exchange with the members of the society. Thus, in one or other form, they will carry out **payable information service (carrying out information services for some income)** [Ivanova et al, 2001]. This is the background of ***Information Market***.

Knowledge Information Objects

The usual understanding of the verb "**to know**" is: "*to have in the mind as the result of experience or of being informed, or because one has learned*"; "to *have personal experience of something*” etc. The concept "**knowledge**" usually is connected to concepts "understanding" and "familiarity gained by experience; range of information" [Hornby et al, 1987] or "organized body of information" [Hawkins, 1982].

V.P. Gladun correctly remarks that the concept “*knowledge*” does not have common meaning, especially after beginning of it’s using in technical lexicon in 70-ies years of the last century. Usually, when we talk about the human knowledge we envisage all information one has in his mind.

Another understanding sets the “knowledge” against the “data”. We talk about data when we are solving any problem or are making logical inference. Usually the concrete values of given quantities are used both as data and descriptions of objects and interconnections between objects, situations, events, etc.

During decision making or logical inference we operate with data involving some other information like descriptions of the solving methods, rules for inference of corollaries, models of actions from which the decision plan is formed, strategies for creating decision plans, and general characteristics of objects, situations, and events.

In accordance with this understanding, the "knowledge" is information about processes of decision making, logical inference, regularities, etc., which, applied to the data, creates any new information [Gladun, 1994].

The knowledge is a structured or organized body of information models, i.e. the knowledge is information model, which concerns a set of information models and interconnections between them.

Let remember, in general, the information model is a set of reflections, which are structured by Subject and, from his point of view, represents any entity [Markov et al, 2001].

The information objects, which contain information models, are called ***"knowledge information objects"***.

Knowledge Market

The growth of societies shows that the knowledge information objects become important and necessary articles of trade. The open social environment and market attitudes of society lead to arising of *knowledge customers* and *knowledge sellers*, which step-by-step form "Knowledge Markets" [Markov et al, 2002].

As the other markets, the ***Knowledge Market*** *is organized aggregate of participants, who operate following common rules and principles*. The knowledge market structure is formed by a combination of mutually-connected elements with simultaneously shared joint resources.

Staple commodities of knowledge market are knowledge information objects.

The knowledge information bases and tools for processing the knowledge information objects, such as tools for collecting, storing, distributing, etc., form the knowledge market environment. The network information technologies enable to construct uniform *global knowledge market environment*. It is very important, it to be friendly for all knowledge market participants and open for all layers of the population without dependence from a nationality, social status, language of dialogue, place of residing. The decision of this task becomes a crucial step of humanization of all world commonwealths.

In the global information society, on the basis of modern electronics, the construction of the global knowledge market, adapted to the purposes, tasks and individual needs of the knowledge market participants is quite feasible, but the achievement of this purpose is connected to the decision of a number of scientific, organizational and financial problems. For instance, the usual talk is that *at the Knowledge Market one can buy knowledge*. But, from our point of view, *this is not so correct*.

In global information society, the e-commerce becomes fundamental for the Knowledge Market. The advantages of e-commerce are obvious. In the same time there exist many risks for beginners at this kind of market. From this point of view, the society needs to provide many tasks for training the citizens to use properly opportunities of the new environment [Markov, 1999]. Let consider an example.

When an architect develops any constructive plan for future building, he creates a concrete “*information object*”. Of course, he will sell this plan. This is a transaction in area of the *Information Market*.

Another question is: from where does architect have received the skills to prepare such plans? It is easy to answer – he has studied hardly for many years and received knowledge is the base for his business. Textbooks and scientific articles are not concrete information for building concrete house, but they contain the knowledge needed for creating such plans.

The scientific books and papers written by the researchers (lecturers) in the architectural academy are special kind of “information objects” which contain special generalized information models. They are “*knowledge information objects*” which have been sold to students and architects.

Here we have a kind of transactions at the “*Knowledge Market*”.

We have to take into consideration the *difference between responsibility* of architect and lecturer (researcher).

If the building collapses, the first who will be responsible is architect, *but never lecturer!*

In beginning of the XX-th century, the great Bulgarian poet Pencho Slaveikov wrote:

"The speaker doesn't deliver his thought to listener, but his sounds and performances provoke thought of the listener. Between them, a process performs like lighting the candle, where the flame of the first candle is not transmitted to another flame, but only cause it."

If one buys a candle what does he really buy – "wax" or "light" of candle? The light is not for sale in the store… But one really may see the example how the candle works and how it may be used. Based on this he/she may decide whether to buy the candle or not.

We came to the main problem we need to point – t*he authors and publishers are not responsible for what they sold to the customers*. Pros and Cons of (electronic) Publishing are discussed many times (see for instance [NLC, 2004]). From customers' point of view, it is difficult to discover what really we will receive if we will buy one (electronic) publication. The title and announcement of the publications are not their content. The customers could not claim damage if the content is not what it is needed. To regulate this process we need specialized rules and standards for knowledge markets as well as corresponded laws for *authors' and publishers' responsibility*.

The scientific work usually is reported as series of publications in scientific journals. The practice is to delegate social rights to editors and reviewers to evaluate the quality of reported results.

And here we see serious problem – *is their evaluation enough? Of course,* ***it isn't!***

Because of this, counting of citations became important. But, the citations may be of different types including negative ones. We need methodology for evaluating *Usefulness of the Scientific Contributions (USC).*

Structure of the Knowledge Market

The Structure of the Knowledge Market was presented in [Markov et al, 2002]. The updated scheme of the basic structure of Knowledge Market is outlined on Figure 1 below.

Let's remember basic elements of the knowledge market.

Employer (Er) is the initial component of the Knowledge Market whose investments support providing the scientific research. The concept of Employer means men or enterprise, which need to buy manpower for the purposes of the given business. A special case is the government of the state which may be assumed as representative of the society as Employer. In addition, different scientific or not scientific foundations, social organizations, etc., may invest in scientific activities and this way to become Employers.

The concept of the ***Employee (Ee)*** means a man who is already taken as a worker in the given business or is potentially to be taken in it. The main interest of the employee is to sell his received knowledge and skills. The main goal of the Employee is to receive maximal financial or other effects from already received knowledge and skills. This means that the Employee is not internally motivated to extend them if this knowledge and skills are enough for chosen work activity. From other point of view the Employee motivation closely depends to future expectations for his social status. The Employee became as converter of the learned knowledge and skills into real results of his workplace. Let remark, that scientific organizations, institutes, groups, etc. may be employed to fulfill some scientific projects and to be in the role of Employee at the KM.

In other words, *Employer* hires *Employees*. During the work processes, the knowledge and skills of Employees are transformed in real products or services. This process is served by the Manpower Market. Employees, even owning a high education level, need additional knowledge to solve new tasks of the Employers. Still, they are **customers of new knowledge**, who arouse necessity of the Knowledge Market, which should rapidly react to the customers’ requests. In other words, the Manpower's Market causes activity of the Knowledge Market (KM). These two members of KM are main its components – *the knowledge customers*.

It is clear that the business needs the high-skilled workers. The employer buys the final result of the cycle in the Knowledge Market - the educated and skilled workers. The continuous changing of technological and social status of the society leads to appearance of new category – industrial **Researchers (R)** – peoples/organizations, who have two main tasks:

- To invent and/or promote new technologies to Employers in convenient way to implement them in practice;
- To determine the educational methods for training the staff for using the new technologies.

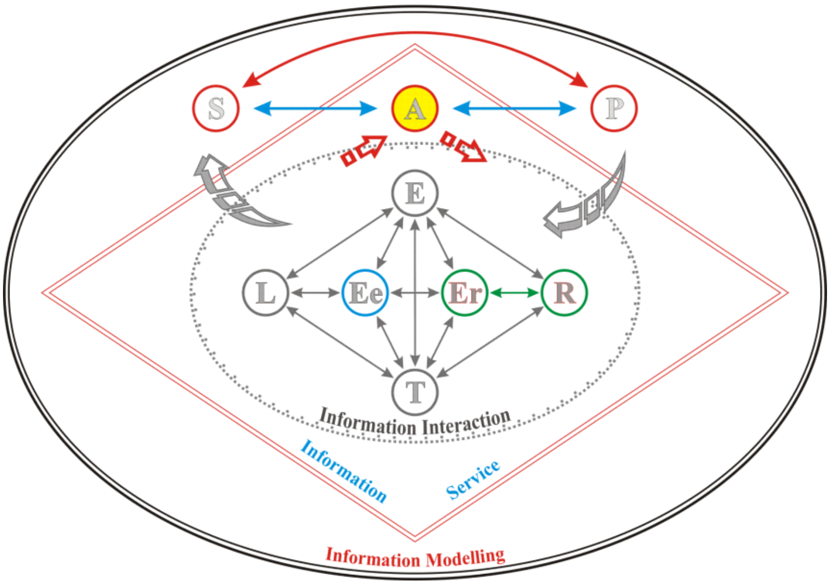


Figure 1. Structure of the Knowledge Market

The educational process is carried out by the **Lecturers (L)**, who transforms new scientific knowledge into pedagogical grounded lessons and exercises. During realizing concrete educational process, Lecturers are assisted by **Tutors (T)** who organize the educational process and supports the Employees to receive the new knowledge and to master theirs skills. At the end of the educational process, a new participant of KM appears – **Examiners (E)** – who test results of education and answer to the question "have the necessary knowledge and skills been received".

These six components of the Knowledge Market, which contact each other via global information network, form the first knowledge market level called

“*information interaction*”. As far as these components are too much and distributed in the world space, the organization and co-ordination of theirs information interaction needs adequate **“*information service*”**. It is provided by a new component called **Administrators (A)**. Usually the Administrators are Internet and/or Intranet providers or organizations. They ***collect, advertize and sell knowledge objects, sometimes without understanding what really they content*.**

The rising activity of knowledge market creates need of developing new general or specific knowledge as well as modern tools for the information service in frame of the global information network. This causes the appearance of high knowledge market level, which allows observing processes, as well as inventing, developing and implementing new knowledge and corresponded systems for information service. This is the **“*information modeling*”** level. It consists of two important components – the academic researchers called here **Scientists (S)** and the **Publishers (P)**. In this paper we will discuss more deeply characteristics and activities of both of them.

Of course, the Knowledge Market as a kind of Market follows rules and laws given by social environment. The interrelation between government, social structures, and Knowledge Market need to be studied in separate investigation. In several papers we have already investigate different problems of the Knowledge Market [Ivanova et al, 2001; Markov et al, 2002; Ivanova et al, 2003; Markov et al, 2003].

For years we have seen that the Knowledge Market is very important for growth of science and in the same time it is important scientific area and need to be investigated.

Science, Publishing, and Knowledge Market

Preparing this survey, we have collected more than hundred definitions of terms “*science*” and “*scientific methodology*”. Analyzing them we chose the one of the Britain's Science Council, which has spent a year working out a new definition of the word “science”. The Science Council is a membership organization that brings together learned societies and professional bodies across science and its applications. It was established under Royal Charter in October 2003 and was registered as a charity with the Charity Commission in September 2009. The principal activity of Science Council is to promote advancement and dissemination of knowledge and education in science, pure and applied, for public benefit [BSC, 2013].

The Science Council definition focuses on the pursuit of knowledge rather than established knowledge. It may be the first "official definition of science" ever published. Here's what they've come up with:

> ***"Science is the pursuit of knowledge and understanding of the natural and social world following a systematic methodology based on evidence"*** [BSC, 2013].

It defines science as a pursuit, an activity, related to the creation of new knowledge, rather than established knowledge itself. Science is seen as a species of research.

Scientific methodology includes the following [BSC, 2013]:

- Objective observation: measurement and data (possibly although not necessarily using mathematics as a tool);
- Evidence;
- Experiment and/or observation as benchmarks for testing hypotheses;
- Induction: reasoning to establish general rules or conclusions drawn from facts or examples;
- Repetition;
- Critical analysis;
- Verification and testing: critical exposure to scrutiny, peer review and assessment.

The last point is closely connected to publishing activities which are the main way to provide critical exposure to scrutiny, peer review and assessment. In addition, previous published research results have to be taken in account and current results have to be compared and evaluated in accordance to them.

Due to very great number of results to be published, *scientific publishing activities became an industrial branch*. Nowadays, the scientific publishing companies (Publishers "**P"** on Figure 1) compete with others at the knowledge markets in two main areas:

- Collecting original scientific results to be published;
- Market shares where the publications may be sold.

The basic difference between knowledge markets and other kinds of markets consists in the following.

To publish the results of their research is an obligation that professional scientists are compelled to fulfill [Merton, 1957b]. New knowledge, updated by researchers, has to be transformed into information made available to the scientific community. Not only do scientists have to make their work available to the public at large, but they in turn are supposed to have access to the work of their peers. Research is carried out in a context of "*exchange*". Even so, the fact that the system of scientific publication has survived in modern science is due, paradoxically, to scientists' desire to protect their intellectual property. New scientific knowledge is a researcher's personal creation, and claim to its discovery can be laid only through publication [Merton, 1957a].

The "reward system", based on the recognition of work, merely underscores the importance of publication: the only way to spread the results of research throughout the world is to have them published. Publication therefore has three objectives: *to spread scientific findings, protect intellectual property and gain fame* [Okubo, 1997].

The academic researchers (Scientists "**S**" on Figure 1) who produce the new knowledge (presented by knowledge objects to be published) are, in the same time, ***main clients***. In other words, *the source and target groups partially coincide* but they are distributed all over the world. Because of this, information about the published results is accumulated by knowledge market organizers (Administrators "**A**" on Figure 1) who, using special kinds of data bases, serve the interactions between scientists and publishers as well as between both of them and the rest participants of the knowledge markets.

Due to serious *competition between publishers*, the administrators play an extra role – to *range* those using different criteria and this way *to control the* *knowledge objects’ flows*. This is a play for billions of Dollars, Euros, etc. Let see an example from our practice.

We were invited to write a chapter in a scientific monograph to be published by a leading scientific publishing company [Markov et al, 2013a]. The book was published and it became as a staple commodity at the knowledge market. Depending of the format, its price varies between $195 and $390 [Naidenova & Ignatov, 2013]. We were glad to understand that our chapter was evaluated as a good one to be included in an encyclopedic four volumes comprehensive collection of research on the latest advancements and developments [Markov et al, 2013b]. Again, depending of format, the price of the collection varies between $2050 and $4100 [AIRM, 2013].

Let see what income will be received if we assume that the editions have only 250 exemplars and if the editions have 1000 exemplars sold.

In the case with 250 exemplars sold, the income is:

- min: 195x250 + 2050x250 = 48750 + 512500 = 561250 USD;
- max: 390x250 + 4100x250 = 97500 + 1025000 = 1122500 USD.

In the case with 1000 exemplars sold, the income is:

- min: 195x1000 + 2050x1000 = 195000 + 2050000 = 2245000 USD;
- max: 390x1000 + 4100x1000 = 390000 + 4100000 = 4490000 USD.

Concluding this hypothetical accounting we may say that expected income may vary between **500 thousands** and **4.5 millions** of Dollars. Because of this, it is very important to be a “leading” publisher who publishes new and useful results which can be sold. *Unfortunately our income from these editions was **0 (zero) cents**.*

National and International Knowledge Markets

One may remark that for our scientific work we had received salaries, society spend resources for supporting our research via buildings, service workers, etc. Yes, it is truth. But let analyze the situation according the scheme on Figure 1. Two variants of knowledge markets are shown on Figure 2 and Figure 3. The first one is “*national*” KM and second – “*international*” KM. Let analyze them step by step.

The ***National knowledge market*** (Figure 2) is included in the clear boundaries and all processes are connected.

1. The society, via government subsidies and/or concrete national projects, provides financial and organizational support of the scientists and their work.
2. The received results are published and indexed again on the base of financial and organizational support of government subsidies and concrete national projects.
3. Selling the results as printed publications and implementations in practical realizations as well as via the tax mechanism, the society receives some income which in some degree covers the initial expenses.

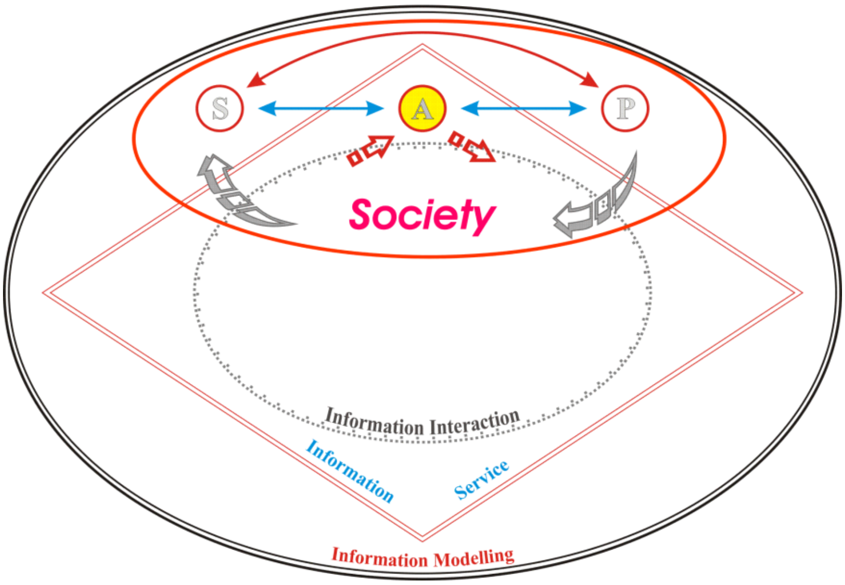


Figure 2. National Knowledge Market

The ***International knowledge market*** (Figure 3) is distributed in the boundaries of separated societies and all processes are financially disconnected.

1. The Society 1, via government subsidies and/or concrete national or international projects, provides financial and organizational support of the scientists and their work.
2. The received results are published in Society 3 and indexed in Society 2 on the base of financial and organizational support of government subsidies and concrete national or international projects.
3. Selling the results as printed publications and implementations in practical realizations as well as via the tax mechanism, the Society 3 receives some income which covers its initial expenses and realizes some profit.
4. Selling informational services based on indexed publications, Society 2 covers its initial expenses and realizes some profit.
5. Only Society 1 has *no profit* but some losses because it spends resources for supporting its scientists but the *surplus value* of their work is accumulated in Society 2 and Society 3.
6. Finally, Society 1 became poor and slowly perishes, but Society 2 and Society 3 became rich and grow.

It is important to comment the role of ***international scientific projects***. They give some financial support to the Society 1 but in the same time they orientate scientists towards interests of sponsoring society, usually it is Society 2 or Society 3, both two societies together or one and the same society which plays both roles. As result, the national knowledge market of Society 1 will be *destroyed* and its rebuilding becomes impossible. In opposite, the national knowledge markets of other societies will grow.

Now the main question is “*How to influence to the Society 1 to participate in such unequal battle?*”

The answer is: *By using the power of*

- *Developed national knowledge markets;*
- *Advertising, mainly indirect.*

The best influence is ***the developed national knowledge market*** with participants who are high level specialists in their area. This generates the willingness to join, to be part of them. As more people are involved so great is the influence to other societies. Opening the national knowledge market is very important step. Possibility to be published on such authoritative level is a possible dream. And the result is total influence. In addition, opening the manpower market for specialists from abroad make this dream reality and many scientists start working following the rules of this national knowledge market to ensure possibility for immigration. Finally, they influence on developing the own national knowledge markets to be organized in the same manner and rules as of the prototype one *without taking in account the national specifics and interests*.

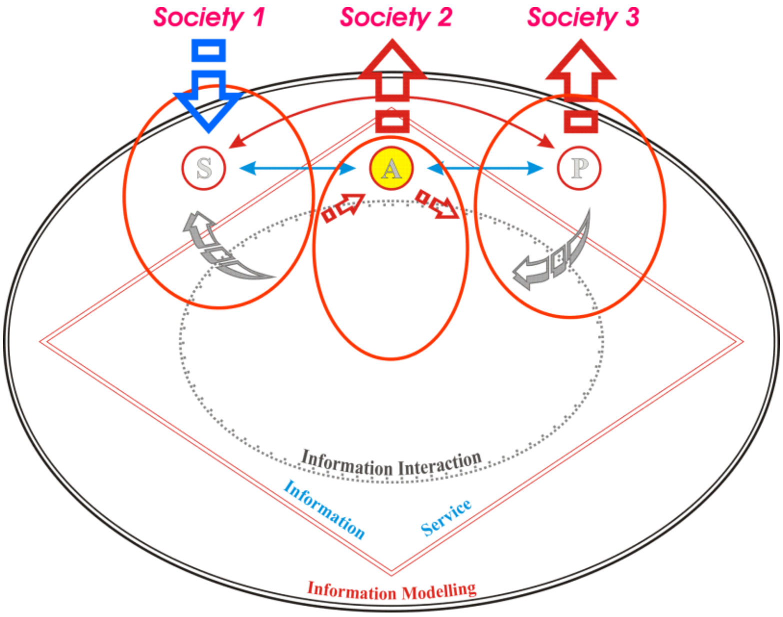


Figure 3. International Knowledge Market

The ***advertizing*** (mainly – indirect) of developed national knowledge markets increase their influence. *Advertising* was originated from a Latin term — “*advertire”*, which means — “***to turn to***”. The American Marketing Association (AMA) has defined *Advertising* as — the placement of announcements and persuasive messages in time or space purchased in any of the mass media by business firms, nonprofit organizations, government agencies, and individuals who seek to inform and/or persuade members of a particular target market or audience about their products, services, organizations, or ideas [AMA, 2013].

Indirect advertising is a form of marketing that does not use the formal everyday methods such as newspapers and magazines. This type of advertising uses: a product in a television show**;** giving a product away for free**;** sponsoring of events or activities (= paying for them)**;** etc. [Jeeves, 2013; CBED, 2013].

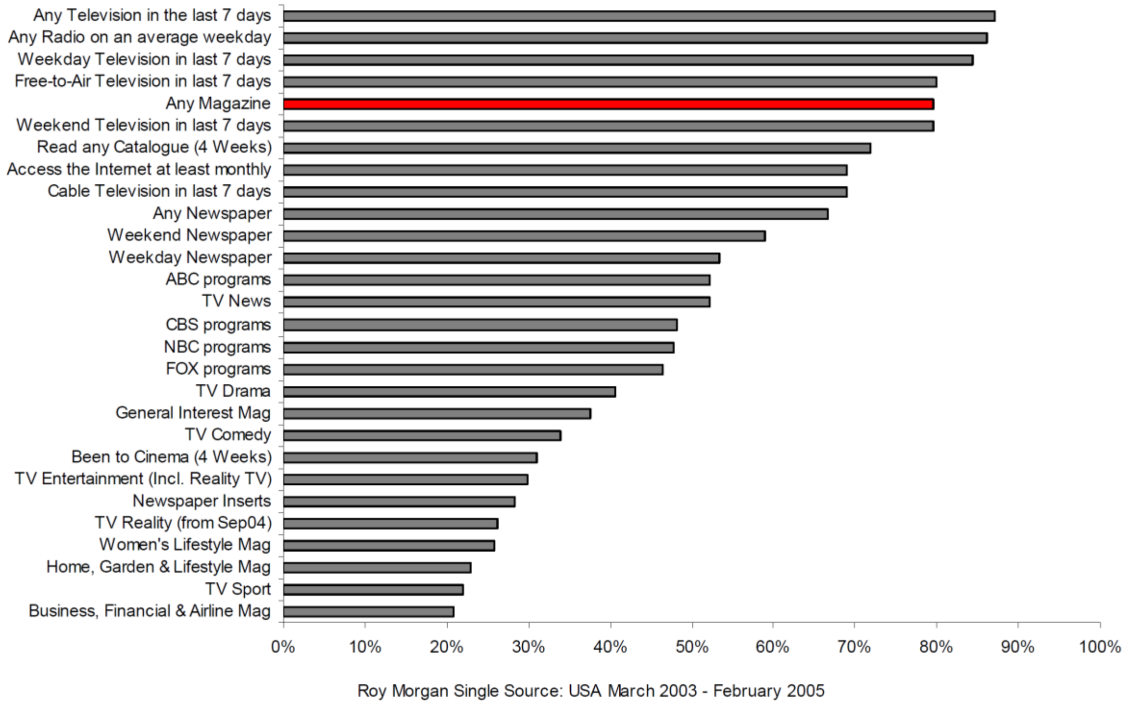
*"**Audience reach measures**" have been used to determine how many people see the advertising and how often.* Measurement systems exist across the globe that determine how many people in total read certain magazines and newspapers, watch TV programs, listen to radio stations, etc.

For instance, in the US, Roy Morgan Single Source shows that, in year 2005, television is still the most widely used medium (see Figure 4). However, magazines, as a group, reach as many people as 'free to air' TV, and more people than newspapers or the Internet. Of course, specific magazines or genres of magazines often outperform specific television 'shows' [Levine et al, 2005].

One of the movements happening on the internet is that of indirect marketing and advertising. Publishers and manufactures are catching on to what customers want, which is proof that they must invest having a business. Indirect advertising and marketing is often a technique to obtain this, as in most circumstances it supplies something of worth upfront *for totally free*. You are going to see this with no cost eBooks, blogs, and videos all dedicated to helping the visitor.

If the content delivers enough enable, the visitor may just check out the rest of the site and sign up for membership region or buy their premium book. Indirect marketing makes use of a funnel pointing toward the location where the business can make money. Another instance is often observed with no cost apps tied to movies. By downloading the app, you might just want to go see or obtain the movie [EzineMark, 2013].

In order to determine how to create an effective advertising campaign decision makers in the industry use a range of ***measures*** to try to predict the outcome of the campaign. Those who make decisions each year about where to place billions of dollars in advertising have focused in the past primarily on audience or “opportunity-to-see” measures – the task being to create chance that target audience will see advertisement with assumption that everything else will run its course.



Roy Morgan Single Source: USA March 2003 - February 2005

Figure 4. Media Usage in USA for year 2005

Bibliometrics, Scientometrics, Informetrics and Webometrics

The advertisers need to know their audience and ***to measure results achieved*** – ***shifts in sales or shifts in attitude*** among the intended audience. Today all marketing and advertising people are judged by the overall performance of their company, each quarter of every year. Research and information is not a substitute for ingenuity. But ignoring intelligent and reliable research and information altogether is a luxury nobody can afford! [Levine et al, 2005]. At the knowledge markets there are two main kinds of indirect advertizing:

- *Ranging selected journals* and this way to raise the income of publishers of these journals and Society 3;
- *Counting citations and computing scientific indexes* based only on digital libraries of collected papers from selected journals and this way to raise income of administrators of these libraries and Society 2.

Measuring science has become an “industry”. Governments and their statistical offices have conducted regular surveys of resources devoted to research and development (R&D) since the 1950s. A new science had raised – *Scientometrics*.

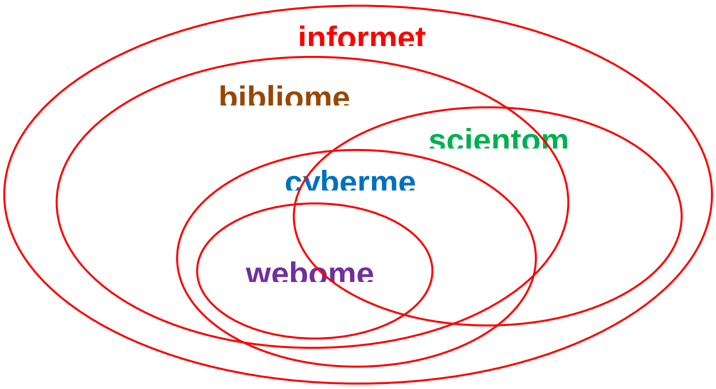
“Scientometrics” is the English translation of the title word of Nalimov’s classic monograph “**Naukometriya**” in 1969, which was relatively unknown to western scholars even after it was translated into English. Without access to the internet and limited distribution, it was rarely cited. However, the term became better known once the journal “Scientometrics” appeared in 1978 [Garfield, 2007] and term has grown in popularity and is used to describe the study of science: growth, structure, interrelationships and productivity [Mooghali et al, 2011].

Scientometrics is related to and has overlapping interests with Bibliometrics and Informetrics. The terms Bibliometrics, Scientometrics, and Informetrics refer to component fields related to the study of the dynamics of disciplines as reflected in the production of their literature [Hood & Wilson, 2001]. A whole community of researchers concerned with counting papers and citations called themselves bibliometricians [Godin, 2005].

Among the many statistical analyses of scientific publications, bibliometrics holds a privileged place for counting scientific papers. Bibliometrics is one of the sub-fields concerned with measuring the output of scientific publications. Bibliometrics owes its systematic development mainly to the works of its founders V.V. Naliv, D.J. D. Price and Eugene Garfield in the 1950s. Since 1958 Bibliometrics has evolved as a field, taught in library and information science schools and it emerged as a tool for scientific evaluation for a number research groups around the world. This process was made possible by the work of Eugene Garfield and his “Science Citation Index”. Castell, an American psychologist, was credited with the launching of Scientometrics, when he produced statistics on a number of scientists and their geographical distribution, and ranked the scientists according to their performance. He introduced two dimensions into the measurements of science, namely, *quantity* and *quality*. The term informetrics was introduced by Blackert, Siegel and Nacke in 1979, but gained popularity by the launch of the international informertics conferences in 1987. A recent development in informetrics called the webometrics/cybermetrics, has become a part of the main stream library and information science research area. The term webometrics refers to the quantitative studies of the nature of scientific communication over the internet and its impact on diffusion of ideas and information. The inter-relations between Infor-, biblio-, sciento-, cyber-, and webometrics are illustrated on Figure 5 [Thelwall, 2006].

Dirk Tunger gave the next definitions [Tunger, 2007]:

- ***Bibliometrics*** is a study or measurement of formal aspects of texts, documents, books and information;
- ***Scientometrics*** analyses the quantitative aspects of the production, dissemination and use of scientific information with the aim of achieving a better understanding of the mechanisms of scientific research as a social activity;
- ***Informetrics*** is a sub-discipline of information sciences and is defined as the application of mathematical methods to the content of information science;
- ***Webometrics*** is the application of informetrical methods to the World Wide Web (WWW).



Firure 5. Infor-, biblio-, sciento-, cyber-, and webometrics**.**

The sizes of the overlapping ellipses are made for sake of clarity only. [Thelwall, 2006]

Citation tracking and Evaluation of Research

Citation tracking is very important. It allows for tracking of authors own influence, and therefore the influence of organization. It allows tracking the development of a technology, which may be the basis for progress undreamt of when a paper is written. Citation tracking provides information on other organizations and authors who are doing similar work, potentially for collaboration, and identifies publications that cover similar topics. Finally, tracking back in time can find the seminal works in a field [Fingerman, 2006].

The use of scientometric indicators in ***research evaluation*** emerged in the 1960s and 1970s, first in the United States and then also in various European countries. Before that time, research evaluation had not been formalized other than through the peer review system, on the one hand, and through economic indicators which could only be used at the macro-level of a national system, on the other.

The ***economic indicators*** (e.g., percentage of GDP spent on R&D) have internationally been developed by the Organization of Economic Co-operation and Development (OECD) in Paris. For example, the Frascati Manual for the Measurement of Scientific and Technical Activities form 1963 (or its new edition [Frascati Manual, 2002]) can be considered as response to the increased economic importance of science and technology which had become visible in economic statistics during the 1950s.

The idea that scientific knowledge can be organized deliberately and controlled from a mission perspective (for example, for military purposes) was a result of World War II. Before that time the intellectual organization of knowledge had largely been left to the internal mechanisms of discipline formation and specialist communications. The military impact of science and technology through knowledge-based development and mission-oriented research during World War II (e.g., the Manhattan project) made it necessary in 1945 to formulate a new science and technology policy under peacetime conditions.

In 1945, Vannevar Bush’s report to the U.S. President entitled The Endless Frontier contained a plea for a return to a liberal organization of science. ***Quality control should be left to the internal mechanisms of the scientific elite, for example, through the peer review system.*** The model of the U.S. National Science Foundation from 1947 was followed by other Western countries. For example, the Netherlands created its foundation for Fundamental

Scientific Research (ZWO) in 1950. With hindsight, one can consider this period as the institutional phase of science policies: the main policy instrument was the support of science with institutions to control its funding [Okubo, 1997].

The attention for the measurement of scientific communication originated from *an interest other than research evaluation*. During the 1950s and 1960s, the scientific community itself had become increasingly aware of the seemingly uncontrolled expansion of scientific information and literature during the postwar period. In addition to its use in information retrieval, the Science Citation Index produced by Eugene Garfield’s Institute of Scientific Information came soon to be recognized as a means to objectify standards [Price, 1963; Elkana et al, 1978]. The gradual introduction of output indicators (e.g., numbers of publications and citations) could be legitimated both at the level of society - because it enables policy makers and science administrators to use arguments of economic efficiency - and internally, because quality control across disciplinary frameworks becomes difficult to legitimate unless objectified standards can be made available in addition to the peer review process [Leydesdorff, 2005].

In 1976 Francis Narin’s pioneering study “Evaluative Bibliometrics” [Narin, 1976] was published under the auspices (not incidentally) of the U.S. National Science Foundation. In 1973 Henry Small had proposed a method for mapping the sciences based on the co-citations of scientific articles. While Small’s approach tried to agglomerate specialties into disciplinary structures, Narin focused on hierarchical structures that operate top-down [Carpenter & Narin, 1973; Pinski & Narin, 1976]. This program appealed to funding agencies like the N.S.F. and N.I.H. that faced difficult decisions in allocating budgets across disciplinary frameworks [Leydesdorff, 2005].

Recent years have seen quantitative bibliometric indicators being increasingly used as a central element in the assessment of the performance of scientists, either individually or as groups, and as an important factor in evaluating and scoring research proposals.

These indicators are varied (see [bibliometric, 2012]), and include e.g.:

- Citation counts of individual papers published by researchers;

- Journal metrics (the impact factors of the journals);
- Measures that quantify personal research contributions over an extended period.

Journal metrics

Journal metrics measure the performance and/or impact of ***scholarly journals***. Each metric has its own particular features, but in general, *they all follow the theories and practices of advertizing and aim to provide rankings* and insight into journal performance *based on citation analysis* (very similar to “audience reach measures” and rankings).

They start from the basic premise that a citation to a paper is a form of endorsement, and the most basic analysis can be done by simply counting the number of citations that a particular paper attracts: more citations to a specific paper means that more people consider that paper to be important.

Citations to journals (via the papers they publish) can also be counted, thus indicating how important a particular journal is to its community, and in comparison to other journals. Different journal metrics use different methodologies and data sources, thus offering different perspectives on the scholarly publishing landscape, and bibliometricians use different metrics depending on what features they wish to study [Elsevier, 2011].

For example, let remember four metrics:

- Journal Impact Factor (IF);
- SCImago Journal Rank (SJR);
- Eigenfactor;
- Source-Normalized Impact per Paper (SNIP).

Journal ***Impact Factor***; is a measure of a journal’s average citations per article. The impact factor was computed by dividing the number of citations by the number of articles contained in the journal. This made it possible to eliminate any bias stemming from a journal’s size, rendering citation proportional to the number of articles.

The Impact Factor (IF) is the brainchild of Dr. Eugene Garfield, who devised a system of quantifying the number of times a manuscript is referenced in the literature [Teixeira da Silva & Van, 2013]. As indicated by Thomson Reuters (http://thomsonreuters.com/products_services/science/free/essays/impact_factor/), the IF is calculated as an extremely simple equation:

Year impact factor ***IF = C/N***, where ***C*** = Cites to articles published in two previous years (Year-1) and (Year-2) (this is a subset of total cites in current Year); ***N*** = number (sum) of articles published in Year-1 and Year-2.

Developed by Professor Félix de Moya, ***SCImago Journal Rank*** (SJR) [SCI, 2013] is a prestige metric based on the idea that “*all citations are not created equal*”. With SJR, the subject field, quality, and reputation of the journal have a direct impact on the value of a citation. This means that a citation from a source with a relatively high SJR is worth more than a citation from a source with a lower SJR.

The essential idea underlying the application of these arguments to the evaluation of scholarly journals is to assign weights to bibliographic citations based on the importance of the journals that issued them, so that citations issued by more important journals will be more valuable than those issued by less important ones. This "importance" will be computed recursively, i.e., the important journals will be those which in turn receive many citations from other important journals [González-Pereira et al, 2009].

SJR assigns relative scores to all of the sources in a citation network. Its methodology is inspired by the Google PageRank algorithm, in that not all citations are equal. A source transfers its own ‘prestige’, or status, to another source through the act of citing it. A citation from a source with a relatively high SJR is worth more than a citation from a source with a lower SJR. A source’s prestige for a particular year is shared equally over all the citations that it makes in that year; this is important because it corrects for the fact that typical citation counts vary widely between subject fields. The SJR of a source in a field with a high likelihood of citing is shared over a lot of citations, so each citation is worth relatively little. The SJR of a source in a field with a low likelihood if citing is shared over few citations, so each citation is worth relatively much. The result is to even out the differences in citation practice between subject fields, and facilitate direct comparisons of sources. SJR emphasizes those sources that are used by prestigious titles [Elsevier, 2011].

The ***Eigenfactor® score*** of a journal is an estimate of the percentage of time that library users spend with that journal. The Eigenfactor algorithm corresponds to a simple model of research in which readers follow chains of citations as they move from journal to journal. Imagine that a researcher goes to the library and selects a journal article at random. After reading the article, the researcher selects at random one of the citations from the article. She then proceeds to the journal that was cited, reads a random article there, and selects a citation to direct her to her next journal volume. The researcher does this *ad infinitum*.

The amount of time that the researcher spends with each journal gives us a measure of that journal's importance within network of academic citations. Moreover, if real researchers find a sizable fraction of the articles that they read by following citation chains, the amount of time that our random researcher spends with each journal gives us an estimate of the amount of time that real researchers spend with each journal. While we cannot carry out this experiment in practice, we can use mathematics to simulate this process [Bergstrom, 2007].

Source-Normalized Impact per Paper (SNIP) corrects for differences in the frequency of citation across research fields. SNIP measures a source's contextual citation impact. It takes into account characteristics of the source's subject field, especially the frequency at which authors cite other papers in their reference lists, the speed at which citation impact matures, and the extent to which the database used in the assessment covers the field's literature. SNIP is the ratio of a source's average citation count per paper, and the 'citation potential' of its subject field. It aims to allow direct comparison of sources in different subject fields.

A source's subject field is the set of documents citing that source. The citation potential of a source's subject field is the average number of references per document citing that source. It represents the likelihood of being cited for documents in a particular field. A source in a field with a high citation potential will tend to have a high impact per paper.

Citation potential is important because it accounts for the fact that typical citation counts vary widely between research disciplines – they tend to be higher in Life Sciences than in Mathematics or Social Sciences, for example. If papers in one subject field contain on average 40 cited references while those in another contain on average 10, then the former field has a citation potential that is four times higher than that of the latter. Citation potential also varies between subject fields within a discipline. For instance, basic journals tend to show higher citation potentials than applied or clinical journals, and journals covering emerging topics tend to have higher citation potentials than periodicals in well established areas.

For sources in subject fields in which the citation potential is equal to the average of the whole database, SNIP has the same value as the 'standard' impact per paper. But in fields with a higher citation potential – for instance, a topical field well covered in the database – SNIP is lower than the impact per paper. In fields in which the citation potential is lower – for instance, more classical fields, or those with moderate database coverage – SNIP tends to be higher than the impact per paper. In this way, SNIP allows you to rank your own customized set of sources, regardless of their subject fields [Elsevier, 2011].

Concluding this chapter we have to remember that a metric in business is a measure used to gauge some quantifiable component of an organization's performance, such as return on investment (ROI), or revenues. Metrics are part of the broad area of business intelligence used to help business leaders make more informed decisions. Organizations often use metrics to develop a systematic approach to transform an organization's mission statement and strategy into quantifiable goals, and to monitor the organization's performance in terms of meeting those goals [GPM, 2010]. At the knowledge market, the journal metrics are aimed for quantitative evaluation the popularity and importance of the journals as well as their impact. These metrics have to be used carefully. They are useful for publishers, librarians and administrators, but are not applicable for evaluating of personal scientific contributions. At first, the quantity personal measures were introduced to achieve this goal.

Quantity measures

Quantity measures that quantify personal research contributions over an extended period are based mainly on the idea of [Hirsch, 2005]. Several papers related to research indices were proposed to assess the quality of the academic research publications. Each one of those indices has its own strengths and weaknesses. The idea of having research indices started when J. Hirsh proposed the H-index [Hirsch, 2005].

Although the H-index has many limitations and seems biased or unfair in many cases, the other proposed indices such as: G-, H(2)-, HG-, Q^2 -, AR-, M-quotient, M-, W-, H_w- ,E-, A-, R- , W-, J-index, etc. considered H-index as a suitable base to produce those other indices with some behavioral enhancements in order to overcome its limitations. In fact, all the other indices are calculated based on the number of citations (originally proposed in H-index) which the authors' papers received. The differences between those indices can be shown through how the index deals with the citations number, as in H-index, G-index, W-index, or in adding new attributes such as time, average…etc as in Contemporary H-index, M-quotient, and AR- index [Maabreh & Alsmadi, 2012]. A review focused in h-Index variants, computation and standardization for different scientific fields is given in [Alonso et al, 2009]. Following [Bornmann et al, 2008] in Table 1 below we remember some definitions of popular indexes.

Table 1. Definitions of the h index and its variants [Bornmann et al, 2008]

Index	Definition
N/yr	Total number of publications (N) divided by years of publishing (yr)
N_{pr}/yr	Number of peer-reviewed publications (N_{pr}) divided by years of publishing (yr)
Cit	Total number of citations (Cit) received by an author
Cit/N	Citations per publication
H index [Hirsch, 2005]	A scientist has index *h* if *h* of his or her N_p published papers have at least *h* citations each and the other (N_p - *h*) papers have fewer than ≤ *h* citations each"

Index	Definition
M quotient [Hirsch, 2005]	$\frac{h}{y}$ where *h* = *h* index, *y* = number of years since publishing the first paper
G index [Egghe, 2006]	“The highest number *g* of papers that together received g^2 or more citations”
H(2) index [Kosmulski, 2006]	“A scientist’s *h*(2) index is defined as the highest natural number such that his *h*(2) most-cited papers received each at least $[h(2)]^2$ citations”
A index [Jin, 2006]	$\frac{1}{h}\sum_{j=1}^{h} cit_j$ where *h* = *h* index, *cit* = citation counts
M index [Bornmann et al, 2008]	The median number of citations received by papers in the Hirsch core (this is the papers ranking smaller than or equal to *h*)
R index [Jin et al, 2007]	$\sqrt{\sum_{j=1}^{h} cit_j}$ where *h* = *h* index, *cit* = citation counts
AR index [Jin et al, 2007]	$\sqrt{\sum_{j=1}^{h} \frac{cit_j}{a_j}}$ where *h* = *h* index, *cit* = citation counts, *a* = number of years since publishing
H_w index [Egghe & Rousseau, 2008]	$\sqrt{\sum_{j=1}^{r_o} cit_j}$ where *cit* = citation counts, r_o = the largest row index *j* such that $r_w(j) \leq cit_j$
Creativity index (C_a) [Soler, 2007]	$\sum_{i=1}^{N_p} \frac{c(n_i, m_i)}{a_i}$ where: N_p=Number of published papers; n_i=Number of references for paper “i”; m_i=Number of citations for paper “i”; a_i=Number of authors for paper “i”; c=not clearly defined in reference

Disadvantages of journal metrics and quantitative measures

At the first glance, the variety of scientific measures seems to be very great and with great differences.

Really, they all are based on counting the citations and similar formulas based or not on additional criteria like prestige of the journals, time periods, number of authors, etc.

The indexes for quantifying personal research contributions are based on same idea of the Hirsh with modifications.

The subject of limitations in research indices is still evolving and with all proposed indices, there are still limitations and weaknesses. Moreover, the large number of available indices may lead to the dispersion of the evaluation, and therefore produce differences in values among research communities or even countries [Maabreh & Alsmadi, 2012].

References may also be negative. An author may be cited for research of a controversial nature or for an error of methodology. Here too, citation does not always measure the quality of research but rather the impact of a particular piece of work or of an individual scientist [Okubo, 1997].

At the end, if an academic shows good citation metrics, it is very likely that he or she has made a significant impact on the field. However, the reverse is not necessarily true. If an academic shows weak citation metrics, this may be caused a lack of impact on the field. However, it may also be caused by: working in a small field; publishing in a language other than English (LOTE); or publishing mainly (in) books [Harzing, 2008].

Sites and tools that are interested in the evaluation of researchers and research publications may have to calculate and display all the indices, and this may cause two issues [Maabreh & Alsmadi, 2012]:

- Large number of indices, if used, may clutter pages and make them unreadable;
- Since most likely values will be different among those indices, and in some cases they may even contradict with each other, such information will be misleading to the reader rather than being helpful or informative.

From the beginning, the quantitative measuring of scientific work has been criticized due to problems raised during evaluation of scientific results. Let point one of the earliest papers "Why the impact factor of journals should not be used for evaluating research" [Seglen, 1997]. Its arguments are still valid:

Problems associated with the use of journal impact factors [Seglen, 1997]

- Journal impact factors are not statistically representative of individual journal articles;
- Journal impact factors correlate poorly with actual citations of individual articles;

- Authors use many criteria other than impact when submitting to journals;
- Citations to "non-citable" items are erroneously included in the database;
- Self citations are not corrected for;
- Review articles are heavily cited and inflate the impact factor of journals;
- Long articles collect many citations and give high journal impact factors;
- Short publication lag allows many short term journal self citations and gives a high journal impact factor;
- Citations in the national language of the journal are preferred by the journal's authors;
- Selective journal self citation: articles tend to preferentially cite other articles in the same journal;
- Coverage of the database is not complete;
- Books are not included in the database as a source for citations;
- Database has an English language bias;
- Database is dominated by American publications;
- Journal set in database may vary from year to year;
- Impact factor is a function of the number of references per article in the research field;
- Research fields with literature that rapidly becomes obsolete are favored;
- Impact factor depends on dynamics (expansion or contraction) of the research field;
- Small research fields tend to lack journals with high impact;
- Relations between fields (clinical v basic research, for example) strongly determine the journal impact factor;
- Citation rate of article determines journal impact, but not vice versa;

Summary points [Seglen, 1997]:

- Use of journal impact factors conceals the difference in article citation rates (articles in the most cited half of articles in a journal are cited 10 times as often as the least cited half);
- Journals' impact factors are determined by technicalities unrelated to the scientific quality of their articles;
- Journal impact factors depend on the research field: high impact factors are likely in journals covering large areas of basic research with a rapidly expanding but short lived literature that use many references per article;

- Article citation rates determine the journal impact factor, not vice versa.

These problems still exist and are object for current discussions. For example, the major disadvantage of the Web of Science is that it may provide a substantial underestimation of an individual academic’s actual citation impact. This is true equally for the two functions most generally used to perform citation analyses – for the “general search” and for the Web of Science “cited reference”. However, the Web of Science “general search” function performs more poorly in this respect than the “cited reference” function. There are a number of reasons for the underestimation of citation impact by Thomson ISI Web of Science, for instance [Harzing, 2008]:

- Web of Science General Search is limited to ISI-listed journals - In the General Search function Web of Science only includes citations to journal articles published in ISI listed journals [Roediger, 2006]. Citations to books, book chapters, dissertations, theses, working papers, reports, conference papers, and journal articles published in non-ISI journals are not included;
- Web of Science Cited Reference is limited to citations from ISI-listed journals - In the Cited Reference function Web of Science does include citations to non-ISI publications. However, it only includes citations from journals that are ISI-listed.

Both Google Scholar and Thomson ISI Web of Science have problems with academics that have names including either diacritics (e.g. Özbilgin or Olivas-Luján) or apostrophes (e.g. O'Rourke) [Harzing, 2008]:

- In Thomson ISI Web of Science a search with diacritics provides an error message and no results;
- In Google Scholar a search for the name with diacritics will generally not provide any results either.
- For both databases doing a search without the diacritic will generally provide the best result.

The popularity and the wide use of the h-index have raised a lot of criticism.

The most notable and well-documented example of critical view on the h-index (and other “simple” measures of research performance) is the report by the joint

Committee on Quantitative Assessment of Research [Adler et al, 2008]. In this report, the authors argue strongly against the use (or misuse) of citation metrics (e.g., the impact factor or the h-index) alone as a tool for assessing quality of research, and encourage the use of more complex methods for judging scientists, journals or disciplines, that combine both citation metrics as well as other criteria such as memberships on editorial boards, awards, invitations or peer reviews. With regard to the h-index (and associated modifications), specifically, [Adler et al, 2008] stress that its simplicity is a reason for failing to capture the complicated citation records of researchers, loosing thus crucial information essential for the assessment of a scientist’s research. The lack of mathematical/statistical analysis on the properties and behavior of the h-index is also mentioned. This is in contrast to the rather remarkable focus of many articles to demonstrate correlations of h-index with other publication/citation metrics (i.e. published papers or citations received), a result which according to the authors is self-evident, since all these variables are essentially functions of the same basic phenomenon, i.e. publications [Panaretos & Malesios, 2009].

Besides the above-mentioned works, there are many more articles referring to disadvantages of the h-index. In what follows we list some of the most important disadvantages of the h-index [Panaretos & Malesios, 2009]:

- The h-index is bounded by the total number of publications. This means that scientists with a short career (or at the beginning of their career), are at an inherent disadvantage, regardless of the importance of their discoveries. In other words, it puts newcomers at a disadvantage since both publication output and citation rates will be relatively low for them;
- Some authors have also argued that the h-index is influenced by self-citations. Many self-citations would give a false impression that the scientists’ work is widely accepted by the scientific community. Both self-citations and “real” (independent) citations are usually used in the calculation of the h-index. In this context, the emerging problem is that scientists with many co-operating partners may receive many self-citations, in contrast to scientists that publish alone;
- The h-index has slightly less predictive accuracy and precision than the simpler measure of mean citations per paper;
- Another problem is that the h-index puts small but highly-cited scientific outputs at a disadvantage. While the h-index de-emphasizes singular

successful publications in favor of sustained productivity, it may do so too strongly. Two scientists may have the same h-index, say, h = 30, i.e., they both have 30 articles with at least 30 citations each. However, one may have 20 of these papers that have been cited more than 1000 times and the other may have all of his/hers h-core papers receiving just above 30 citations each. It is evident that the scientific work of the former scientist is more influential;

- Limitations/differences of the citation data bases may also affect the h-index. Some automated searching processes find citations to papers going back many years, while others find only recent papers or citations;
- Another database related problem often occurring with a significant effect on the correct calculation of the h-index, is that of name similarities between researchers. It is almost impossible to find a scientist with a unique combination of family name and initials while searching the most known citation databases. As a result, in many cases the h-index will be overestimated, since in its calculation the works of more than one researcher are added;
- It seems that the h-index cannot be utilized for comparing scientists working in different scientific fields. It has been observed that average citation numbers differ widely among different fields;
- General problems associated with any bibliometric index, namely the necessity to measure scientific impact by a single number, apply here as well. While the h-index is one 'measure' of scientific productivity, some object to the practice of taking a human activity as complex as the formal acquisition of knowledge and condense it to a single number. Two potential dangers of this have been noted:

(a) Career progression and other aspects of a human's life may be damaged by the use of a simple metric in a decision-making process by someone who has *neither the time nor the intelligence* to consider more appropriate decision metrics;

(b) Scientists may respond to this by maximizing their h-index to the detriment of doing more quality work.

This effect of using simple metrics for making management decisions has often been found to be an unintended consequence of metric-based decision taking; for instance, governments routinely operate policies designed to minimize crime figures and not crime itself.

The disadvantages of the h-index may be seen in the indices which inherit its properties. For instance, some advantages and disadvantages of quantity metrics were outlined by [Thompson, 2009] (see Table 2).

Table 2. Some advantages and disadvantages of quantity metrics [Thompson, 2009]

Metric	Advantages	Disadvantages
N/yr	Measures gross productivity	Definition of “publication” can be arbitrary; No insight into the importance or impact of published works
N_{pr}/yr	Measures gross productivity Eliminates marginal publications	No insight into the importance or impact of published work
Cit	Measures total impact of a body of work	Can be inflated by a small number of papers with high citation counts.
Cit/N	Measures total impact of a body of work normalized by the number of published papers.	Tends to reward low productivity Can penalize high productivity
h-index	Combines quantitative (publication numbers) and impact (citation counts) into a simple whole number. Identifies a set of core, high performance journal articles (“Hirsch core”)	Insensitive to highly cited work
M quotient	Allows h-index comparisons between faculty that differ in seniority	Insensitive to highly cited work
G index	Once a paper makes the Hirsh core, additional citations in this group are not counted further; the g index takes these further citations into account	Gives more weight to highly cited papers
H(2) index	Since h(2) index is always smaller then h-index, it is less open to problems of citations accuracy	Possibly overly sensitive to a few highly cited papers
A index	Calculates the average number of citations in the Hirsch core	Emphasizes more of the impact of the Hirsch core than quantity. Can be very sensitive to a few highly cited papers
M index	Median value may be a better measure of central tendency because of the skewed nature of citation counts	Emphasizes more of the impact of Hirsch core than the quantity.

Metric	Advantages	Disadvantages
R index	Involves the Hirsch core but does not “punish” an author for having a high h-index unlike the a-index	Emphasizes more of the impact of the Hirsch core than quantity. Can be very sensitive to a few highly cited papers
AR index	Normalizes the r index by the number of years publishing allowing comparison of younger and more seasoned faculty	Similar to r index
Creativity index (C_a)	Only scholarship metric that proposes to measure creativity	Insufficient data to validate this metric at present. The calculation of the creativity index is not simple, however the author of paper has a free download of a program that will calculate the index

Very important disadvantage of quantitative measures is that they are applicable only to cited papers.

In 1991, David A. Pendlebury of the Philadelphia-based Institute for Scientific Information had published the startling conclusion that

> ***55% of the papers published in journals covered by ISI's citation database did not receive a single citation in the 5 years after they were published*** [Hamilton, 1991].

In his further publication, Pendlebury gave more concrete data. He had written [Pendlebury, 1991]:

“The figures -- 47.4% un-cited for the sciences, 74.7% for the social sciences, and 98.0% for the arts and humanities -- are indeed correct.

These statistics represent every type of article that appears in journals indexed by the Institute for Scientific Information (ISI) in its Science Citation Index, Social Sciences Citation Index, and Arts & Humanities Citation Index. The journals' ISI indexes contain not only articles, reviews, and notes, but also meeting abstracts, editorials, obituaries, letters like this one, and other marginalia, which one might expect to be largely un-cited.
In 1984, about 27% of the items indexed in the Science Citation Index were

such marginalia. The comparable figures for the social sciences and arts and humanities were 48% and 69%, respectively.

If one analyzes the data more narrowly and examines the extent of un-cited articles alone, the figures shrink, some more than others: **22.4%** of 1984 science articles remained un-cited by the end of 1988, as did **48.0%** of social sciences articles and **93.1%** of articles in arts and humanities journals.

If one restricts the analysis even further and examines the extent of un-cited articles by U.S. authors alone, the numbers are even less "worrisome."

Only 14.7% of 1984 science articles by U.S. authors were left un-cited by the end of 1988.

We estimate the share of un-cited 1984 articles by non-U.S. scientists to be about 28%” [Pendlebury, 1991].

Authors from developing countries

Whatever performance metrics we may use, it appears that ***authors from developing countries*** do face certain constraints in terms of achieving higher performance indices and therefore recognition for themselves and their country. *It is quite possible that authors from advanced countries may tend to cite publications from organizations located in their own countries, leading to a disadvantage for authors working in difficult situations, with less funding opportunities* Since there is a limited page budget and increased competition in many “high-profile” journals, it is *not always possible to publish in these journals*.

One way to overcome this problem is to encourage and give value to papers published in national journals. There are many scientists from developing countries such as India working in highly developed countries with advanced scientific infrastructure and huge funding. These scientists should seriously consider publishing their work in journals originating from their native countries. This will bring an international flavor to the national journals, attracting more international authors and ultimately making them mainstream international journals. When these journals become more visible and easily accessible through their online versions, there is a chance that papers published in these journals are more often cited [Kumar, 2009].

In other words, *developing national knowledge markets became mission important and considerable*.

Mentoring abilities

In addition, we should measure the ***mentoring abilities*** of a scientist. Scientists do research and also mentor younger colleagues. Good mentoring should be a significant consideration of one's contribution to science. The h-index might measure research productivity, but currently there does not appear to be a "*mentoring index*" [Jeang, 2008]. If the coauthors of a scientist are his or her own trainees or students and if they continue to make a scientific impact after leaving their supervisor, it does point to the quality of the mentoring by the scientist and to the impact made by the scientist, as a result of his/her mentoring abilities, in a given area during a given period. This is a very important but totally neglected aspect of the contribution made by a scientist or an academic.

However, *we do not yet have a well–worked out formula to measure such mentoring abilities* [Kumar, 2009].

Evaluation of Scientific Contributions

The products of science are not objects but ***ideas***, means of communication and reactions to the ideas of others. While it is possible simultaneously to track scientists and money invested, it is far more difficult to measure *science as a body of ideas*, or to grasp its interface with the economic and social system. For now, indicators remain essentially a unit of measure based on observations of science and technology as a system of activities rather than as a body of specific knowledge [National Science Foundation, 1989].

Research papers and publications are important indicators for the ability of an author or an education community to conduct research projects in the different human science fields. In general, the number of publications and the increase in this number is a direct indicator of the size or the volume of research activities for a particular author or university. Nonetheless, the number of publications merely, is showed to be a limited indicator to show the impact of those publications. The number of citations for a particular paper is shown to be more relevant and important in comparison to the number of publications. This is why early citation indices such as H-index and G-index gave more weight and important to the number of citations in comparison to the number of publications [Maabreh & Alsmadi, 2012].

Each indicator has its advantages and its limitations, and care must be taken not to consider them as “absolute” indices [Atanassov & Detcheva, 2012; Atanassov & Detcheva, 2013]. The “convergence” of indicators has to be tested in order to put the information they convey into perspective [Martin & Irvine, 1985]

Usefulness of Scientific Contribution

The Main Phases of the Science are

(1) Creation of a Scientific Result;

(2) Registration of the Scientific Result;

(3) Implementation and Using of the Scientific Result.

The bibliometric indexes analyze the second phase – registration of scientific result as (primary) publications and as (secondary) citations. The first and third phases are out of bibliometric scope. This way the evaluating of scientific work became partial and not significant. Practically, the evaluation of scientific results is closed in the contours of the Knowledge Markets (KM) shown at Figure 2 and/or Figure 3, i.e. without taking in account the main knowledge customers of the KM.

A possible step, to counterbalance and to extend consideration to all KM elements shown at Figure 1, is to analyze the publications and citations from point of view of the third phase – ***implementation and using the scientific results*** *by the members of KM.*

A wide spread understanding is that only high qualified ***academic researchers*** (Scientists (S), Figure 1) can evaluate published ideas. They have knowledge and skills to continue research and developing of proposed ideas and via citations they recognize previous research done by other scientists or by themselves. In accordance to *usefulness of cited ideas*, we may separate academic citations on three main groups:

— ***Substantial citations***, which applied or supported the citing work indicating implementation and using the citied results, including "mentoring impact";

— ***Casual citations***, which noted only or reviewed the citing work;

— ***Refuting citations,*** which indicate that the citing work (possibly) has no scientific added value.

Regarding ***industrial researchers*** (Researchers (R), Figure 1) we may make the similar consideration. They have knowledge and skills to implement the published ideas and to evaluate their usefulness for industrial applications. Here the citations are mainly in two groups:

— *Substantial citations*, which applied or supported the citing work indicating implementation and using the citied results, including "mentoring impact";

— *Refuting citations*, which indicate that the citing work (possibly) has no scientific added value to be implemented.

Further analysis of the KM-scheme concerns the educational cycle done by ***Lecturers*** ((L), Figure 1), ***Tutors*** ((T), Figure 1) and ***Examiners*** ((E), Figure 1). Their main goal is to assist Employees in learning of the published ideas. In this cycle, the citations are in text-books, methodical or other supporting publications, and educational learning materials. All such citations we may classify as:

— *Casual citations*, which noted only or reviewed the citing work.

The ***Employees*** ((Ee), Figure 1) may use the received knowledge in their everyday activities. During educational process they may create some new knowledge information objects with or without new ideas. For instance, they may prepare different theses, surveys, guides, papers, etc. In such case, the types of citations may vary, i.e. it may be:

— *Substantial citations*, which applied or supported the citing work indicating implementation and using the citied results, including “mentoring impact”;

— *Casual citations*, which noted only or reviewed the citing work;

— *Refuting citations,* which indicate that the citing work (possibly) has no scientific added value.

The ***Employers*** ((Er), Figure 1) are the most important members of KM. They invest both in developing man power as well as in research activities. In both cases the evaluation of usefulness of scientific results is not by citations in papers but by amount of invested assets. This way their citations may be classified only as

— *Substantial citations*, which applied or supported the citing work indicating implementation and using the citied results, including “mentoring impact”

if the amount of investments is over some normalized limit. Usually the investments are provided by scientific or educational projects and because of this we may assume that one project corresponds to one substantial citation.

At the end we have to pay attention to two main distributors of knowledge ***Publishers*** ((P), Figure 1) and ***Administrators*** ((A), Figure 1). After first publishing of the knowledge information objects (papers, books, etc.), Publishers start selling and corresponded advertizing. Main part of advertizing activities is indexing of published materials by different scientific digital libraries and data bases which are inherent for Administrators. All their citations may be classified as:

— *Casual citations*, which noted only or reviewed the citing work.

Transitive citations

The useful scientific results may cause a chain of publications which further use and develop them. This way, transitive citations will exist. Citation chain has to start from a substantial citation and to continue by same type citations because casual citations could not generate such citation chain.

The influence of the scientific ideas is greatest when citation chains exist. Because of this, the *transitive substantial citations* have to be counted as native characteristic of the scientific publications. It is correct to assume that a transitive substantial citation is equal to direct one.

Temporal dimension

There is also a ***temporal dimension*** to the citation process. An article may first be cited for substantial reasons (e.g., its content has been used). Later when a paper is widely known and has obtained many citations the importance of the other mechanisms will increase (authors citing authoritative papers, the bandwagon effect, etc.). In other words, ***visibility dynamics*** become more important over with time because of the self-intensifying mechanisms that are involved. This explains why the relative differences in citation rates between poorly cited and highly cited papers increase over time. Another temporal effect is the phenomenon termed "obliteration by incorporation", meaning that basic theoretical knowledge is not cited anymore. As a consequence, the most basic and important findings may not be among the most highly cited papers because they have been rapidly incorporated into the common body of accepted knowledge [Aksnes, 2005].

Concluding this short survey we have to draw attention to one very important fact.

A great number of publications have no chance to be viewed and further studied because they are published in media with limited and/or payable access. In this case only well-known authors have chance to be recognized and possibly – cited.

Only what is needed is publications to be included in different digital libraries with open access and *as more such libraries exist* in the world so greatest chance these publications have to become useful. The variety of digital libraries and index data bases with open access to scientific publications and reviews is a crucial factor for further grow of the science. One may say that such practice will destroy the knowledge markets. This is partially true. The societies invest in science by direct or indirect financing and further business with scientific results is not admissible

USC-methodology

Following considerations discussed above, we assume that for evaluating of usefulness of scientific contributions more-less important are:

- ***p*** – Number of the papers;
- ***q*** – Number of monographs;
- ***s*** – Number of the substantial citations;
- ***c*** – Number of the casual citations;
- ***r*** – Number of the refuting citations;
- $\boldsymbol{Y} = y_e - y_b + 1$ – Length of the interval of publications;
- $\boldsymbol{z} = y_c - y_b$ – Length of the interval of citations,

where

- y_b – starting year (beginning) of the period of publications;
- y_e – last year (end) of the period of publications;
- y_c – last year (end) of the period of citations.

In this list we have three different types of values which we have to reduce to common measurement unit. We propose to use "paper" as such unit because it may be assumed that *one paper represents a single idea*.

In accordance with this, we propose to use four coefficients of correlation:

- ***m*** – coefficient of the monograph correlation:
  - ⇒ ***m** : **1 monograph*** = ***m papers**;* example: if 1 monograph = 5 papers than m = 5;
- ***a*** – coefficient of the substantial citation correlation:
  - ⇒ ***a : 1 substantial citation*** = ***1/a paper**;* example: if 5 substantial citations = 1 paper than a=5;
- ***b*** – coefficient of the casual citation correlation:
  - ⇒ ***b : 1 casual citation*** = ***1/b paper**;* example: if 10 casual citations = 1 paper than b = 10;
- ***v*** – coefficient of the refuting citation correlation:
  - ⇒ ***v : 1 refuting citation*** = ***1/v paper**;* example: if 10 refuting citation = 1 paper than v = 10.

This way we have the methodological formula for *Usefulness of Scientific Contributions* (***usc-index***):

$$usc = \frac{p + mq + z}{Y} + \frac{s}{aY} + \frac{c}{bY} - \frac{r}{vY}$$

This formula is ***only a formal representation*** of the understanding that the ***scientific contributions have to be evaluated completely*** taking in account as more parameters as possible. All types of publications have to be included in the evaluation process as well as mentoring activities, learning materials, and all types of citations including transitive citations, implementations, scientific projects, received funding, etc.

Special comment is needed for ***substantial self-citations***. They are indicator that the scientists provide longtime investigation and step by step publish new results. This is normal cycle of science. Ignoring this means that we expect receiving the results in one “genius” invention. In addition, mentoring students and young researchers lead to publishing of co-authored papers which cause ***substantial citations from co-authors*** in further their independent work and publications. As the received knowledge is more qualitative so more important are the further citations from co-authors. Ignoring this means that we do not acknowledge the high level skills and leading ideas of the advisors.

Example

Results from an experiment with real data taken from DBLP (http://dblp.uni-trier.de/) are presented in Table 3.
In the real data there was no data for monographs and refuting citations. Because of this the corresponded columns contain zeroes.

Table 3. Experimental data for usc-index

scientist	**usc**	**y_b**	**y_e**	**y_c**	**Y**	**z**	**m**	**a**	**b**	**v**	**p**	**q**	**s**	**c**	**r**
S1	**26.07**	1991	2011	2009	**21**	18	5	5	10	10	405	0	15	1215	0
S2	**13.74**	1983	2011	2011	**29**	28	5	5	10	10	109	0	208	2200	0
S3	**13.52**	1995	2011	2011	**17**	16	5	5	10	10	110	0	32	975	0
S4	**11.66**	1981	2011	2011	**31**	30	5	5	10	10	181	0	50	1406	0
S5	**8.48**	1999	2011	2010	**13**	11	5	5	10	10	44	0	8	537	0
S6	**8.23**	1972	2011	2011	**40**	39	5	5	10	10	98	0	66	1789	0
S7	**6.68**	2000	2011	2007	**12**	7	5	5	10	10	53	0	10	182	0
S8	**5.57**	1985	2011	2011	**27**	26	5	5	10	10	68	0	22	520	0

S9	**4.36**	2007	2011	2010	**5**	3	5	5	10	10	16	0	1	26	0
S10	**3.87**	1991	2010	2010	**20**	19	5	5	10	10	44	0	1	142	0
S11	**3.71**	2003	2011	2008	**9**	5	5	5	10	10	26	0	0	24	0
S12	**3.62**	2004	2009	2011	**6**	7	5	5	10	10	8	0	0	67	0
S13	**3.62**	1983	2009	2011	**27**	28	5	5	10	10	47	0	2	223	0
S14	**3.54**	1973	1986	2008	**14**	35	5	5	10	10	11	0	2	32	0
S15	**3.33**	2009	2011	2010	**3**	1	5	5	10	10	8	0	1	8	0
S16	**3.16**	1995	2009	2011	**15**	16	5	5	10	10	18	0	2	130	0
S17	**2.42**	1986	2011	2006	**26**	20	5	5	10	10	34	0	2	85	0
S18	**2.35**	2008	2011	2011	**4**	3	5	5	10	10	6	0	1	2	0
S19	**1.63**	2001	2011	2008	**11**	7	5	5	10	10	10	0	1	7	0
S20	**0.96**	1991	2006	2001	**16**	10	5	5	10	10	5	0	1	1	0

USC-index reflects the dynamics of scientific development during the analyzed period. For instance, scientist S2 has more long scientific career and more citations than S1 but his usc-index is less than that of S1 due to less number of papers for longer period.

It is important to remark: *periods have different lengths (column Y) and for further analysis it has to be accounted*.

It is complicated to compute usc-index for all scientists of a given organization and many times more complicated to do this for all researchers from given scientific area. Because of this, the computer linguistic analysis of the scientific publications (to obtain values of the main parameters of usc-index) is serious scientific problem which has to be solved. Some preliminary considerations about possibility for solving it may be done. For instance, it is typical that the introduction of a scientific article is structured as a progression from the general to the particular. References have been found to be most frequent in the introductory section of paper. Thus, in the introduction, an article typically refers to more general or basic works within a field. The net effect of many articles referring to the same general works, therefore, is that such contributions get a very large number of citations. References to highly cited publications seemed to occur more in the introduction than anywhere else in the articles. Similarly, since most scientific articles contain a methodology section in which the methods applied in the study are documented, authors typically cite the basic papers describing these methods. This may explain why some papers containing commonly used methods sometimes receive a very large number of citations [Aksnes, 2005].

Conclusion

Starting point of our consideration was the introduction of the “Information Market” as a payable information exchange and based on it information interaction. In addition, special kind of Information Markets - the Knowledge Markets (KM) were outlined. Basic understanding of our work is that we have to evaluate the usefulness of scientific contributions from point of view of those for whom the results are created. This is not simple task because the KM customers are of many kinds.

The identifying of the staple commodities of the knowledge markets was a step of the process of investigation of contemporary situation in the global knowledge environment. Investigation of the staple commodities of the knowledge markets is very difficult but useful task. We have introduced them as kind of information objects, called “knowledge information objects”. The main their distinctive characteristic is that they contain information models, which concerns sets of information models and interconnections between them.

We belong to the modern knowledge market and perhaps we shall agree that “à la marché comme à la marché” ("at the market as at the market"). In the world of science, there exist commercial interests that set the trends to redistribute the money given for science by the societies. Unfortunately, for instance, the "impact factor" is just such trend, borrowed from advertising industry, to force scientists to invest in selected retailer chains.

It is not permissible to replace the quality of a scientific publication, with qualities of the media in which it has been published.

In science, the incorrect management decisions lead to a decline in its development. *If a complete scientific "industry" is not developed, the "complete" administrative attitude to science grows, which inevitably will kill it.* Exuberant dependence on single numbers to quantify scientists’ contribution and make administrative decisions can affect their career progression or may force people to somehow enhance their h-index instead of focusing on their more legitimate activity, i.e., doing good science. Considering the complex issues associated with the calculation of scientific performance metrics, it is clear that a comprehensive approach should be used to evaluate the research worth of a scientist. We should not rely excessively on a single metric [Kumar, 2009].

Although the use of such quantitative measures may be considered at first glance to introduce objectivity into assessment, the exclusive use of such indicators to measure science “quality” can cause severe bias in the assessment process when applied simplistically and without appropriate benchmarking to the research environment being considered. Funding agencies are aware of this, nevertheless experience shows that the reviewing of both individuals and projects on the national and European level is still relying excessively on the use of these numerical parameters in evaluation. This is a problem of much concern in the scientific community, and there has been extensive debate and discussion worldwide on this topic [bibliometric, 2012].

Since the very first applications of bibliometric indicators in this way, scientists and science organizations have taken strong positions against such purely numerical assessment. Various organizations in Europe have published studies on their potential adverse consequences on the quality of funded scientific research. A prime example is the publication of the Académie des Sciences of the Institut de France that has presented clear recommendations on the correct use of bibliometric indices [IDF, 2011]. Other publications have addressed the role of peer review in the assessment of scientists and research projects e.g. the European Science Foundation Peer Review Guide published in 2011 [ESF, 2011a] with recommendations for good practices in peer review following an extensive European survey on peer review practices [ESF, 2011b]. Other recent examples are a study of peer review in publications by the Scientific and Technology Committee of the House of Commons in the UK [STC, 2011], the peer review guide of the Research Information Network in the UK [RIN, 2010] and the recommendations formulated at a workshop dedicated to quality assessment in peer review of the Swedish Research Council [SRC, 2009].

A common conclusion of these studies is the recognition of the important role of peer review in the quality assessment of research, and the recommendation to apply bibliometric performance indicators with great caution, and only by peers from the particular discipline being reviewed [bibliometric, 2012].

A considerable step toward this goal is ***The San Francisco Declaration on Research Assessment*** (DORA), [DORA, 2012] initiated by the American Society for Cell Biology (ASCB) together with a group of editors and publishers of scholarly journals, who recognize the need to improve the ways in which the outputs of scientific research are evaluated. The group met in December 2012 during the ASCB Annual Meeting in San Francisco and subsequently circulated a draft declaration among various stakeholders. DORA as it now stands has benefited from input by many of the original signers. It is a worldwide initiative covering all scholarly disciplines.

A special press release of *Initiative for Science in Europe (ISE)* called ***“Initiative to put an end to the misuse of the journal impact factor (JIP)”*** has been published [ISE, 2012]. We have kind permission of ISE to reprint text:

> “Major European science organizations have joined the "San Francisco Declaration On Research Assessment" which was released today by the American Society for Cell Biology (ASCB). Signatories in Europe include the European Mathematical Society, EUCheMS, European Sociology Association, European Education Research Association, FEBS, EMBO and other societies and organizations that are organized under the umbrella of the Initiative for Science in Europe (ISE).
>
> The increasing reliance on journal based metrics for research assessment, hiring, promotion or funding decisions has been criticized by experts for a number of years. The "San Francisco Declaration On Research Assessment" for the first time unites researchers, journals, institutions and funders to address the problems of an overreliance on the journal impact factor and to work for change of the current system of research assessment.
>
> The declaration formulates concrete recommendations for different stakeholder groups. It calls publishers to "greatly reduce emphasis on the journal impact factor as a promotional tool", funding agencies and institutions to consider "the value and impact of all research outputs" for purpose of research assessment, "including qualitative indicators of research impact" and researchers to make "decisions about funding,

> hiring, tenure, or promotion, [..] based on scientific content rather than publication metrics" when involved in assessment committees. It also invites organizations that supply metrics to "[b]e open and transparent by providing data and methods used to calculate all metrics".
>
> The San Francisco Declaration on Research Assessment was drafted by a group of editors and publishers of scholarly journals that met at the Annual Meeting of The American Society for Cell Biology (ASCB) in San Francisco in December 2012. It has since developed into a worldwide initiative welcoming all scientific disciplines including the social sciences and humanities.
>
> Scientists and institutions alike are invited to express their commitment and support for the initiative at http://ascb.org/SFdeclaration.html" [ISE, 2012].

Endorsing DORA, the Association for Computers and the Humanities (ACH) remarked that it is a set of recommendations for applying more nuanced, accurate ways to evaluate research than the Journal Impact Factor (JIF). DORA makes eighteen recommendations for researchers, funders, research institutions, organizations that provide metrics, and publishers, such as focusing evaluation on the content of a paper, applying article-based rather than journal-based metrics, incorporating research outputs such as datasets and software in evaluating impact, and promoting the reuse of reference lists through the adoption of Creative Commons Public Domain Dedication licenses.

In addition, we have to underline that the variety of digital libraries and index data bases with open access to scientific publications and reviews is a crucial factor for further grow of the science. One may say that such practice will destroy the knowledge markets. This is only partially true because the societies invest in science by direct or indirect financing and further business with scientific results is not admissible

Following the considerations given above, this paper was aimed to present a new *usc-methodology* for evaluating the scientific contribution of a scientist or a scientific group (organization).

It consists in proposing three main groups of citations: ***Substantial citations***, ***Casual citations***, and ***Refuting citations***, which all have *temporal dimensions*.

In addition, due to existence of different types of values (for monographs, papers and citations), a common measurement unit (“idea” or “paper”) and four coefficients (for monographs, substantial, casual, and refuting citations) of correlation to measurement unit (paper) have been proposed.

The problem of automatic linguistic analysis of scientific publications, in accordance with usc-methodology and computing of its ***usc-index*** for different target scientific structures has been outlined.

Finally, we have to underline, that usc-methodology is aimed only to turn process of evaluation of scientific contributions back to human responsibility of authors, reviewers, and publishers. Modern science is distributed all over the world and concentration of any it’s part in one or two monopolies is absolutely inadmissible. To ensure growing of science we are obligated to provide for growing of variety of possibilities for doing science – financial resources, publishing opportunities, scientific indexing systems, and distributing organizations.

In addition to all printed universe we are obligated to take in account the variety of possibilities for direct contact between scientists in a single place like conferences, seminars, and workshops or distributed geographically like tele-conferences, electronic mailing lists, blogs, etc.

Special comment was done for *substantial self-citations*. They are indicator that the scientists provide longtime investigation and step by step publish new results. In addition, mentoring students and young researchers lead to publishing of co-authored papers which cause *substantial citations from co-authors* in further their independent work and publications. As the received knowledge is more qualitative so more important are the further citations from co-authors. Ignoring this means that we do not acknowledge the high level skills and leading ideas of the researchers and advisors.

This *usc-index* is *only a formal representation* of the understanding that the *scientific contributions have to be evaluated completely* taking in account as more parameters as possible. All types of publications as well as mentoring activities, learning materials, and all types of citations including substantial self-citations, substantial citations from co-authors, transitive citations, implementations, scientific projects, received funding, etc. have to be included in the evaluation of usefulness of scientific contributions.

Acknowledgements

Authors are indebted to acad. Alexander Palagin, prof. Levon Aslanyan and colleagues from the scientific seminar on “Information Modeling” for fruitful discussions and recommendations.

This paper is partially supported by the project **ITHEA-XXI**.

Bibliography

[Adler et al, 2008] Adler, R., Ewing, J. and Taylor, P. “Citation Statistics”, Joint IMU/ICIAM/IMSCommittee on Quantitative Assessment of Research, 2008 http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf (accessed at 09.05.2013).

[AIRM, 2013] Association, Information Resources Management, "Data Mining: Concepts, Methodologies, Tools, and Applications (4 Volumes)." IGI Global, 2013, 1-2120. Web, 9 May. 2013, doi: 10.4018/978-1-4666-2455-9, http://www.igi-global.com/book/data-mining-concepts-methodologies-tools/68176 (accessed at 09.05.2013).

[Aksnes, 2005] Aksnes D.W. Citations and their use as indicators in science policy. // Dissertation for the doctoral degree of the University of Twente, March 2005.

[Alonso et al, 2009] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera. h-Index: A review focused in its variants, computation and standardization for different scientific fields. // Journal of Informetrics 3 (2009) pp. 273–289. http://sci2s.ugr.es/hindex/pdf/JOI-3-4-273-289.pdf (accessed at 29.05.2013).

[AMA, 2013] American Marketing Association, “Dictionary - MarketingPower, inc. 2013”, http://www.marketingpower.com/_layouts/Dictionary.aspx?dLetter=A (accessed at 09.05.2013).

[Atanassov & Detcheva, 2012] V. Atanassov, E. Detcheva, Theoretical Analisis of Empirical Relationships for Pareto-Distributed Scientometric Data, Int. J.

"Information Models and Analyses", Vol.1/2012, ISSN 1314-6416, pp.271-282.

[Atanassov & Detcheva, 2013] V. Atanassov, E. Detcheva, Citation-Paper Rank Distributions and Associated Scientometric Indicators – A Survey, Int. J. "Information Models and Analyses", Vol.2/2013, ISSN 1314-6416 (in print).

[Bergstrom, 2007] Carl Bergstrom, “Eigenfactor: Measuring the value and prestige of scholarly journals”. College & Research Libraries News, 2007, 68(5), 314–316, http://octavia.zoology.washington.edu/publications/Bergstrom07.pdf, (accessed at 18.05.2013).

[bibliometric, 2012] “On the use of bibliometric indices during assessment” EPS Statement // 5-11 June 2012 http://c.ymcdn.com/sites/www.eps.org/resource/collection/B77D91E8-2370-43C3-9814-250C65E13549/EPS_statement_June2012.pdf (accessed at 09.05.2013).

[Bornmann et al, 2008] Lutz Bornmann, Rüdiger Mutz, Hans-Dieter Daniel. Are There Better Indices for Evaluation Purposes than the h Index? A Comparison of Nine Different Variants of the h Index Using Data from Biomedicine, Journal of the American society for information science and technology, 59(5):830–837, 2008.

[BSC, 2013] Britain's Science Council, Definition of “Science” http://www.sciencecouncil.org/definition (accessed 09.05.2013)

[Carpenter & Narin, 1973] Carpenter, M. P and Narin, F. “Clustering of scientific journals”, J. Am. Soc. Inf. Sci., 1973, 24, 425–436.

[CBED, 2013] Definition of indirect advertising noun from the Cambridge Business English Dictionary, Cambridge University Press, 2013, http://dictionary.cambridge.org/dictionary/business-english/indirect-advertising?q=indirect+advertising (accessed at 09.05.2013).

[DORA, 2012] San Francisco Declaration on Research Assessment // American Society for Cell Biology (ASCB). http://am.ascb.org/dora/ (accessed at 17.05.2013).

[Egghe & Rousseau, 2008] Egghe, L. and Rousseau, R. “An h-index weighted by citation impact”, Information Processing & Management, 44, 2008, pp. 770–780, doi:10.1016/j.ipm.2007.05.003.

[Egghe, 2006] Egghe, L. “An improvement of the h-index: the g-index”. ISSI Newsletter, 2(1), 2006b, pp. 8–9.

[Elkana et al, 1978] Elkana, Y., Lederberg, J., Merton, R. K., Thackray, A. and Zuckerman, H., Toward a Metric of Science: The Advent of Science Indicators, Wiley, New York, 1978.

[Elsevier, 2011] Elsevier B.V. “Frequently asked questions” Research analytics redefined Copyright © 2011 All rights reserved. Scopus is a registered trademark of Elsevier B.V. http://www.journalmetrics.com/faq.php (accessed at 09.05.2013).

[ESF, 2011a] European Science Foundation, European Peer Review Guide, Integrating Policies and Practices for Coherent Procedures, Member Organisation Forum, 2011. http://www.esf.org/fileadmin/Public_documents/Publications/European_Peer_Review_Guide.pdf (accessed at 26.05.2013).

[ESF, 2011b] European Science Foundation, Survey Analysis Report on Peer Review Practices, Member Organisation Forum, 2011, http://www.esf.org/fileadmin/Public_documents/Publications/pr_guide_survey.pdf (accessed at 26.05.2013)

[EzineMark, 2013] Ezine Mark. Direct Vs Indirect Marketing And Advertising. http://advertising.ezinemark.com/direct-vs-indirect-marketing-and-advertising-31ebf6babc8.html (accessed at 09.05.2013).

[Fingerman, 2006] Susan Fingerman “Web of Science and Scopus: Current Features and Capabilities” Electronic Resources Reviews Copyright 2006, Used with permission http://www.istl.org/06-fall/electronic2.html (accessed at 09.05.2013).

[Frascati Manual, 2002] Frascati Manual “The Measurement of Scientific and Technological Activities”, Proposed Standard Practice for Surveys on Research and Experimental Development. 6th edition (2002), pp. 266, // Organisation for Economic Co-operation and Development (OECD), Paris, France. 2002. ISBN 978-92-64-19903-9. http://www.oecd.org/innovation/inno/frascatimanualproposedstandardpracticeforsurveysonresearchandexperimentaldevelopment6thedition.htm. Publié en français sous le titre: Manuel de Frascati 2002. Méthode type proposée pour les enquêtes sur la recherche et le développement experimental, OECD Publications, Paris, France.

[Garfield, 1977] Garfield, E. Can Citation Indexing Be Automated? // Essay of an Information Scientist, vol. 1 (Vol. 1). Philadelphia: ISI Press. 1977.

[Garfield, 2007] Eugene Garfield, The evolution of the Science Citation Index, International Microbiology, 2007, 10:65-69, DOI: 10.2436/20.1501.01.10, ISSN: 1139-6709, http://garfield.library.upenn.edu/papers/barcelona2007.pdf, (accessed at 09.05.2013)

[Gladun, 1994] Gladun V. P. The processes of knowledge creation, (“Процессы формирования знаний”, Педагог 6), Sofia, 1994 (in Russian).

[Godin, 2005] Benoît Godin, From Eugenics to Scientometrics: Galton, Cattell and Men of Science. Project on the History and Sociology of S&T Statistics, Working Paper No. 32, Canadian Science and Innovation Indicators Consortium (CSIIC). 2006

[González-Pereira et al, 2009] Borja González-Pereira, Vicente P. Guerrero-Bote and Félix Moya-Anegón “The SJR indicator: A new indicator of journals' scientific prestige”, 2009, http://arxiv.org/ftp/arxiv/papers/0912/0912.4141.pdf (accessed at 09.05.2013).

[GPM, 2010] State Government of Victoria State Services Authority, A guide to people metrics// The State Services Authority, Melbourne, 2010.

[Hamilton, 1991] David P. Hamilton,” Research Papers: Who's Uncited Now?”, Science, 251:25, 1991, http://garfield.library.upenn.edu/papers/hamilton2.html (accessed at 09.05.2013).

[Harzing, 2008] Anne-Wil Harzing “Google Scholar - a new data source for citation analysis”, Research in International Management Products & Services for Academics, 2008, http://www.harzing.com/pop_gs.htm (accessed at 09.05.2013).

[Harzing, 2012] Harzing A.-W. How to become an author of ESI Highly Cited Papers? // University of Melbourne, 2012. http://www.harzing.com/esi_highcite.htm (accessed at 09.05.2013).

[Hawkins, 1982] Hawkins J.M., “*The Oxford Paperback Dictionary”,* Oxford University Press, 1982, ISBN 0-19-281209-2.

[Hirsch, 2005] Hirsch, J. E., “An index to quantify an individual’s scientific research output”, Proceedings of the National Academy of Sciences of the United States of America, 102(46), 2005, pp. 16569–16572, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1283832/pdf/pnas-0507655102.pdf (accessed at 09.05.2013)

[Hood & Wilson, 2001] William W. Hood, Conceptión S. Wilson. “The literature of bibliometrics, scientometrics, and informetrics”, Scientometrics, vol. 52, No. 2 (2001) 291–314, http://faculty.kfupm.edu.sa/MATH/kabbaj/Benchmarks/HoodWilson2001.pdf (accessed at 09.05.2013).

[Hornby et al, 1987] Hornby A.S., A.P. Cowie, A.C. Gimson, “*Oxford Advanced Learner’s Dictionary”,* Oxford University Press 1987, ISBN 0-19-431106-6

[IDF, 2011] Institut de France, Académie des Sciences, On the Proper use of Bibliometrics to Evaluate Individual Researchers, Rapport de l’Académie des sciences, 2011, www.academie-sciences.fr/activite/rapport/avis170111gb.pdf (accessed at 26.05.2013).

[ISE, 2012] Join Initiative to put an end to the misuse of the journal impact factor (JIP) // Initiative for Science in Europe (ICE). http://www.no-cuts-on-research.eu/news/dora.html (accessed at 17.05.2013).

[Ivanova et al, 2001] N. Ivanova, K. Ivanova, K. Markov, A. Danilov, K. Boikatchev. *The Open Education Environment on the Threshold of the Global Information Society,* IJ ITA, 2001, V.8, No.1 pp.3-12. (Presented at Int. Conf. KDS 2001 Sankt Petersburg, 2001, pp.272-280, in Russian, Presented at Int. Conf. ICT&P 2001, Sofia, pp.193-203)

[Ivanova et al, 2003] Kr. Ivanova, N. Ivanova, A. Danilov, I. Mitov, Kr. Markov, “Education of adult on the professional knowledge market”, (*Обучение взрослых на рынке профессиональных знаний,* Сборник доклади на Национална научна конференция “Информационни изследвания, приложения и обучение”), i.TECH-2003, Varna, Bulgaria, 2003, pp. 35-41, (in Russian).

[Ivanova et al, 2006] Kr. Ivanova, N. Ivanova, A. Danilov, I. Mitov, Kr. Markov. Basic Interactions between Members of the Knowledge Market. Int. Journal “Information Theories and Applications”, Vol.13/2006, No.:1, pp.19-30.

[Jeang, 2008] Kuan-Teh Jeang, “H-index, mentoring-index, highly-cited and highly accessed: how to evaluate scientists?” Retro virology, vol. 5, Article Number: 106, Nov 2008. BioMed Central, Springer. http://link.springer.com/article/10.1186%2F1742-4690-5-106 (accessed at 09.05.2013).

[Jeeves, 2013] Ask Jeeves. What is Indirect Advertising? http://uk.ask.com/question/what-is-indirect-advertising(accessed at 09.05.2013).

[Jin et al, 2007] Jin, B., Liang, L., Rousseau, R., & Egghe, L. “The R- and AR-indices: Complementing the h-index”, Chinese Science Bulletin, 52(6), 2007, pp. 855–863.

[Jin, 2006] Jin, B “h-index: an evaluation indicator proposed by scientist”, Science Focus, 1(1), 2006, pp. 8–9.

[Kosmulski, 2006] Kosmulski, M. A new Hirsch-type index saves time and works equally well as the original h-index. ISSI Newsletter, 2(3), 2006, pp. 4–6.

[Kumar, 2009] M. Jagadesh Kumar Evaluating Scientists: Citations, Impact Factor, h-Index, Online Page Hits and What Else? IETE TECHNICAL REVIEW, Vol. 26, ISSUE 3, 2009, DOI: 10.4103/0256-4602.50699; Paper No TR 81_09; Copyright © 2009 by the IETE, pp. 165 - 168

[Levine et al, 2005] Michele Levine, Gary Morgan, Angela Brooks, Marcus Tarrant, Howard Seccombe, “Advertising-Adverteasing-Advertiring?”, What sort of ROI can you expect from Print Advertising – unless your ad performs?” 12th Worldwide Readership Research Symposium, Prague - October 23-26, 2005, Roy Morgan International Melbourne, Australia, 2005

[Leydesdorff, 2005] Loet Leydesdorff, “The Evaluation of Research and the Evolution of Science Indicators”, Current Science, 2005, Vol. 89, No. 9, pp. 1510-1517, http://arxiv.org/ftp/arxiv/papers/0911/0911.4298.pdf (accessed at 09.05.2013)

[Maabreh & Alsmadi, 2012] Majdi Maabreh and Izzat M. Alsmadi “A Survey of Impact and Citation Indices: Limitations and Issues“, International Journal of Advanced Science and Technology, Vol. 40, March, 2012, pp. 35-54

[Markov et al, 2001] K. Markov, P. Mateev, K. Ivanova, I. Mitov, S. Poryazov. “*The Information Model”,* IJ ITA, 2001, V.8, No.2 pp.59-69 (Presented at Int. Conf. KDS 2001 Sankt Petersburg, 2001, pp.465-472)

[Markov et al, 2002] K. Markov, K. Ivanova, I. Mitov, N. Ivanova, A. Danilov, K. Boikatchev. *Basic Structure of the Knowledge Marke*t”, IJ ITA, 2002, V.9, No.4, pp.123-134 .

[Markov et al, 2003] Kr.Markov, Kr.Ivanova, I.Mitov, “*General Information Theory”, Basic Formulations,* FOI-COMMERCE, Sofia, 2003, ISBN 954-16-0024-1

[Markov et al, 2013] Kr. Markov, Kr. Ivanova, V. Velychko. “Evaluation of Scientific Contributions” // International Scientific Conference “Modern Informatics: Problems, Achievements, and Prospects of Development,” devoted to the 90th anniversary of academician V. M. Glushkov. Kiev, Ukraine, 2013 (in print).

[Markov et al, 2013a] Markov, Krassimir, Koen Vanhoof, Iliya Mitov, Benoit Depaire, Krassimira Ivanova, Vitalii Velychko and Victor Gladun. "Intelligent Data Processing Based on Multi-Dimensional Numbered Memory Structures", Diagnostic Test Approaches to Machine Learning and Commonsense Reasoning Systems, IGI Global, 2013. 156-184, Web. 7 May. 2013. doi:10.4018/978-1-4666-1900-5.ch007

[Markov et al, 2013b] Markov, Krassimir, Koen Vanhoof, Iliya Mitov, Benoit Depaire, Krassimira Ivanova, Vitalii Velychko and Victor Gladun. "Intelligent Data Processing Based on Multi-Dimensional Numbered Memory Structures." Data Mining: Concepts, Methodologies, Tools, and Applications, IGI Global, 2013, 445-473, Web. 6 May. 2013. doi:10.4018/978-1-4666-2455-9.ch022

[Markov et al., 2006] Kr. Markov, Kr. Ivanova, I. Mitov. “The Staple Commodities of the Knowledge Market”, Int. Journal “Information Theories and Applications”, Vol.13/2006, No.:1, pp.5-19.

[Markov, 1999] Kr. Markov. About harms from e-commerce (Относно вредите от електронната търговия), IECC'99: International e-Commerce Conference, ADIS & VIA EXPO, Bulgaria, Sofia, 1999 (in Bulgarian).

[Martin & Irvine, 1985] Br. Martin and J. Irvine, 'Evaluating the Evaluators: A Reply to Our Critics', Social Studies of Science, 1985, 15, pp. 558-75.

[Merton, 1957a] Merton, R.K. "Priorities in Scientific Discovery", American Sociological Review, Vol. 22, 1957, p. 635.

[Merton, 1957b] Merton, R. K. "Social and Democratic Social Structure", in Social Theory and Social Structure, Free Press, New York, 1957, pp. 550-561.

[Mooghali et al, 2011] A. Mooghali, R. Alijani, N. Karami, A. Khasseh "Scientometric Analysis of the Scientometric Literature" International Journal of Information Science and Management, Vol. 9, No. 1 January / June 2011, pp. 21-31

[Naidenova & Ignatov, 2013] Naidenova, Xenia and Dmitry I. Ignatov, "Diagnostic Test Approaches to Machine Learning and Commonsense Reasoning Systems", IGI Global, 2013, 1-367, Web 9, May 2013, doi:10.4018/978-1-4666-1900-5 http://www.igi-global.com/book/diagnostic-test-approaches-machine-learning/63869 (accessed at 09.05.2013).

[Narin, 1976] Francis Narin. Evaluative Bibliometrics: The use of Publication and Citation Analysis in the Evaluation of Scientific Activity. Computer Horizons, Inc. Project No. 704R, March 31, 1976. http://yunus.hacettepe.edu.tr/~tonta/courses/spring2011/bby704/narin_1975_eval-bibliometrics_images.pdf (accessed at 09.05.2013).

[National Science Foundation, 1989] National Science Foundation, "Science & Engineering Indicators" 1989, National Science Foundation, Washington, DC

[NLC, 2004] *Electronic Publishing: Guide to Best Practices for Canadian Publishers*. National Library of Canada, Ottawa, Created: 2001-10-03. Updated: 2004-03-0, http://www.nlc-bnc.ca/9/13/index-e.html, http://www.collectionscanada.ca/obj/p13/f2/01-e.pdf (accessed at 09.05.2013).

[Okubo, 1997] Okubo, Y. (1997), "Bibliometric Indicators and Analysis of Research Systems: Methods and Examples", OECD Science, Technology and Industry Working Papers, 1997/01, OECD Publishing. http://dx.doi.org/10.1787/208277770603 (accessed at 09.05.2013).

[Panaretos & Malesios, 2009] John Panaretos and Chrisovaladis Malesios, "Assessing scientist research performance and impact with single indices", MPRA Paper No. 12842, 2008, http://mpra.ub.uni-muenchen.de/12842/, Scientometrics, December 2009, Volume 81, Issue 3, pp 635-670.

http://link.springer.com/article/10.1007%2Fs11192-008-2174-9 (accessed at 09.05.2013).

[Pendlebury, 1991] David A. Pendlebury, "Science, Citation, and Funding", Science, 251:1410-1411, 1991. http://garfield.library.upenn.edu/papers/pendlebury.html (accessed at 09.05.2013).

[Peper, 2009] Peper M. Publish or Perish. May 22, 2009. http://blogs.library.duke.edu/libraryhacks/2009/05/22/publish-or-perish/ (accessed at 09.05.2013).

[Pinski & Narin, 1976] Pinski, G. and Narin, F., Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Inf. Proc. Manage., 1976, 12(5), 297–312.

[Price, 1963] Price, D.S., Little Science, Big Science, Columbia University Press, New York, 1963.

[RIN, 2010] Peer review, A guide for researchers, Research Information Network, UK, March 2010. www.rin.ac.uk/our-work/communicating-and-disseminating-research/peer-review-guide-researchers) (accessed at 26.05.2013).

[Roediger, 2006] Roediger III, H.L. The h index in Science: A New Measure of Scholarly Contribution, APS Observer: The Academic Observer, 2006, vol. 19, no. 4.

[SCI, 2013] Science Citation Index http://thomsonreuters.com/products_services/science/science_products/a-z/science_citation_index/#tab1 (accessed at 09.05.2013).

[Seglen, 1997] Per O Seglen "Why the impact factor of journals should not be used for evaluating research". BMJ 1997, 314:498–502.

[Small, 1982] Small, H.. Citation Context Analysis. // In B. Dervin & M.-. Voigt (Eds.), Progress in communication sciences (Vol. 3, pp. 287-310). Norwood: Ablex. 1982.

[Soler, 2007] Jose M. Soler. A Rational Indicator of Scientific Creativity. // Journal of Informetrics, Volume 1, Issue 2, April 2007, Pages 123‑130. http://www.sciencedirect.com/science/article/pii/S1751157706000253. also available as eprint arXiv:physics/0608006v1 [physics.soc-ph]. (accessed at 29.05.2013)

[SRC, 2009] Swedish Research Council, Quality Assessment in Peer Review, Vetenskapsrådet, 2009, ISBN 978-91-7307-190-1, www.cm.se/webbshop_vr/pdfer/2011_01L.pdf (accessed at 26.05.2013)

[STC, 2011] Peer review in scientific publications, Science and Technology Committee, House of Commons, UK, 18 July 2011.

www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/856/85602.htm (accessed at 26.05.2013).

[Teixeira da Silva & Van, 2013] Jaime A. Teixeira da Silva and Pham Thanh Van “The Impact of the Impact Factor®: Survey among Plant Scientists”, Received: 1 October, 2011. Accepted: 29 October, 2012, the Asian and Australasian Journal of Plant Science and Biotechnology©2013 Global Science Books.

[Thelwall, 2006] Mike Thelwall, Liwen Vaughan & Lennart Bjorneborn: Webometrics in: Ann. Rev. Of Information Science & Technology (2006), p. 84, http://www.academia.edu/709949/Webometrics (accessed at 09.05.2013).

[Thompson, 2009] Dennis F. Thompson, Erin C. Callen, Milap C. Nahata, “New Indices in Scholarship. Assessment”, American Journal of Pharmaceutical Education 2009; 73 (6) Article 111. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2769533/pdf/ajpe111.pdf (accessed at 09.05.2013).

[Tunger, 2007] Dirk Tunger, “Webometrics, Informetrics and Bibliometrics – How useful are these Indicators for Measuring Knowledge?”; ESSHRA June, 12-13, 2007 http://www.euresearch.ch/fileadmin/documents/events2007/ESSRHA07/ball_herget_tunger_esshra_june2007.pdf (accessed at 09.05.2013).

Authors' Information

Markov Krassimir *– Institute of Mathematics and Informatics, BAS, Bulgaria; IJ ITA Editor in Chief. e-mail: markov@foibg.com*
Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems

Ivanova Krassimira *– University of National and World Economy, Sofia, Bulgaria; Institute of Mathematics and Informatics, BAS, Bulgaria;*
e-mail: krasy78@mail.bg
Major Fields of Scientific Research: Software Engineering, Business Informatics, Data Mining, Multidimensional multi-layer data structures in self-structured systems

Velychko Vitalii – *Institute of Cybernetics, NASU, Kiev, Ukraine*
e-mail: Velychko@rambler.ru
Major Fields of Scientific Research: Data Mining, Natural Language Processing

STORING RDF GRAPHS USING NL-ADDRESSING[1]

Krassimira Ivanova, Vitalii Velychko, Krassimir Markov

Abstract: *NL-addressing is a possibility to access information using natural language words as addresses of the information stored in the multi-dimensional numbered information spaces. For this purpose the internal encoding of the letters is used to generate corresponded co-ordinates. The tool for working in such style is named OntoArM. Its main principles, functions and using for storing RDF graphs are outlined in this paper.*

Keywords: *NL-addressing, RDF graphs, ontology representations.*

ACM Classification Keywords: *D.4.2 Storage Management; E.2 Data Storage Representations.*

Introduction

Resource Description Framework (RDF) is the W3C recommendation for semantic annotations in the Semantic Web. RDF is a standard syntax for Semantic Web annotations and languages [Klyne & Carroll, 2004].

The underlying structure of any expression in RDF is a collection of triples, each consisting of a ***subject***, a ***predicate*** and an ***object***. A set of such triples is called an ***RDF graph***. This can be illustrated by a node and directed-arc diagram, in which each triple is represented as a node-arc-node link (hence the term "graph") (Fig.1).

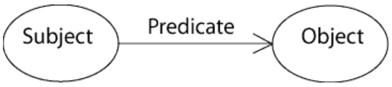


Fig. 1. RDF triple

[1] Reprinted from: G.Setlak, M.Alexandrov, K.Markov (Eds.), Artificial Intelligence Methods and Techniques for Business and Engineering Applications. ITHEA, 2012, Rzeszow, Poland; Sofia, Bulgaria, ISBN: 978-954-16-0057-3 (printed), ISBN: 978-954-16-0058-0 (online). pp. 84-98

Each triple represents a statement of a relationship between the things denoted by the nodes that it links. Each triple has three parts: (1) subject, (2) object, and (3) a predicate (also called a *property*) that denotes a relationship. The direction of the arc is significant: it always points toward the object. The nodes of an RDF graph are its subjects and objects.

The assertion of an RDF triple says that some relationship, indicated by the predicate, holds between the things denoted by subject and object of the triple. The assertion of an RDF graph amounts to asserting all the triples in it, so the meaning of an RDF graph is the conjunction (logical AND) of the statements corresponding to all the triples it contains. A formal account of the meaning of RDF graphs is given in [Hayes, 2004].

The state of the art with respect to existing storage and retrieval technologies for RDF data is given in [Hertel et al, 2009]. Different repositories are imaginable, e.g. main memory, files or databases. RDF schemas and instances can be efficiently accessed and manipulated in main memory. For persistent storage the data can be serialized to files, but for large amounts of data the use of a database management system is more reasonable. Examining currently existing RDF stores we found that they are using relational and object-relational database management systems. Storing RDF data in a relational database requires an appropriate table design. There are different approaches that can be classified in (1) generic schemas, i.e. schemas that do not depend on the ontology, and (2) ontology specific schemas.

In the following we will present a new approach for organizing graph data bases, called Natural Language Addressing (NL-Addressing) and will illustrate it for the most important ontological table designs.

Natural Language Addressing (NL-Addressing)

The idea of Natural Language Addressing (NL-Addressing) is very simple. It is based on the computer internal representation of the word as strings of codes in any system of encoding (ASCII, UNICODE, etc.).

For example, the ASCII encoding of the word „accession” has the next representation: 97 99 99 101 115 115 105 111 110. It may be used as co-ordinate array, which indicates a point in the multidimensional information space, where the corresponded information may be stored.

It is clear, the words have different lengths and, in addition, some phrases may be assumed as single concepts. This means that we need a tool for managing multidimensional information spaces with possibility to support all needed dimensions in one integrated structure.

The independence of dimensionality limitations is very important for developing new intelligent systems aimed to process high-dimensional data. To achieve this, we need information models and corresponding access methods to cross the boundary of the dimensional limitations and to obtain the possibility to work with information spaces with variable and practically unlimited number of dimensions. Such possibility is given by the Multi-Dimensional Information Model (MDIM) [Markov, 2004] and correspond Multi-Dimensional Access Method (MDAM) [Markov, 1984]. Its advantages have been demonstrated in many practical realizations during more than twenty-five years. In recent years, this kind of memory organization has been implemented in the area of intelligent systems memory structuring for several data mining tasks and especially in the area of association rules mining [Mitov et al, 2009]. Here we will show its applicability for organizing of RDF stores.

Multi-dimensional numbered information spaces

Main structures of Multi-Dimensional Information Model (MDIM) are *basic information elements, information spaces, indexes* and *meta-indexes,* and *aggregates*. The definitions of these structures are remembered below.

The ***basic information element*** ($\boldsymbol{BIE}$) of MDIM is an arbitrary long string of machine codes (bytes). When it is necessary, the string may be parceled out by lines. The length of the lines may be variable.

Let ***the universal set*** $\boldsymbol{UBIE}$ be the set of all $\boldsymbol{BIE}$.

Let $\boldsymbol{E}_1$ be a set of basic information elements. Let μ_1 be a function, which defines a biunique correspondence between elements of the set $\boldsymbol{E}_1$ and elements of the set $\boldsymbol{C}_1$ of positive integer numbers, i.e.:

$$\boldsymbol{E}_1 = \{e_i \,|\, e_i \in \boldsymbol{UBIE}\ , i=1,\dots, m_1\},\ \ \boldsymbol{C}_1 = \{c_1 \,|\, c_i \in \boldsymbol{N}, i=1,\dots, m_1\};\ \ \boldsymbol{\mu_1 : E_1 \leftrightarrow C_1}$$

The elements of $\boldsymbol{C}_1$ are said to be numbers (co-ordinates) of the elements of $\boldsymbol{E}_1$.

The triple $\boldsymbol{S_1 = (E_1, \mu_1, C_1)}$ is said to be a ***numbered information space of range 1*** (one-dimensional or one-domain information space).

Let $\boldsymbol{NIS}_1$ be a set of all one-dimensional information spaces.

The triple $\boldsymbol{S_2 = (E_2, \mu_2, C_2)}$ is said to be a ***numbered information space of range 2*** (two-dimensional or multi-domain information space of range two) iff the elements of $\boldsymbol{E}_2$ are numbered information spaces of range one (i.e. belong to the set $\boldsymbol{NIS}_1$) and μ_2 is a function which defines a biunique correspondence between elements of $\boldsymbol{E}_2$ and elements of the set $\boldsymbol{C}_2$ of positive integer numbers, i.e.:

$$\boldsymbol{E}_2 = \{e_i \,|\, e_i \in \boldsymbol{NIS}_1\ , i=1,\dots, m_2\},\ \ \boldsymbol{C}_2 = \{c_i \,|\, c_i \in \boldsymbol{N}, i=1,\dots, m_2\};\ \ \boldsymbol{\mu_2 : E_2 \leftrightarrow C_2}$$

Let $\boldsymbol{NIS}_{n-1}$ be a set of all (n-1)-dimensional information spaces.

The triple $\boldsymbol{S_n = (E_n, \mu_n, C_n)}$ is said to be a ***numbered information space of range n*** (n- dimensional or multi-domain information space) iff the elements of $\boldsymbol{E}_n$ are numbered information spaces of range n-1 (belong to the set $\boldsymbol{NIS}_{n-1}$) and μ_n is a function which defines a biunique correspondence between elements of $\boldsymbol{E}_n$ and elements of the set $\boldsymbol{C}_n$ of positive integer numbers, i.e.:

$$\boldsymbol{E}_n = \{e_j \,|\, e_j \in \boldsymbol{NIS}_{n-1}\ , j=1,\dots, m_n\},\ \ \boldsymbol{C}_n = \{c_j \,|\, c_j \in \boldsymbol{N}, j=1,\dots, m_n\};\ \ \boldsymbol{\mu_n : E_n \leftrightarrow C_n}$$

The information space $\boldsymbol{S}_n$, which contains all information spaces of a given application is called ***information base*** of range **n**. The concept information base without indication of the range is used as generalized concept to denote all available information spaces.

The sequence $\boldsymbol{A} = (c_n, c_{n-1}, \dots, c_1)$, where $c_i \in \boldsymbol{C}_i$, $i=1, \dots, n$ is called ***multidimensional space address*** of range ***n*** of a basic information element. Every space address of range ***m***, ***m < n***, may be extended to space address of range ***n*** by adding leading *n-m* zero codes. Every sequence of space addresses $A_1, A_2, \ldots, A_k$, where ***k*** is arbitrary positive number, is said to be a ***space index***.

Every index may be considered as a basic information element, i.e. as a string, and may be stored in a point of any information space. In such case, it will have a multidimensional space address, which may be pointed in the other indexes, and, this way, we may build a hierarchy of indexes. Therefore, every index, which points only to indexes, is called ***meta-index***.

The approach of representing the interconnections between elements of the information spaces using (hierarchies) of meta-indexes is called ***poly-indexation*.**

Let $G = \{S_i \mid i=1, \ldots, n\}$ be a set of numbered information spaces.

Let $\boldsymbol{\tau} = \{v_{ij} : S_i \rightarrow S_j \mid i=\text{const}, j=1, \ldots, n\}$ be a set of mappings of one "main" numbered information space $\boldsymbol{S}_i \in G \mid i=\text{const}$, into the others $\boldsymbol{S}_J \in G, j=1, \ldots, n$, and, in particular, into itself.

The couple: $D = (G, \boldsymbol{\tau})$ is said to be an "***aggregate***".

It is clear, we can build **m** aggregates using the set G because every information space $\boldsymbol{S}_J \in G, j=1, \ldots, n$, may be chosen to be the main information space.

Operations in the MDIM

After presenting the information structures, we need to remember the operations, which are admissible in the model. In MDIM, we assume that ***all** information elements of **all** information spaces **exist***.

If for any $\boldsymbol{S}_i : \boldsymbol{E}_i = \varnothing \wedge \boldsymbol{C}_i = \varnothing$, than it is called ***empty***.

Usually, most of the information elements and spaces are empty. This is very important for practical realizations.

Because of the rule that all structures exist, we need only two operations with a ***BIE***: updating and getting the value and two service operations: getting the length of a ***BIE*** and positioning in a ***BIE***.

Updating, or simply – ***writing*** the element, has several modifications with obvious meaning: writing as a whole; appending/inserting; cutting/replacing a part; deleting.

There is only one operation for getting the value of a ***BIE***, i.e. ***read*** a portion from a ***BIE*** starting from given position. We may receive the whole ***BIE*** if the starting position is the beginning of ***BIE*** and the length of the portion is equal to the ***BIE*** length.

We have only one operation with a **single space** – *clearing (deleting) the space*, i.e. replacing all ***BIE*** of the space with Ø (empty ***BIE***). After this operation, all ***BIE*** of the space will have zero length. Really, the space is cleared via replacing it with empty space.

We may provide two operations with **two spaces**: (1) *copying* and (2) *moving* the first space in the second. The modifications concern how the ***BIE*** in the recipient space are processed. We may have: copy/move with clearing the recipient space; copy/move with merging the spaces.

The first modifications first clear the recipient space and after that provide a copy or move operation. The second modifications may have two types of processing: destructive or constructive. The ***destructive merging*** may be "conservative" or "alternative". In the conservative approach, the ***BIE*** of recipient space remains in the result if it is with none zero length. In the other approach – the ***BIE*** from donor space remains in the result. In the ***constructive merging*** the result is any composition of the corresponding ***BIE*** of the two spaces.

Of course, the move operation deletes the donor space after the operation.

Special kind of operations concerns the *navigation* in a space. We may receive the space address of the ***next*** or ***previous***, ***empty*** or ***non-empty***, elements of the space starting from any given co-ordinates.

The possibility to count the number of non empty elements of a given space is useful for practical realizations.

Operations with indexes, meta-indexes, and aggregates in the MDIM are based on the classical logical operations – intersection, union, and supplement, but these operations are not so trivial. Because of the complexity of the structure of the information spaces, these operations have two different realizations.

Every information space is built by two sets: the set of co-ordinates and the set of information elements. Because of this, the operations with indexes, meta-indexes, and aggregates may be classified in two main types: (1) operations based only on co-ordinates, regardless of the content of the structures; (2) operations, which take in account the content of the structures:

- The operations based only on the co-ordinates are aimed to support information processing of analytically given information structures. For instance, such structure is the table, which may be represented by an aggregate. Aggregates may be assumed as an extension of the relations in the sense of the model of Codd [Codd, 1970]. The relation may be represented by an aggregate if the aggregation mapping is one-one mapping. Therefore, the aggregate is a more universal structure than the relation and the operations with aggregates include those of relation theory. What is the new is that the mappings of aggregates may be not one-one mappings.
- In the second case, the existence and the content of non empty structures determine the operations, which can be grouped corresponding to the main information structures: elements, spaces, indexes, and meta-indexes. For instance, such operation is the **projection**, which is the analytically given space index of non-empty structures. The projection is given when some coordinates (in arbitrary positions) are fixed and the other coordinates vary for all possible values of coordinates, where non-empty elements exist. Some given values of coordinates may be omitted during processing.

Other operations are transferring from one structure to another, information search, sorting, making reports, generalization, clustering, classification, etc.

OntoArM

The program realization of MDIM is called Multi-Domain Access Method (MDAM). For a long period, it has been used as a basis for organization of various information bases. There exist several realizations of MDAM for different hardware and/or software platforms. The most resent one is the FOI Archive Manager – ArM [Markov et al, 2008]. The newest MDAM realization is called ArM32 (for MS Windows). [Markov, 2004]

The OntoArM is an ontological graph oriented access method but not a middleware in the sense of [Hertel et al, 2009]. It is an upgrade of ArM32.

The OntoArM ontological elements are organized in ontological graph spaces with variable ranges. There is no limit for the ranges of the spaces. Every ontological element may be accessed by a corresponding multidimensional space address (coordinates) given via NL-word or phrase. Therefore, we have two main constructs of the physical organizations of OntoArM – ontological spaces and ontological elements.

In OntoArM the length of the ontological element (string) may vary from 0 up to 1G bytes. There is no limit for the number of strings in an archive but their total length plus internal indexes could not exceed the limited length of the file system for a single file (4G, 8G, etc.). There is no limit for the numbers of files in the information base as well as for theirs dispositions.

OntoArm operations inherited from ArM32

The operations with basic information elements are:

- *ArmRead* (reading a part or a whole element);
- *ArmWrite* (writing a part or a whole element);
- *ArmAppend* (appending a string to an element);
- *ArmInsert* (inserting a string into an element);
- *ArmCut* (removing a part of an element);
- *ArmReplace* (replacing a part of an element);
- *ArmDelete* (deleting an element);
- *ArmLength* (returns the length of the element in bytes).

The operations over the spaces are:

- *ArmDelSpace* (deleting the space),
- *ArmCopySpace* and *ArmMoveSpace* (copying/moving the first space in the second in the frame of one file),
- *ArmExportSpace* (copying one space from one file the other space, which is located in other file).

The operations, aimed to serve the navigation in the information spaces return the space address of the ***next*** or ***previous***, ***empty*** or ***non-empty*** elements of the space starting from any given co-ordinates. They are *ArmNextPresent, ArmPrevPresent, ArmNextEmpty*, and *ArmPrevEmpty*.

The projections’ operations return the space address of the ***next*** or ***previous non-empty*** elements of the projection starting from any given co-ordinates. They are *ArmProjNext* and *ArmProjPrev*.

The operations, which create indexes, are:

- *ArmSpaceIndex* – returns the space index of the non-empty structures in the given information space;
- *ArmProjIndex* – gives the space index of basic information elements of a given projection

The service operations for counting non-empty elements or subspaces are correspondingly:

- *ArmSpaceCount* – returns the number of the non-empty structures in given information space;
- *ArmProjCount* – gives the number of elements of given (hierarchical or arbitrary) projection.

OntoArm RDF graph oriented operations

Converting strings into space addresses

There are two internal operations for conversion:

- *ArmStr2Addr* – converts string to space address. Four ASCII symbols or two UNICODE 16 symbols form one co-ordinate word. This reduces four, respectively – two, times the space’ dimensions. The string is extended with leading zeroes if it is needed.
- *ArmAddr2Str* – converts space address in ASCII or UNICODE string. The leading zeroes are not included in the string.

The operations for conversion are not needed for the end-user because they are used by the upper level operations given below. All OntoArM operations access the information by NL-addresses (given by a NL-words or phrases). Because of this we will not point specially this feature.

OntoArM operations for storing and receiving RDF information

There are two main operations for creating the RDF-store:

- *OntoArmWrite* – writes a buffer (usually NL-string).
- *OntoArmRead* – reads a buffer (usually NL-string).

It is clear; to work easily with RDF graphs, several additional operations are needed:

- *OntoArmAppend (appending a string to an element);*
- *OntoArmInsert (inserting a string into an element);*
- *OntoArmCut (removing a part of an element);*
- *OntoArmReplace (replacing a part of an element);*
- *OntoArmDelete (deleting an element);*
- *OntoArmLength (returns the length of the element in bytes).*

OntoArM operations for graph navigation

The operations, aimed to serve the navigation in the graph are context depended – the format of the elements is important for the navigation. If the element is an NL-index, the navigation operation may take its ***next*** or ***previous*** NL-word for further processing. If the element has more complicated structure, the navigation operations have to be accommodated to it. In general, these operations are usual ones for navigating in the graph structures.

NL-Addressing for ontology generic schemas

Vertical representation

The simplest RDF generic schema is the triple store with only one table required in the database. The table contains three columns named *Subject*, *Predicate* and *Object*, thus reflecting the triple nature of RDF statements. This corresponds to the *vertical representation* for storing objects in a table [Agrawal et al, 2001].

The greatest advantage of this schema is that no restructuring is required if the ontology changes. Adding new classes and properties to the ontology can be realized by a simple INSERT command in the table. On the other hand, performing a query means searching the whole database and queries involving joins become very expensive. Another aspect is that the class hierarchy cannot be modeled in this schema, what makes queries for all instances of a class rather complex [Hertel et al, 2009].

It is easy to store this schema via OntoArM. The *Subject* will be the address and all its couples (*Predicate*, *Object*) may be stored at one and the same address. This way with one operation all arcs of the node of the graph will be received. There exists another variant of organization where the *Predicate* may be additional co-ordinate or name of the archive. In this case, additional operations for reading arcs will be needed. Nevertheless, in all cases the OntoArM will have linear complexity O(max_L), where max_L is the maximal length of the word or phrases, used for NL-addressing. In the same time, the relational table has complexity at least O(n log n), where n is number of all indexed elements (words), if we will take in account supporting indexing and binary search. Of course, the memory for binary indexes exceeds the OntoArM memory for internal indexes. At the end, the time for direct access is many times less then via binary search. The speed experiments with *Firebird* relation data base had showed about 30-ty times for reading and more than 90-ty times for writing in ArM’s favor [Markov et al, 2008].

Normalized triple store

The triple store can be used in its pure form [Oldakowski et al, 2005], but most existing systems add several modifications to improve performance or maintainability. A common approach, the so-called *normalized triple store*, is adding two further tables to store resource URIs and literals separately as shown in Fig. 2, which requires significantly less storage space [Harris & Gibbins, 2003]. Furthermore, a hybrid of the simple and the normalized triple store can be used, allowing storing the values themselves either in the triple table or in the resources table [Jena2, 2012].

Trilpes:

Subject	Predicate	IsLiteral	Object
r1	*r2*	*False*	*r3*
r1	*r4*	*True*	*l1*
…	…	…	…

Resources:

ID	URI
r1	*…#1*
r2	*…#2*
…	…

Literals:

ID	Value
l1	*Value1*
…	…
…	…

Fig. 2. Normalized triple store

In a further refinement, the Triples table can be split horizontally into several tables, each modeling an RDF(S) property:

- SubConcept for the rdfs:subClassOf property, storing the class hierarchy

- SubProperty for the rdfs:subPropertyOf property, storing the property hierarchy
- PropertyDomain for the rdfs:domain property, storing the domains and cardinalities of properties
- PropertyRange for the rdfs:range property, storing the ranges of properties
- ConceptInstances for the rdf:type property, storing class instances
- PropertyInstances for the rdf:type property, storing property instances
- AttributeInstances for the rdf:type property, storing instances of properties with literal values

These tables only need two columns for *Subject* and *Object*. The table names implicitly contain the predicates. This schema separates the ontology schema from its instances, explicitly models class and property hierarchies and distinguishes between class-valued and literal-valued properties [Broekstra, 2005; Gabel et al, 2004].

The normalized triple store is ready for representing via OntoArM. Only what we have to do is to take in account the representing all arcs from a node by one space NL-index and the representing all properties as an aggregate. The *Subject* will be the NL-address and only *Object* will be saved. Possibility to concatenate all *Objects* for a *Subject* reduces the size of memory and time. There are different approaches for building the aggregate – using additional co-ordinate to the *Subjects’* values or to use separate archives for storing the information.

In all cases, the OntoArM has linear complexity O(max_L), the relation data base – at least O(n log n).

NL-Addressing for ontology specific schemas

Horizontal representation

Ontology specific schemas are changing when the ontology changes, i.e. when classes or properties are added or removed. The basic schema consists of one table with one column for the instance ID, one for the class name and one for each property in the ontology. Thus, one row in the table corresponds to one instance. This schema is corresponding to the *horizontal representation*

[Agrawal et al, 2001] and obviously has several drawbacks: large number of columns, high sparsity, inability to handle multi-valued properties and the need to add columns to the table when adding new properties to the ontology, just to name a few.

Horizontally splitting this schema results in the so called one-table-per class schema - one table for each class in the ontology is created. A class table provides columns for all properties whose domain contains this class. This is tending to the classic entity-relationship-model in database design and benefits queries about all attributes and properties of an instance.

However, in this form the schema still lacks the ability to handle multi-valued properties, and properties that do not define an explicit domain must then be included in each table. Furthermore, adding new properties to the ontology again requires restructuring existing tables [Hertel et al, 2009].

The horizontal representation is an example of a set of aggregates in the sense of OntoArM. Storing every class in a separate archive gives possibility to add properties without restructuring existing tables because the aggregate may be described by a meta-index. Again, NL-addressing in OntoArM has linear complexity O(max_L), the relation data base representation – at least O(n log n).

Decomposition storage model

Another approach is vertically splitting the schema, what results in the one-table-per-property schema, also called the *decomposition storage model*.

In this schema one table for each property is created with only two columns for *Subject* and *Object*. RDF(S) properties are also stored in such tables, e.g. the table for rdf:type contains the relationships between instances and their classes.

This approach is reflecting the particular aspect of RDF that properties are not defined inside a class. However, complex queries considering many properties have to perform many joins, and queries for all instances of a class are similarly expensive as in the generic triple schema [Hertel et al, 2009].

In practice, a hybrid schema combining the table-per-class and table-per property schemas is used to benefit from the advantages of both of them. This schema contains one table for each class, only storing there a unique ID for the specific instance. This replaces the modeling of the rdf:type property. For all other properties tables are created as described in the table-per-property approach (Fig. 3) [Pan & Heflin, 2004]. Thus, changes to the ontology do not require changing existing tables, as adding a new class or property results in creating a new table in the database.

ClassA:

ID
…#1
…

Property1:

Subject	Object
…#1	…#3
…	…

ClassB:

ID
…#3
…

Fig. 3. Hybrid schema

A possible modification of this schema is separating the ontology from the instances. In this case, only instances are stored in the tables described above.

Information about the ontology schema is stored separately in four additional tables *Class*, *Property*, *SubClass* and *SubProperty* [Alexaki et al, 2001]. These tables can be further refined storing only the property ID in the Property table and the domain and range of the property in own tables Domain and Range [Broekstra, 2005]. This approach is similar to the refined generic schema, where the ontology is stored the same way and only the storage of instances is different.

To reduce the number of tables, single-valued properties with a literal as range can be stored in the class tables. Adding new attributes would then require changing existing tables. Another variation is to store all class instances in one table called Instances. This is especially useful for ontologies where there is a large number of classes with only few or no instances [Alexaki et al, 2001].

The decomposition storage model is memory and time consuming due to duplicating the information and generation of too much binary search indexes. It is very near to the OntoArM style and may be directly implemented using NL-addressing but this will be not efficient. NL-addressing permits new possibilities due to omitting of explicit given information – names as well as binary indexes. The feature tables may be replaced by NL-addressing access to corresponded points of the information space where all information about given *Subject* will exist. This way we will reduce the needed memory and time. At the end, let point again, that NL-addressing has linear complexity O(max_L) and the relation data base representation – at least O(n log n).

Conclusion

NL-addressing is a possibility to access information using natural language words as addresses of the information stored in the multi-dimensional numbered information spaces. For this purpose the internal encoding of the letters is used to generate corresponded co-ordinates. The tool for working in such style is named OntoArM. Its main principles, functions and using for storing RDF graph were outlined in this paper.

There are further issues not pointed above, which may require an extension of the triple-based schemas and thus are affecting the design of the database: (1) Storing multiple ontologies in one database; (2) Storing statements from multiple documents in one database.

Both points are concerning the aspect of provenance, which means keeping track of the source an RDF statement is coming from. When storing multiple ontologies in one database it should be considered that classes, and consequently the corresponding tables, can have the same name. Therefore, either the tables have to be named with a prefix referring to the source ontology [Pan & Heflin, 2004] or this reference is stored in an additional attribute for every statement. A similar situation arises for storing multiple documents in one database. Especially, when there are contradicting statements it is important to know the source of each statement. Again, an additional attribute denoting the source document helps solving the problem [Pan & Heflin, 2004].

The concept of named graphs [Caroll et al, 2004] is including both issues. The main idea is that each document or ontology is modeled as a graph with a distinct name, mostly a URI. This name is stored as an additional attribute, thus extending RDF statements from triples to so-called quads. For the database schemas described above this means adding a fourth column to the tables and potentially storing the names of all graphs in a further table.

All these problems can be solved by OntoArM, because a separated ontology may be represented in one single archive. In addition, the NL-addressing permits accessing the equal names in different ontologies without any additional indexing or using of pointers, identification and etc. Only the NL-words or phrases are enough to access all information in all existing ontologies (resp. graphs).

The linear complexity O(max_L) of NL-addressing is very important for realizing very large triple stores.

OntoArM is implemented in the Institute of Cybernetics V.M. Glushkov at the National Academy of Sciences of Ukraine, Kiev (IC NASU). It has been used for storing ontology information about multiple documents from own data bases as well as from different internet sources.

The further work is concerned to implementing OntoArM for storing multiple ontologies in the libraries of the "Instrumental Complex with Ontological Purpose", which is under developing in the IC NASU.

Acknowledgements

The paper is partially financed by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA and the Consortium FOI Bulgaria (www.ithea.org, www.foibg.com).

Bibliography

[Agrawal et al, 2001] Agrawal R, Somani A, Xu Y Storage and querying of e-commerce data. In: Proceedings of the 27th Conference on Very Large Data Bases, VLDB 2001,Roma, Italy.

[Alexaki et al, 2001] Alexaki S, Christophides V, Karvounarakis G, Plexousakis D, Tolle K (2001) The ICS-FORTH RDFSuite: Managing voluminous RDF description bases. In: Proceedings of the 2nd International Workshop on the Semantic Web, Hongkong.

[Broekstra, 2005] Broekstra J. Storage, querying and inferencing for Semantic Web languages. PhD Thesis, Vrije Universiteit, Amsterdam (2005).

[Caroll et al, 2004] Caroll J, Bizer C, Hayes P, Stickler P (2004) Semantic Web publishing using named graphs. In: Proceedings of Workshop on Trust, Security, and Reputation on the SemanticWeb, at the 3rd International SemanticWeb Conference, ISWC 2004, Hiroshima, Japan.

[Codd, 1970] Codd, E.: A relation model of data for large shared data banks. Magazine Communications of the ACM, 13/6, 1970, pp.377 387.

[Gabel et al, 2004] Gabel T, Sure Y, Voelker J (2004) KAON – An overview. Insititute AIFB, University of Karlsruhe. http://kaon.semanticweb.org/main kaonOverview.pdf.

[Harris & Gibbins, 2003] Harris S, Gibbins N 3store: Efficient bulk RDF storage. In: Proceedings of the 1st International Workshop on Practical and Scalable Semantic Systems, PSSS 2003, Sanibel Island, FL, USA.

[Hayes, 2004] Patrick Hayes, Editor, *RDF Semantics*, W3C Recommendation, 10 February 2004, http://www.w3.org/TR/2004/REC-rdf-mt-20040210/ . Latest version available at http://www.w3.org/TR/rdf-mt/ .

[Hertel et al, 2009] Alice Hertel, Jeen Broekstra, and Heiner Stuckenschmidt. RDF Storage and Retrieval Systems. In: S. Staab and R. Studer (eds.), Handbook on Ontologies, International Handbooks on Information Systems, DOI 10.1007/978-3-540-92673-3, Springer-Verlag Berlin Heidelberg 2009. pp 489-508.

[Jena2, 2012] Jena2 database interface – database layout. http://jena.sourceforge.net/DB/layout.html. (visited at 22.08.2012)

[Klyne & Carroll, 2004] Graham Klyne and Jeremy J. Carroll, Editors, *Resource Description Framework (RDF): Concepts and Abstract Syntax*, W3C Recommendation, 10 February 2004, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ . Latest version available at http://www.w3.org/TR/rdf-concepts/ .

[Markov et al, 2008] Markov K, Ivanova, K., Mitov, I., & Karastanev, S. Advance of the access methods. Int. J. Information Technologies and Knowledge, 2/2, 2008, pp.123-135

[Markov, 1984] Кr.Markov. A Multi-domain Access Method. // Proceedings of the International Conference on Computer Based Scientific Research. Plovdiv, 1984. pp. 558-563.

[Markov, 2004] Markov, K. Multi-domain information model. Int. J. Information Theories and Applications, 11/4, 2004, pp.303-308.

[Mitov et al, 2009] Mitov, I., Ivanova, K., Markov, K., Velychko, V., Vanhoof. K., Stanchev, P. "PaGaNe" – A classification machine learning system based on the multidimensional numbered information spaces. In World Scientific Proc. Series on Computer Engineering and Information Science, No.2, pp.279 286.

[Oldakowski et al, 2005] Oldakowski R, Bizer C, Westphal D RAP: RDF API for PHP. In: Proceedings of Workshop on Scripting for the Semantic Web, SFSW 2005, at 2nd European Semantic Web Conference, ESWC 2005, Heraklion, Greece.

[Pan & Heflin, 2004] Pan Z, Heflin J (2004) DLDB: Extending relational databases to support Semantic Web queries. Technical Report LU-CSE-04-006, Department of Computer Science and Engineering, Lehigh University.

Authors' Information

***Krassimira Ivanova** – University of National and World Economy, Sofia, Bulgaria*
e-mail: krasy78@mail.bg
Major Fields of Scientific Research: Data Mining

***Vitalii Velychko** – Institute of Cybernetics, NASU, Kiev, Ukraine*
e-mail: Velychko@rambler.ru
Major Fields of Scientific Research: Data Mining, Natural Language Processing

***Krassimir Markov** – Institute of Mathematics and Informatics at BAS, Sofia, Bulgaria;*
e-mail: markov@foibg.com
Major Fields of Scientific Research: Multi-dimensional information systems, Data Mining

TABLE OF CONTENTS OF IJ ITA VOL. 29, NO.1

TABLE OF CONTENTS OF IJ ITA VOL. 29, NO.2

TABLE OF CONTENTS OF IJ ITA VOL. 29, NO.3

TABLE OF CONTENTS OF IJ ITA VOL. 29, NO.4