UDC 004.514.62+004.048
DOI: 10.15587/2312-8372.2017.103177

Dubinsky A.

DEVELOPMENT OF PROTOTYPE FOR USER INTERFACE OF INFORMATION SYSTEM

Запропоновано структуру інформаційної системи для зменшення інформаційного перевантаження користувача при роботі з великим числом джерел документів з інтернет. Дан короткий список призначених для користувача історій. Отримання документів відбувається за допомогою RSS-каналів. Інтерфейс системи виконаний як інтелектуальна карта (mindmap). Вибірка документів для вузлів карти виконується на основі правил, записаних на SQL-подібній мові запитів.

Ключові слова: *діаграма зв'язків, mindmap, ієрархічна кластеризація, інформаційне перевантаження, інтерфейс користувача.*

1. Introduction

The amount of information created and processed by the humankind is continuously growing. This phenomenon was named «the information explosion». It was realized as an actual problem in the last century. At that time, it was a problem for comparatively small communities of scientists and researchers. Today the constant work with wide flows of information from open sources is necessary for any intellectual activity.

Currently, the level of Internet penetration is high enough both in developed and in developing countries [1]. The main part of information flows passes through the Internet. In the last decade, a significant part of Internet channels is built on the «Web 2.0» ideology. The audience of Facebook, Twitter, and other social networks is growing all over the world. Private individuals, entrepreneurs, business, public organizations and government structures use these networks.

The Internet network was initially created for using links between HTML-pages. The ability to browsing through links often leads to significant losses of working time. Advanced intellectual workers normally avoid unnecessary web surfing.

Web sites and thematic channels of Web 2.0 often show advertisements, «paid» posts and other unwanted content. Context or banner adverts are added by companies which own the info environment (social networks, search engines, etc.).

Viewing of additional non-targeted information results in unproductive time consuming, drawing of user's attention and increasing fatigue. Cutting out irrelevant pages will reduce the time lost and increase productivity. In order to avoid showing irrelevant documents, first, we need to automatically evaluate the relevance (usefulness) of documents.

2. The object of research and its technological audit

The object of research is the interface and structure of the information system, which is designed to pre-filter information flow on topics specified by the user.

Existing information systems can be divided into several main types. The first one is the search engines [2]: Google, Yandex, Bing, Baidu, etc. They scan all of the Internet sources, perform a search in their index at the user's request. These systems are free and available online. The second type is the Web 2.0 networks: Facebook, Vk, Qzone, Twitter, etc. They work only with the user content placed inside the network. These systems are free as well, but user authentication is required for the full access. The third one is represented by professional systems for monitoring of Internet sources and document analysis. A survey of monitoring systems is given in [3]. The document analysis systems are described in [4]. Search engines and social networks have hundreds of millions of active users [5]. The professional system has up to tens of thousands installations.

Let's go to the list of the main disadvantages of abovementioned types of information systems. Search engines: only text interface, a large list of search results, insufficiently effective filtering of irrelevant results. Web 2.0 systems: poor search capabilities, operating with internal content only, insufficient accessibility of content beyond the system. Professional systems: rather high costs, complexity of customization, weakness of the built-in API (Application Programming Interface).

3. The aim and objectives of research

The main aim of research is to determine the type of information system interface. This interface should simplify and facilitate the user's work.

To achieve this aim it is necessary to accomplish the following tasks:

1. Define the structure of the information system. The components of the system will be based on ready open source software solutions.
2. Define the transformation of the mathematical graph of the hierarchy of document clusters.
3. Describe the list of requirements for the user interface.
4. Present the implementation of a user interface, which meets the stated requirements.

4. Research of existing solutions of the problem

To protect the user from information overload, first of all, it is necessary to evaluate the relevance of the set of proposed documents. It should be done avoiding showing this document set to the user. The next step is to separate documents into clusters having similar topics. After that, to generate the report on the set of documents cluster. Finally, to represent the results of clustering in an appropriate form.

The key task is the evaluation of document relevance and the degree of its compliance with the user search query. This task is solved by existing information retrieval systems. The HTML meta tags are created for simple metadata processing. However, without content analysis of the text, they do not protect against the dishonesty actions from the Internet site creators.

The first generation of content methods for relevance evaluation is based on the linguistic and statistical characteristics of the texts [2]. The next generation of methods uses the analysis of an array of hyperlinks of the web [6]. These methods have allowed improving the quality of search when a larger amount of sources is available. The creators of Internet sites use a wide arsenal of search engine optimization methods, for the higher ranking place in the Search Engine Result Page [7].

The search engines can evaluate the content relevancy based on the user behavior [8]. Analysis of user behavior and the identification of similar user's preferences allowed to create the collaborative filtering methods [9]. Modern search engines use a complex combination of these methods to improve the quality of search results. The latest generation of methods for computing of document relevance is based on semantic methods [10].

One of the most exciting issues of the work of modern social networks is the algorithm of the newsfeed formation by default. Usually, this algorithm is the secret and it is known only in general. The most interesting thing is the algorithm for making Facebook news [11] because this network has the largest audience of users [5].

The known solution is preloading documents before showing them to the user [12]. It was proposed in the early years of the Internet development, when the main problem was the waiting time. A good survey of the clustering text documents methods is given in [13]. The clustering method based on the user information request is described in [14]. Therefore, the main task is to develop a user-friendly interface of the information system.

5. Methods of research

Let's use the constructive approach of the theory of systems, so that the structure of the system is formed by the set of functional parts of the system. Therefore, the structure of the system is determined by the necessary target result. According to the number of elements of the system, it may be classified as a simple system. To solve the research task, there is no need to formulate and analyze the mathematical model of the system. Determination of the optimal system structure is out of the scope of this paper.

During the information system design, it is also necessary to determine the deployment scheme. The prospective solution is the using of microservices architecture [15] in the cloud. However, the simple prototype can be created as a monolithic web-based solution.

During software development, it is necessary to create the set of user interface requirements. Let's make the detailed descriptions of the user's functional requirements. At this level, there is no need to use the formal notation or UML diagrams. The methods for working with user's requirements are described in the Software Engineering Body of Knowledge (SWEBOK), also known as ISO/IEC TR 19759:2005 standard.

The proposed prototype of the user interface is based on mind maps method of visualization. The initial view of multiple documents can be obtained using automatically methods of hierarchical clustering. Mathematical apparatus of graph theory help to makes the transition from the document clusters hierarchy to the connected network of map nodes.

6. Research results

6.1. Structure of the information system. The structure of the information system is shown in Fig. 1.

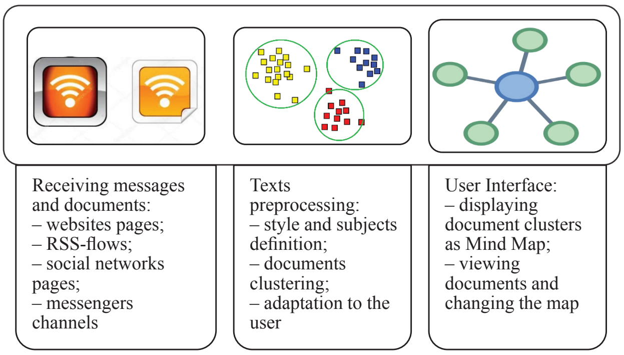

Fig. 1. The structure of the information system

The first block provides an interface with the sources of documents or messages. These sources are sites, social networks and messengers.

The most convenient mechanism for delivering the frequently updated information to users is the RSS standards family. Unfortunately, the largest Web 2.0 environments, such as Facebook and Twitter, do not support external processing of the users content. Therefore, in a long time, it is impossible to guarantee a steady receipt of updates of information from such social networks. There are solutions by third-party services, but they also do not guarantee full functionality after new updating of the content access policy. Therefore, the only available palliative solution is creation of a separate application for the each social network. However, an information division to the several windows by the origin source will be inconvenient for users.

Recently, online messengers have gained considerable popularity. Business accounts, open API, the wide opportunity of creating bots and a high level of user satisfaction allow creating many channels with thousands of subscribers. Viber, WhatsApp, Telegram, and others have become new effective channels for information dissemination.

Today almost all professional media sources are supporting several distribution channels. Therefore, in most cases, the information system can receive the information from the sources.

The task of the second block is preliminary selection of the most relevant messages (documents). It can be done by document clustering. The mathematical methods of clustering objects have been developed for more than 50 years [13]. The actual way of clustering is an adaptive approach based on neural network models. First, let's note the method [16] for automatic clustering of web documents by relevance to the information needs, using a hybrid neural network. A good survey of the approaches of using neural networks for constructing clusters is available in [17]. So the fine-tuning of the neural network parameters is needed for efficiently clustering for different initial data.

The third block of the information system provides a user interface for viewing documents or messages clusters. The idea of using effective visualization for convenient display of large arrays of documents is well-known. For example, in [18] it is proposed to use Kohonen self-organizing maps to graphical display the category map in a conceptual space for navigating the user in the set of search engine results. The using of conceptual maps for processing of knowledge and information is surveyed in [19]. In [20], the usage of MindMap-visualization as an interface for documents collections is described.

6.2. The hierarchy graph of document clusters transformation. The result of the hierarchical clustering of a set of documents is a connected acyclic graph – the binary rooted tree G_2. Tree leafs are separate documents. The internal vertexes are the clusters. The root represents the all entire set of documents.

A binary tree is convenient for computer processing. However, it does not allow getting a compact clusters visualization on the screen. It is well known that a human can effectively interact with only a small number of objects. This value is in the range of 7 ± 2 (the Yngve-Miller's number). The specific value may vary for different people. So, the binary tree G_2 must be transformed into the n-ary tree G_n before rendering to the screen. Let the user opportunity to specify the convenient value of n.

The depth of the tree should also be limited for the same considerations.

We will use the hierarchical clustering method, which generates a balanced tree. Then the initial number of documents should not exceed the value of n^n. Therefore, we ignored this restriction, since a stronger restriction on the documents quantity due to the computational complexity $O(n^2)$ of hierarchical clustering methods.

The graph clusters transformation to the n-ary tree is easily carried out for $n=2^2$ or $n=2^3$. The number of hierarchy levels is reduced by 2 or 3 times, respectively. If we select another value of n, several vertexes are drawn to one vertex in the new tree. In the initial binary tree, those nodes occupied several different levels. So for $n=6$, two vertexes of level i+1, four at level i+2 and four vertexes on i+3 level will be contracted to one vertex of level i (Fig. 2).

The number of documents is a random value, so it is usually not a power of two. Therefore, the original binary tree cannot be perfectly balanced. The selection of the drawn branches will be made so that the result tree gets more balanced.

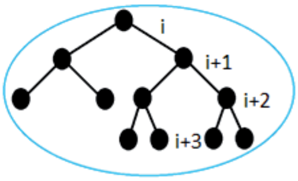

Fig. 2. The set of graph vertexes that is drawn into one vertex, for $n = 6$

6.3. The user interface. Development of an effective interface begins with the creation of user scenarios (use case). The information system will be used regularly to review information messages of different subjects. New topics (clusters) will usually be stored for repeated using time after time. Therefore, the main user stories are: viewing new documents and messages from this cluster. Other sets of user stories: viewing the documents and messages, configuring the clusters visualization and setting the view of map. This is the list of the main user stories:

1. Viewing new documents and messages:
– selecting the cluster and showing the list of documents;
– checking source updates;
– marking and exporting the selected document;
– adding comments to the selected document.
2. Configuring the selected map node:
– viewing the list of sources;
– viewing the tag cloud, and cluster keywords;
– editing the rules for documents selection for the current cluster;
– renaming the map node, adding comments to the node.
3. Changing the nodes and map links:
– splitting the cluster into several ones, adding child nodes;
– merging selected clusters into the one;
– changing the placing of map nodes.

The prototype of the information system interface is shown in Fig. 3. This prototype made by JavaScript free source code from [21].

First of all, the user selects the map branch for viewing. All the child nodes in this active branch are displayed. Each node shows (in square brackets) the number of new, not scanned documents. In the rest map branches, the nodes are hidden and the number of documents is not visible.

The other user stories will start through the context menu by left clicking the mouse.

There is no one-to-one correspondence between the automatically generated clusters set and the set of map nodes, determined by the user. Therefore, we need an opportunity for fine-tuning of map visualization. In particular, the way to determine the rules for clusters creation. So, the user will be able to specify the combination of clusters and select their contents via the set of conditions (Fig. 3).

Use of the SQL language subset sing the subset is proposed for setting the rules of selection. In the SELECT statement, instead of the list of fields (table columns), let's write the identifiers of clusters and/or data sources. In the FROM let's write the identifier of the parent node in the map tree. In the WHERE, let's define the correspondence of a set of attributes, for example, the substring for searching in the text document. In the ORDER BY let's define the kind of sorting documents.

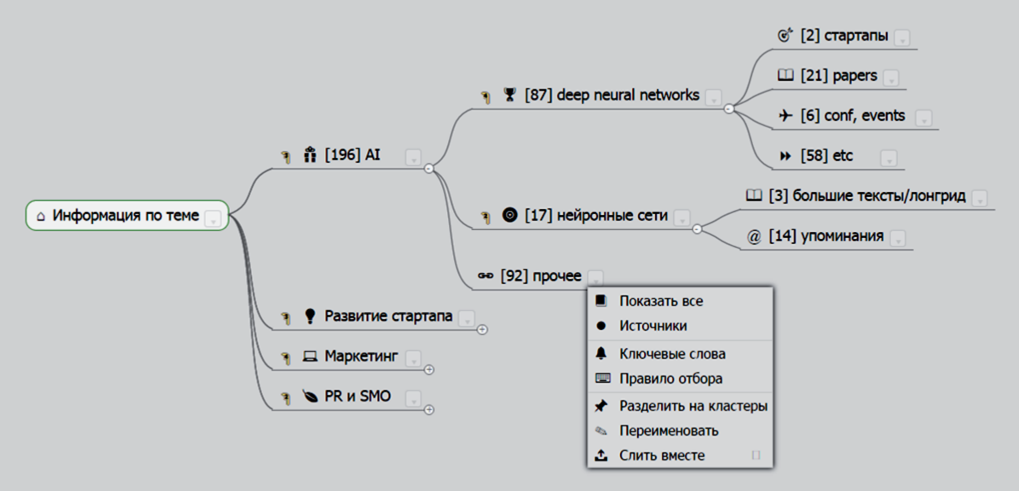

Fig. 3. The prototype of mind map user interface

The last child node will collect all documents, which were not added to any others child nodes by rules. In Fig. 3 such nodes are «etc», «mentions» and «other». For example, the node «other» includes all documents selected in the «AI» branch, but not selected according to the rules of the nodes «deep neural networks» and «neural networks». The user can set the name of the nodes. The system will offer the default node name based on automatically formulated tags derived from the document content.

6.4. The tasks for further research. The next stage of research is the experiment series of tests with sets of real data. At this time, there is no need to use all possible types of sources. So it is necessary to use only RSS feeds processing. Let's use one of the open sources free RSS readers, which is distributed free under GNU General Public License. After the test results, the list of user stories will be improved and redefined.

In addition, it is necessary to find the basic (the most frequently used) templates for cluster creating rules defined by users. They should be used for transmitting the automatically generated clusters to nodes on a visual map.

In the future, it is necessary to adopt the elements of machine learning for facilitating the adjustment of the system to a specific user's needs.

Tracking of user behavior will allow dividing the documents into following subsets:

a) documents selected by the user for detailed study;

b) documents that aren't interested for user;

c) documents that are closed very quickly after viewing.

There is no simple solution for defining and brief recording the patterns or rules for separation into these subsets. The sets of these documents subsets can use for neural network training. Then after some number of training cycles, this neural network will be able to predict the relevance of new documents. This will allow the sorting new documents by relevance or automatically create a cluster with the most interesting content.

7. SWOT analysis of research results

Strengths. The visual interface for the system is a two-dimensional plane instead of a linear list. It will allow to:

1) reduce the time spent looking at irrelevant documents;

2) increase the information retrieval completeness, since the user can view more documents that are necessary for the same fixed time;

3) the above will improve the quality of the analytical reports and management decisions made.

The structure of the information system involves the use of existing solutions as parts of the system. This will significantly shorten the time for creation and the costs of software development. Therefore, it significantly reduces the risks of project failure.

Rules for clusters remapping to the mind map nodes will be written in the SQL dialect that allows to flexibly configuring the knowledge map that is visible on the screen. The user will be able to give the map a more convenient appearance, which will reduce its fatigue and increase the processing speed of information.

The created map can be saved. One-time setup and regular re-use save users time and effort.

Weaknesses. The main weaknesses of the proposed solution are:

- It takes the user much time to explore a new interface, learn how to write down SQL rules, and so on.
- The automatic clustering system needs initial setting. The cold start problem: adaptation of the system to user by machine learning methods gives a result only after time.
- The solution is created for intellectual workers who are not specialists in monitoring information. They will have to overcome the cognitive barrier and lack of motivation for starting using this system.

Opportunities. The main opportunities for further improvement are in the Web 2.0 approach. Thousands of other people will use the results of the first flow of users. First of all, we are talking about identifying patterns of strong selection SQL rules. The best rules, turned out to be convenient and effective, will be used by default by new users of the system.

The mind maps are successfully used for learning and knowledge transferring and delivering. A map compiled by a qualified expert will be very useful to other users.

An additional opportunity: to expand the area of application of mind map concept to other areas of intellectual activity, in particular, to manage corporate knowledge, improve the skills of employees, etc.

Threats. The main threats include:

1. Popularity of AR and VR interfaces (Virtual Reality, Augmented Reality), which it is necessary for solving other task. The two-dimensional interface of the proposed system may seem as «not progressive enough».

2. Lack of information culture. The new generation of people with its «clip-on thinking» may not be able to make efforts to systematize information flows.

3. Search engines and social networks can increase access restrictions to content. Then it will be impossible to receive updates of many important sources and the purpose of the system will become unattainable.

8. Conclusions

1. This article presents the structure of the information system, which is designed to improve the processing efficiency of large information flows. The system consists of three parts, which implements such tasks: obtaining documents from sources, clustering documents, displaying clusters with a mind map.

2. The transformation of the hierarchy graph of document clusters is determined. It allows displaying the set of documents in a user-friendly form.

3. The set of user interface requirements is created. This set of simple user cases can be implemented in the software product. These user stories describe working with documents, setting up the selected map node and changing the map.

4. The implementation of the user interface is proposed. The prototype interface is written in JavaScript and is available as a webpage on the user's device screen – computer, pad or smartphone.

References

1. Econ Stats: All Economic Indicators for All Countries [Electronic resource] // Economy Watch. – Available at: \www/URL: http://www.economywatch.com/economic-statistics/economic-indicators/. – 20.03.2017.
2. Manning, C. D. Introduction to Information Retrieval [Text] / C. D. Manning, P. Raghavan, H. Schutze. – Cambridge: Cambridge University Press, 2008. – 482 p. doi:10.1017/cbo9780511809071
3. Tanatar, N. V. Intellektual'nye poiskovo-analiticheskie sistemy monitoringa SMI [Text] / N. V. Tanatar, A. G. Fedorchuk // Biblioteki Natsional'nyh akademii nauk: problemy funktsionirovaniia, tendentsii razvitiia. – 2008. – Vol. 6. – P. 205–219.
4. Lande, D. V. Instrumentarii analitika [Text] / D. V. Lande // Telekom. – 2010. – № 4. – P. 36–40.
5. Smith, C. How Many People Use the Top Social Media, Apps & Services? [Electronic resource] // C. Smith // DMR. – 23.05.2017. – Available at: \www/URL: http://expandedramblings.com/?s=How+Many+People+Use+the+Top+Social+Media%2C+Apps+%26+Services%3F+
6. Brin, S. The anatomy of a large-scale hypertextual Web search engine [Text] / S. Brin, L. Page // Computer Networks and ISDN Systems. – 1998. – Vol. 30, № 1-7. – P. 107–117. doi:10.1016/s0169-7552(98)00110-x
7. Enge, E. The Art of SEO: Mastering Search Engine Optimization (Theory in Practice) [Text] / E. Enge, S. Spencer, R. Fishkin, J. Stricchiola. – O'Reilly Media, 2009. – 608 p.
8. Liu, Y. BrowseRank [Text] / Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, H. Li // Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR'08. – 2008. – P. 451–458. doi:10.1145/1390334.1390412
9. Su, X. A Survey of Collaborative Filtering Techniques [Text] / X. Su, T. M. Khoshgoftaar // Advances in Artificial Intelligence. – 2009. – Vol. 2009. – P. 1–19. doi:10.1155/2009/421425
10. Turdakov, D. Texterra: A Framework for Text Analysis [Text] / D. Turdakov, N. Astrakhantsev, Y. Nedumov, A. Sysoev, I. Andrianov, V. Mayorov, D. Fedorenko, A. Korshunov, S. Kuznetsov // Proceedings of the Institute for System Programming of RAS. – 2014. – Vol. 26, № 1. – P. 421–438. doi:10.15514/ispras-2014-26(1)-18
11. DeVito, M. A. From Editors to Algorithms [Text] / M. A. DeVito // Digital Journalism. – 2016. – P. 1–21. doi:10.1080/21670811.2016.1178592
12. Wang, Z. Prefetching in World Wide Web [Text] / Z. Wang, J. Crowcroft // Proceedings of GLOBECOM'96. 1996 IEEE Global Telecommunications Conference. – 1996. – P. 28–32. doi:10.1109/glocom.1996.586110
13. Aggarwal, C. C. A Survey of Text Clustering Algorithms [Text] / C. C. Aggarwal, C. X. Zhai // Mining Text Data. – Springer US, 2012. – P. 77–128. doi:10.1007/978-1-4614-3223-4_4
14. Method for the clusterization of a set of objects by using a reference [Electronic resource]: Patent of Ukraine № 72720, MPK G06F 7/00, G06F 17/30, G06F 7/16 / Dubinsky A. – Appl. № 20031212875; Filed 29.12.2003; Publ. 15.03.2005, Bull. № 3. – Available at: \www/URL: http://uapatents.com/9-72720-sposib-klasterizuvannya-naboru-obehktiv-z-vikoristannyam-zrazka.html
15. Lewis, J. Microservices: a definition of this new architectural term [Electronic resource] / J. Lewis, M. Fowler // Martin Fowler. – 25 March 2014. – Available at: \www/URL: http://martinfowler.com/articles/microservices.html
16. Khan, M. S. Web document clustering using a hybrid neural network [Text] / M. S. Khan, S. W. Khor // Applied Soft Computing. – 2004. – Vol. 4, № 4. – P. 423–432. doi:10.1016/j.asoc.2004.02.003
17. Du, K.-L. Clustering: A neural network approach [Text] / K.-L. Du // Neural Networks. – 2010. –Vol. 23, № 1. – P. 89–107. doi:10.1016/j.neunet.2009.08.007
18. Chen, H. Internet browsing and searching: User evaluations of category map and concept space techniques [Text] / H. Chen, A. L. Houston, R. R. Sewell, B. R. Schatz // Journal of the American Society for Information Science. – 1998. – Vol. 49, № 7. – P. 582–603. doi:10.1002/(sici)1097-4571(19980515)49:7<582::aid-asi2>3.0.co;2-x
19. Canas, A. J. A Summary of Literature Pertaining to the Use of Concept Mapping Techniques and Technologies for Education and Performance Support [Text]: Report to the Chief of Naval Education and Training / A. J. Canas, J. W. Coffey, M. J. Carnot, P. Feltovich, R. R. Hoffman, J. Feltovich, J. D. Novak. – Pensacola, Florida: The Institute for Human and Machine Cognition, 2003. – 108 p.
20. Spangler, S. MindMap: utilizing multiple taxonomies and visualization to understand a document collection [Text] / S. Spangler, J. T. Kreulen, J. Lessler // Proceedings of the 35th Annual Hawaii International Conference on System Sciences. – 2002. – P. 1170–1179. doi:10.1109/hicss.2002.994039
21. Karta uma – dlia teh, kto izuchaet Javascript [Electronic resource] // GitHub, Inc. – Available at: \www/URL: https://github.com/Imater/mindmap

РАЗРАБОТКА ПРОТОТИПА ИНТЕРФЕЙСА ПОЛЬЗОВАТЕЛЯ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Предложена структура информационной системы для уменьшения информационной перегрузки пользователя при работе с большим числом источников документов из интернет. Дан краткий список пользовательских историй. Получение документов происходит с помощью RSS-каналов. Интерфейс системы выполнен как интеллектуальная карта (mindmap). Выборка документов для узлов карты выполняется на основе правил, записанных на SQL-подобном языке запросов.

Ключевые слова: диаграмма связей, mindmap, иерархическая кластеризация, информационная перегрузка, интерфейс пользователя.

Dubinsky Alexey, *PhD, Associate Professor, Department of Medical-biological Physics and Informatics, State Establishment «Dnipropetrovsk Medical Academy», Dnipro, Ukraine, e-mail: dubinsky@ukr.net, ORCID: http://orcid.org/0000-0001-8536-1603*