ITHEA

International Journal INFORMATION TECHNOLOGIES & KNOWLEDGE

2009 Volume 3 Number 1

International Journal
INFORMATION TECHNOLOGIES & KNOWLEDGE
Volume 3 / 2009, Number 1

IJ ITK is official publisher of the scientific papers of the members of the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITK as well as the **subscription fees** are given on ***www.ithea.org***
The camera-ready copy of the paper should be received by e-mail: info@foibg.com.
Responsibility for papers published in IJ ITK belongs to authors.

General Sponsor of IJ ITK is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Vol.3, Number 1, 2009

Edited by the **Institute of Information Theories and Applications FOI ITHEA®**, Bulgaria,
in collaboration with the **V.M.Glushkov Institute of Cybernetics of NAS, Ukraine**,
and the **Institute of Mathematics and Informatics, BAS, Bulgaria.**

Publisher: **ITHEA®**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Printed in Bulgaria

ISSN 1313-0455 (printed) **ISSN 1313-048X (online)** **ISSN 1313-0501 (CD/DVD)**

THE COMBINED APPROACH TO PRESENTATION OF MULTIMEDIA CONTENT AND TEXTUAL ANNOTATION

Dmytro Nochevnov

Abstract: *The problem of a semantic gap between multimedia content and its textual annotation defined by user is urgent. This paper presents the combined approach to presentation of multimedia semantics which is based on aggregation of several multimedia that are similar by content and textual annotations into the classes. Such classes will contain generalized descriptions of objects and links reflected in the multimedia, and keywords from some thesaurus. To create and update these classes we also describe the hierarchical clustering and machine learning operations. This approach should expand the multimedia search and navigating area due to including the media documents with similar content and text descriptions.*

Keywords: *multimedia data mining, content based multimedia retrieval, text based multimedia retrieval.*

ACM Classification Keywords: *H.3.1 Content Analysis and Indexing - Abstracting methods. H.5.1 Multimedia Information Systems - Evaluation/methodology.*

Introduction

Interest to area of multimedia handling is stipulated by WWW fast growth and arising need of fast media data indexing and search among the big and heterogeneous corpus of documents. Multimedia content producing and using became a usual practice due to existence of effective tools of multimedia indexing and metadata defining. Such metadata can be in machine-readable text format and in the form of visual and audio descriptors (for example in format MPEG-7), and help to handle of multimedia by search engines and intellectual agents [Stamou and Kollias, 2005]. However, the problem of a semantic gap between low level multimedia adjectives, such as color, texture, the scene, etc., and high level concepts like «a mountain landscape», «a misbehavior of the people», which human uses in textual annotation, is still unsolved. Overcoming this gap is one of the Multimedia Data Mining tasks [Stamou and Kollias, 2005], [Petrushin and Khan, 2007]. This issue can be found in traditional multimedia retrieval methods which are usually being divided into two parts [Vihrovsky and Ignatenko, 2006], [Goodrum, 2000]:

1. Search on the basis of textual annotation (Text Based Multimedia Retrieval). Given retrieval methods use high level information as keywords of some thesaurus or textual annotation.
2. Content-Based Multimedia Retrieval which is based on using the low level information such as texture, visual, audio or other kind of patterns in the request.

The main problem of Text Based Retrieval is impracticability of exact and complete multimedia textual annotation which different users can identify in various manners according to their personal experiences and knowledge [Goodrum, 2000]. One of solutions is using the Content-Based Retrieval based on the template of the requested multimedia. However, in some cases such information is insufficient to automatically investigate the information interesting of a seeker, for example the exact time of presented event, the creator of multimedia, objects «behind a frame», three-dimensional relations between the presented objects etc. Also, there is the retrieval precision decreasing which is caused by homonymy, when search results consist of multimedia with different content and similar textual annotations. In turn, retrieval recall decreasing is caused by synonymy which is inversed to homonymy.

Therefore, in our opinion, it is reasonable to combine multimedia low level descriptors and its high level textual annotation for indexing and retrieval [Goodrum, 2000]. It should make more effective multimedia handling and retrieval due to including the information about other multimedia with similar content and textual annotations.

1. Combined Model of Multimedia Knowledge Presentation

It is known that most of the media data contain visual and sound tracks of physical objects written by sensors [Stamou and Kollias, 2005]. If there were several objects of this kind, the information about their space relations (in case of images and video) or time relations (in case of audio and video data) was saved. To mark these objects we can use the existing ontologies or thesauri of objects' names (for example Getty's Art and Architecture Thesaurus [AAT], which consists more than 120000 terms of the art, architecture and other cultural objects, or Library of Congress Thesaurus of Graphic Materials [LCTGM]) to select appropriate terms and save them together with typical objects' metadata. The similar approach to multimedia presentation can be found in [Petridis et al., 2004] where Semantic Web ontology was used to storage of MPEG-7 descriptors. A first order logic can be used to formalize the links between objects.

All of this will allow to automate detection of objects and links between them during indexing of multimedia, and also to expand a search area by engaging of faithful objects.

1.1. Multimedia Content Model

One of approaches to the multimedia presenting is using a semantic net [Petridis et al., 2004], [Dance et al., 1996]. Following it, we will use a directed graph to model objects and links that are reflected in multimedia **m**, and its textual annotation. Such model should consist of:

1) vertex set $v_i \in V_m$, which presents the reflected in multimedia **m** objects; each vertex is associated with some keyword d(v_i) from the thesaurus; named by this keyword objects should be most similar to the represented object by MPEG-7 descriptors' values;

2) edge set $a_{jk} \in A_m$, which corresponds to real links between the reflected in multimedia **m** objects; each edge is associated with predicate $p(a_{jk})$;

3) keyword set T_m, which is defined manually or automatically by indexing of the multimedia textual annotation t_m.

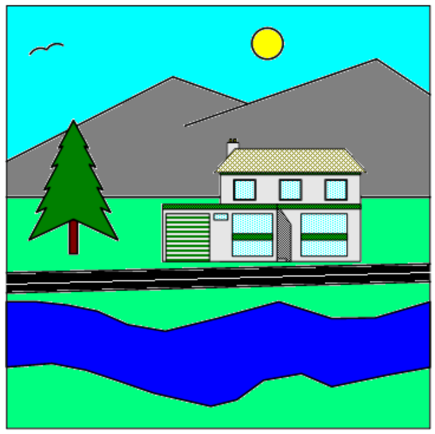
Each keyword d can be associated with one or more vertices of one or more multimedia content models. The same type of link designated by the predicate p, also can be presented in one or more multimedia content models.

Example 1. Let's create a model of image «*Landscape with house in Carpathian Mts*» (see fig.1a). According to the previous definition this multimedia content model $\mathbf{m_1}$ may consist of the next sets (see fig.1b):

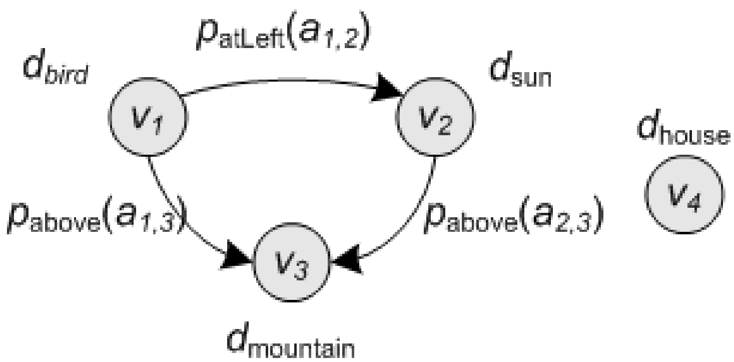
$V_m = \{v_1 \Rightarrow$ MPEG-7("bird"), $v_2 \Rightarrow$ MPEG-7("sun"), $v_3 \Rightarrow$ MPEG-7("mountain"), $v_4 \Rightarrow$ MPEG-7("house")};

$A_m = \{a_{1,2} \Rightarrow p_{atLeft}(a_{1,2}),\ a_{1,3} \Rightarrow p_{above}(a_{1,3}),\ a_{2,3} \Rightarrow p_{above}(a_{2,3})\}$;

T_m = {"landscape", "house", "Carpathian Mts"}.



a) image «Landscape with house in Carpathian Mts»



b) fragment of multimedia content model

Figure 1. Example of multimedia content model $\mathbf{m_1}$

1.2. Multimedia Class Model

For expanding a search area we propose to aggregate the semantically similar multimedia $m_n \in M_c$ into classes $\boldsymbol{c}$; such class will include the generalized information about typical objects and links reflected in multimedia, and keywords from their text annotations.

Let's define a multimedia class as directed graph containing:

1) weighted vertex set $v_i \in V_c$, which describe objects reflected in united multimedia $m_n \in M_c$; each vertex is associated with some keyword $d(v_i)$ from the thesaurus of objects' names, and with weight $w(v_i)$; this weight designates the number of objects in set of united multimedia M_c;

2) weighted edge set $a_{jk} \in A_c$, which describe links between objects reflected in multimedia $m_n \in M_c$; each edge corresponds to some predicate $p(a_{jk})$, and to weight $w(a_{jk})$; this weight designates the number of links in set M_c;

3) textual annotation of a multimedia class T_c, defined by indexing of textual annotations t_m of multimedia $m_n \in M_c$; set T_c consists of weighted keywords; each weight $w(t_i)$ designates the number of keywords within the t_m.

Example 2. Let's show an example of multimedia class model c_1 with single multimedia m_1 from previous example 1 (see fig. 2):

V_c = { {$v_1 \Rightarrow$ MPEG-7("bird"), $w(v_1) = 1$},

{$v_2 \Rightarrow$ MPEG-7("sun"), $w(v_2) = 1$},

{$v_3 \Rightarrow$ MPEG-7("mountain"), $w(v_3) = 1$},

{$v_4 \Rightarrow$ MPEG-7("house"), $w(v_4) = 1$} };

A_c = { {$a_{1,2} \Rightarrow p_{\text{atLeft}}(a_{1,2})$, $w(a_{1,2}) = 1$},

{$a_{1,3} \Rightarrow p_{\text{above}}(a_{1,3})$, $w(a_{1,3}) = 1$},

{$a_{2,3} \Rightarrow p_{\text{above}}(a_{2,3})$, $w(a_{2,3}) = 1$} };

T_c = { {t_1 = "landscape", $w(t_1) = 1$},

{t_2 = "house", $w(t_2) = 1$},

{t_3 = "Carpathian Mts", $w(t_3) = 1$} }.

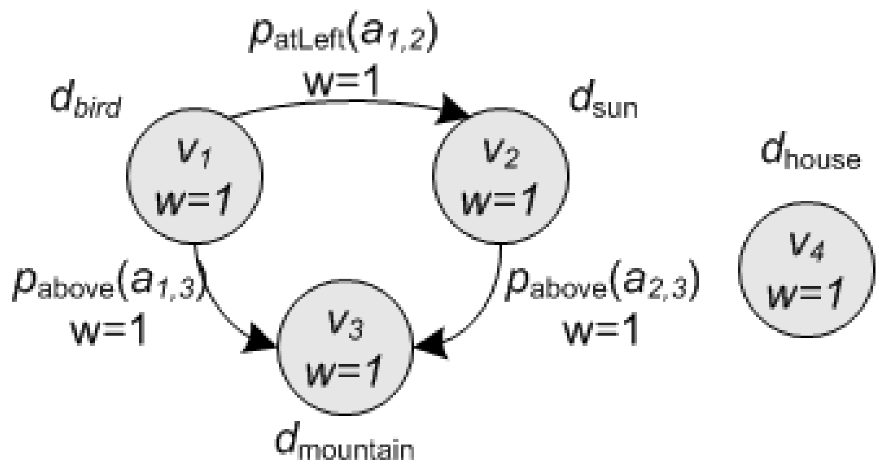


Figure 2. Example of multimedia class model c_1.

2. Multimedia Handling Sequence

We suggest the next operations of multimedia processing:

1. Indexing of a new multimedia and making the multimedia content model *m* in four steps:

a) automatic recognition of objects reflected in multimedia and search the closest by MPEG-7 descriptors values object and take their keywords *d* from the thesaurus of objects;

b) automatic recognition of objects' links and search the appropriate predicates *p*;

c) parse textual annotation t_m to define T_m;

d) aggregate d_m, p_m, T_m into the multimedia content model.

2. Selecting the semantically closest class c_i and updating this one by data about new multimedia; otherwise adding a new class consisting of only new the multimedia.

3. Checking the quality of multimedia class partition and, if necessary, reclustering the semantically homogeneous classes, or all multimedia in corpus.

4. Taking information about the closest by content and annotation multimedia from the selected classes to improve the multimedia retrieval.

The semantically closest class c_i or set of classes ***C*** can be selected manually by user through navigation in the multimedia knowledge base, or automatically by analyzing the distances between multimedia *m* and existing classes *h*(*m*, *c*) by next formula:

$$c = c_k \text{, if } h(m, c_k) = \min(h(m, c_k)) \le h_{\max}, i = \overline{1, N_c} \text{ ,}$$

where N_c - the number of multimedia classes;

h(*m*, *c*) – the distance between multimedia and class; will be defined in (1) below.

3. Distances between Multimedia Classes and Instances

Let's define a method of distance between classes and instances of multimedia evaluation. Taking in mind the structure of multimedia content and class models offered above, we will use the similarity measure based on vector of keywords [Ozkarahan, 1986].

Similarity and distance functions can be described according to the [Duran et al., 1974]:

Definition 1. *Similarity function* is nonnegative real function $z(x, y)$ which satisfies the following properties:

1) $0 \le z(x, y) < 1$ for $x \ne y$,

2) $z(x, x) = 1$,

3) $z(x, y) = z(y, x)$.

Definition 2. *Distance function* is nonnegative real function $h(y, x) = 1 - z(y, x)$.

To evaluate semantically close multimedia class it is necessary to define the distance between a class and an instance of multimedia $h(m, c)$. In this case we will take into account the concordance of objects in models, links between them, keywords in textual annotations and their weights:

$$h(m,c) = 1 - z(m,c) = 1 - \frac{\lambda_v \cdot z_v(m,c) + \lambda_a \cdot z_a(m,c) + \lambda_t \cdot z_t(m,c)}{\lambda_v + \lambda_a + \lambda_t}, \qquad (1)$$

where $\lambda_v, \lambda_a, \lambda_t \in [0,1]$ - the coefficients of impact, which are defined heuristically;

$$z_v(m,c) = \left(1 - \frac{\sum_{\forall v \in (V_m \cap V_c)} n(w_c(v))}{card(V_m \cap V_c)}\right) \cdot \frac{card(V_m \cap V_c)}{card(V_m \cup V_c)}$$ - a similarity by structure of reflected objects;

$$z_a(m,c) = \left(1 - \frac{\sum_{\forall a \in (A_m \cap A_c)} n(w_c(a))}{card(A_m \cap A_c)}\right) \cdot \frac{card(A_m \cap A_c)}{card(A_m \cup A_c)}$$ - a similarity by structure of links between objects;

$$z_t(m,c) = \left(1 - \frac{\sum_{\forall t \in (T_m \cap T_c)} n(w_c(t))}{card(T_m \cap T_c)}\right) \cdot \frac{card(T_m \cap T_c)}{card(T_m \cup T_c)}$$ - a similarity by multimedia textual annotations,

$n(x) = \left|0.1 \cdot \lg(1 + x)\right|_{\bmod 1}$ - function for normalizing the weights;

$\lambda_v = 0, \text{if } card(V_m \cup V_c) = 0$; $\lambda_a = 0, \text{if } card(A_m \cup A_c) = 0$; $\lambda_t = 0, \text{if } card(T_m \cup T_c) = 0$.

To aggregate the similar multimedia in a class it is necessary to evaluate the distance between separate instances of multimedia $h(m_k, m_l)$. To solve this distance we will also taking into account a concordance of objects and links reflected in the multimedia m_k, m_l, and coincidence of keywords form their text annotations:

$$h(m_k, m_l) = 1 - z(m_k, m_l) = 1 - \frac{\lambda_v \cdot z_v(m_k, m_l) + \lambda_a \cdot z_a(m_k, m_l) + \lambda_t \cdot z_t(m_k, m_l)}{\lambda_v + \lambda_a + \lambda_t}, \qquad (2)$$

where $z_v(m_k, m_l) = \dfrac{card(V_{m_k} \cap V_{m_l})}{card(V_{m_k} \cup V_{m_l})}$ - a similarity by structure of reflected objects,

$z_a(m_k, m_l) = \dfrac{card(A_{m_k} \cap A_{m_l})}{card(A_{m_k} \cup A_{m_l})}$ - a similarity by structure of links between objects,

$z_t(m_k, m_l) = \dfrac{card(T_{m_k} \cap T_{m_l})}{card(T_{m_k} \cup T_{m_l})}$ - a similarity by multimedia textual annotations,

$\lambda_v = 0, \text{if } card(V_{m_k} \cup V_{m_l}) = 0$; $\lambda_a = 0, \text{if } card(A_{m_k} \cup A_{m_l}) = 0$; $\lambda_t = 0, \text{if } card(T_{m_k} \cup T_{m_l}) = 0$.

To define the quality of partition of multimedia set by classes may be used the distance between multimedia classes $h(c_i, c_j)$. We will evaluate it in consideration of the class objects concordance, links between them, keywords and their weights:

$$h(c_i, c_j) = 1 - z(c_i, c_j) = 1 - \frac{\lambda_v \cdot z_v(c_i, c_j) + \lambda_a \cdot z_a(c_i, c_j) + \lambda_t \cdot z_t(c_i, c_j)}{\lambda_v + \lambda_a + \lambda_t}, \qquad (3)$$

where $z_v(c_i, c_j) = \left(1 - \frac{\sum\limits_{\forall v \in (V_{c_i} \cap V_{c_j})} n(|w_{c_i}(v) - w_{c_j}(v)|)}{card(V_{c_i} \cap V_{c_j})} \right) \cdot \frac{card(V_{c_i} \cap V_{c_j})}{card(V_{c_i} \cup V_{c_j})}$ - a similarity by structure of reflected objects;

$$z_a(c_i, c_j) = \left(1 - \frac{\sum\limits_{\forall a \in (A_{c_i} \cap A_{c_j})} n(|w_{c_i}(a) - w_{c_j}(a)|)}{card(A_{c_i} \cap A_{c_j})} \right) \cdot \frac{card(A_{c_i} \cap A_{c_j})}{card(A_{c_i} \cup A_{c_j})}$$ - a similarity by structure of links;

$$z_t(c_i, c_j) = \left(1 - \frac{\sum\limits_{\forall t \in (T_{c_i} \cap T_{c_j})} n(|w_{c_i}(t) - w_{c_j}(t)|)}{card(T_{c_i} \cap T_{c_j})} \right) \cdot \frac{card(T_{c_i} \cap T_{c_j})}{card(T_{c_i} \cup T_{c_j})}$$ - a similarity by textual annotations;

$\lambda_v = 0, \text{if } card(V_{c_i} \cup V_{c_j}) = 0$; $\lambda_a = 0, \text{if } card(A_{c_i} \cup A_{c_j}) = 0$; $\lambda_t = 0, \text{if } card(T_{c_i} \cup T_{c_j}) = 0$.

4. Multimedia Class Updating

For adding the multimedia ***m*** into the class ***c*** it is necessary to update the multimedia class model using the machine learning. During this operation the only weights vertices, edges and keywords of class, which is duplicated in the multimedia ***m*** should be increased in the class. According to this rule we should:

1) modify the objects' weights $w_c(v_i)$ from formula:

$$\forall v_i \in V' = V_m \cap V_c \rightarrow w_c(v_i) = w_c(v_i) + 1;$$

2) modify the links' weights $w_c(a_i)$:

$$\forall a_i \in A' = A_m \cap A_c \rightarrow w_c(a_i) = w_c(a_i) + 1;$$

3) modify the keywords' weights $w_c(t_i)$:

$$\forall t_i \in T' = T_m \cap T_c \rightarrow w_c(t_i) = w_c(t_i) + 1;$$

4) add information about non-duplicated objects, links and keywords into the class ***c*** with the weight equal to 1:

$$c = c + (m \setminus c) = \{V_c = V_c + V_m \setminus V_c, A_c = A_c + A_m \setminus A_c, T_c = T_c + T_m \setminus T_c\}.$$

In case of a semantically closed class absence, the information about new multimedia can be added in the knowledge base as a new class with weights = 1.

5. Multimedia Clustering

For automatic multimedia classification and partition into the classes we propose to use an *agglomerative hierarchical method of clustering* [Pedrycz, 2005]. According to this method we will start clustering with unique clusters for each multimedia, and then unite these initial clusters into the new clusters and form two-level hierarchical structure, in which:

- **on first level:** clusters C_i with the minimum distance between clusters h_{min}=1;
- **on second level:** subclusters C_{ij} with the h_{min} <1.

To solve h_{min} you can use expression (2) or (3).

We will break up of multimedia into second level subclusters iteratively by changing h_{min} from 0 to 1, until the appropriate quality will be achieved. For evaluating this quality we suggest to measure the intrinsic homogeneity of a cluster η_o and the measure of clusters' heterogeneousness η_m:

$$\eta_{o_i} = \begin{cases} \dfrac{2}{N_i(N_i-1)} \sum_{j=1}^{N_i} \sum_{k=j+1}^{N_i} h(m_j, m_k), \text{if } N_i > 1 \\ 1, \text{otherwise} \end{cases}$$

, where N_i – the number of multimedia united in cluster C_i;

$$\eta_m = \begin{cases} \dfrac{2}{N_C(N_C-1)} \sum_{j=1}^{N_C} \sum_{k=j+1}^{N_C} h(m_c^{\langle C_j \rangle}, m_c^{\langle C_k \rangle}), \text{if } N_C > 1 \\ 0, \text{otherwise} \end{cases};$$

where $m_c^{\langle C_i \rangle}$ - central multimedia of cluster C_i, with minimum distances to other multimedia of a class: $\sum_j h(m_j, m_c) = \min, \forall m_j \in C_i$;

N_C – the number of parsed clusters;

$h(m_k, m_l)$ was defined in (2) above.

After clustering multimedia sets from each cluster $m_i \in C_j$ is aggregated into the classes c_i by using operation of multimedia class updating described in section 4 of this paper.

Conclusion

While multimedia documents producing and using is became a usual practice today, there is an unsolved problem of the semantic gap between multimedia content and its text descriptions defined by users. To solve this issue in this paper was presented the combined approach to multimedia indexing and retrieval which unite low level descriptors of multimedia and its high-level text annotations into one model. Such approach to presentation of multimedia semantics is based on aggregation of multimedia, that are similar by content and text annotation, into the classes, containing generalized descriptions of objects and links reflected in multimedia, and keywords from some thesaurus. Also we described the hierarchical clustering and machine learning operations which are used to create and update these classes.

Such approach should allow expanding a multimedia search and navigating area due to including of the media documents with a similar content and textual description.

Bibliography

[AAT] Getty's Art and Architecture Thesaurus. - http://www.getty.edu/research/conducting_research/vocabularies/aat/

[Dance et al., 1996] Dance S., Caelli T., Liu Zhi-Qiang. Picture Interpretation: A Symbolic Approach//World Scientific Series In Machine Perception And Artificial Intelligence; Vol. 20 - 1996.

[Duran et al., 1974] Benjamin S. Duran, Patrick L. Odell. Cluster analysis. A survey. - Springer-Verlag. Berlin-Heidelberg-New York, 1974.

[Goodrum, 2000] Goodrum A. A. Image information retrieval: An overview of current research // Informing Science, 3(2):P.63-66, February 2000.

[LCTGM] Thesaurus for Graphic Materials // Library of Congress. - http://www.loc.gov/rr/print/tgm1/

[MPEG-7] MPEG-7 Overview. <http://www.mpeg-7.com> (Industry Focus Group)

[Ozkarahan, 1986] Ozkarahan E. Database machines and database management. - Prentice-Hall, Inc, 1986. - 636p.

[Pedrycz, 2005] Pedrycz W. Knowledge-based clustering: From Data to Information Granules. - John Wiley & Sons, 2005. - 316p.

[Petridis et al., 2004] Petridis K., Kompatsiaris I., Strintzis M.G., Bloehdorn S., Handschuh S., Staab S., Simou N., Tzouvaras V., Avrithis Y. Knowledge representation for semantic multimedia. Content analysis and reasoning. - EWIMT, 2004.

[Petrushin and Khan, 2007] Petrushin V. A. and Khan L. (Eds). Multimedia Data Mining and Knowledge Discovery. - Springer-Verlag London Limited, 2007. – 521 p.

[Stamou and Kollias, 2005] Stamou G. and Kollias S. (Eds). Multimedia Content and the Semantic Web. - John Wiley & Sons Ltd, 2005 - 392 p.

[Vihrovsky and Ignatenko, 2006] Vihrovsky K, Ignatenko A. Using MPEG-7 for visual data classification and retrieval//Internet magazine "the Drawing and Multimedia". - http://cgm.graphicon.ru/content/view/161/61/

Authors' Information

Dmytro Nochevnov *– Associate professor, "Information Technologies of Design" department of Cherkasy State Technological University, Cherkasy, Ukraine, 18006.*

Major Fields of Scientific Research: Data Mining and Knowledge Discovery, Education on Information Technologies.

E-mail: dmitry.ndp@gmail.com.