Indexing and Comparison of Multi-Dimensional Entities in a Recommender System based on Ontological Approach

Maxim Bakaev and Tatiana Avdeenko

Novosibirsk State Technical University, Novosibirsk, Russia

maxis81@gmail.com, tavdeenko@mail.ru

Abstract. The paper describes an application of indexing—the technology currently widely used for processing and comparing textual information—to multi-dimensional entities of knowledge domains. We propose a model for building a frame-based ontology, which contains a domain conceptual model as well as a controlled vocabulary of "base terms" used for indexing. Further, the ontology constitutes the structure for the knowledge base of the recommender system developed by us, whose task is to support human-computer interaction in web applications. The system automatically represents the interaction task being solved as a structured set of base terms, and compares it with the pre-indexed design guidelines representing practical knowledge of the domain. The interaction task context is defined by input data: 1) semi-structured attributes of target users and 2) natural-language requirements for a particular web application. The former are processed mostly via production model rules stored in the knowledge base, while the requirement text is mined for base terms from the controlled vocabulary. As a result of the comparison, the system provides a set of guidelines relevant for a particular interaction task context, seeking to save work effort of interface designers. Also, the proposed approach for indexing multi-dimensional entities can be applied in various recommender and knowledge-based systems.

Keywords. Intellectual system, design guidelines, data indexing, frame ontology.

Indexación y comparación de entidades multidimensionales en un sistema de recomendación basado en el enfoque ontológico

Resumen. Este artículo describe una aplicación de indexación —la tecnología que se usa ampliamente hoy en día para procesar y comparar la información textual— a entidades multidimensionales de dominios de conocimiento. Se propone un modelo para el desarrollo de una ontología basada en marcos (frames). La ontología contiene un modelo conceptual del dominio y un vocabulario controlado de "términos básicos" usados para la indexación. También, la ontología sirve como la estructura para la base de conocimiento del sistema de recomendación desarrollado en este trabajo. El objetivo del sistema de recomendación es apoyar a la interacción "hombre-computadora" en aplicaciones web. El sistema representa automáticamente una tarea dada de interacción como un conjunto estructurado de términos básicos, y lo compara con las instrucciones del diseño indexadas previamente, las cuales representan el conocimiento práctico del dominio. El contexto de la tarea de interacción se define por los datos de entrada: 1) los atributos semi-estructurados de usuarios objetivo, y 2) requerimientos en lenguaje natural para una aplicación web seleccionada. Los atributos se procesan prácticamente mediante las reglas de producción del modelo, y el texto de requerimientos se usa para recuperar los términos básicos del vocabulario controlado. Como el resultado de la comparación el sistema genera un conjunto de instrucciones relevantes para un contexto de una tarea dada de interacción. El objetivo es ahorrar el esfuerzo de trabajo de diseñadores de interfaces. El modelo propuesto para indexar entidades multidimensionales se puede aplicar también en varios sistemas de recomendación y sistemas basados en conocimiento.

Palabras clave. Sistema intelectual, instrucciones del diseño, indexación de datos, ontología de marcos (frames).

1 Introduction

In modern information environment, there is a constant increase of amount and complexity of data being stored and used. The emergence and universal adoption of the Internet, in which most of the resources are hypertextual, led to rapid

development of methods and technologies for natural language processing: summarization, indexing and search, translation, etc. In particular, automatic assessment of text similarity is a highly practical problem in operations with documents, short phrases or search queries [1]. One of the technologies for solving this problem is semantic analysis (e.g., [2]) which includes indexing defined as a description of a text with "index", a set of special terms extracted from the text or taken from a constrained (controlled) dictionary.

In general, any information retrieval system carries out indexing, formulation of query (user's information needs specified in a language understood by the system) and its comparison with the available (indexed) information. Then, the index can be formed based on the following structures [3, p.44]:

1. *Bag of words*. It is a set of unrelated terms (sometimes also called tags) that describe a certain object or information resource. Bag of words indexing/classification is currently widely applied to multidimensional objects such as audio, video or images, in data stores and recommender systems (a review is available in [4]).
2. *Taxonomy*. When terms describing a domain form a hierarchy of categories, a taxonomy is a structure that generally has high clarity and is easy to comprehend due to the fact that only one semantic relation is used in such representation, that is, "parent – child". However, such choice of the relation imposes certain limitations.
3. *Thesaurus*. It is a collection of terms and word combinations grouped into units named concepts. They are organized either hierarchically or with semantic (associative) relations. The chosen relations form a pre-defined and fixed set which generally includes such relations as "parent – child", "part – whole", "cause – effect", or linguistic relationships.
4. *Ontology*. When a domain (field of knowledge) is formally described with concepts (classes), their attributes, relations between them, application axioms, and constraints, this description forms an ontology. Thus, ontologies are more flexible than thesauri since an ontology includes any kind of semantic relations, and at the same time permits a more detailed domain specification due to attributes, constraints, axioms, etc.

Since the 1990s, ontologies have been applied in Information Science for knowledge representation, management and integration; they are the key element in the Semantic Web concept. Currently, the following types of ontologies are identified based on their purpose [3, p.47]: upper ontologies, domain-specific, and task-specific ontologies. The former aim to describe universal knowledge or codify the use of language (e.g., ontology specification language); perhaps, one of the most prominent examples is *CYC*, a common sense knowledge ontology. The scope of domain-specific ontologies is a certain domain of knowledge (such as *Systematized Nomenclature of Medicine – Clinical Terms* in Medicine), while task-specific ontologies are even more concrete and generally built for a particular application.

In our work, we propose an ontology-based approach to indexing complex multi-dimensional entities in knowledge domains, allowing both their classification and similarity estimation. These tasks are important, in particular, in the development of intellectual systems, which still contain (1) a large amount of poorly organized knowledge or (2) interpret the problem in consideration based on incomplete or unstructured input data. In order to deal with these two difficulties, we have developed a recommender system to support human-computer interaction (HCI) engineering for web applications.

Among the deficiencies of the existing systems such as MetroWeb (see an application description in [5] or Bore [6]), is the lack of matching between a particular design context and design guidelines or patterns which constitute the practical knowledge in the domain (see a review in [4]). In this paper, we propose an approach for indexing and comparing multi-dimensional entities, implemented in a recommender system based on a frame ontology which contains both a conceptual model of the domain and a controlled vocabulary of "base terms" used for indexing. The input data for the system are attributes of target

users for the web application and requirements in natural language, the latter are analyzed by a system component designed to fulfill this task. The system forms its interpretation of the problem being solved as an ordered set of base terms (index), which may be called a "query" into the knowledge base (KB), where similarly indexed guidelines are stored. The output for the system are problem-relevant guidelines and automatically generated interface wireframe.

2 Method

The frame ontology developed in this work can be formally represented as

$$O_F =< C, R, S, G, T, D_S, D_G, E > \ , \qquad (1)$$

where C is a finite non-empty set of frame-classes describing domain concepts;

R are binary relationships defined for classes, $R \subseteq C \times C$, $R = \{R_{ISA}\} \cup R_{ASS}$ with R_{ISA} being the hierarchical relationship and R_{ASS}, a set of other associative relationships;

S is a finite set of possible slots (class attributes);

G is a finite set of facets (slot attributes);

E is a finite set of instances (objects created from classes);

T is a finite non-empty set constituting the controlled vocabulary of domain terms, built on the set of base terms B that correspond to the names of the classes:

$$T = \bigcup_{i=1}^{n} T_i, T_i = \{b_i\} \cup Eq(b_i), \bigcap_{i=1}^{n} T_i = \varnothing .$$

Eq(bi) are synonymous terms, each of which is related to the base term $bi \in B$, D_S is a finite set of possible slot types, D_G is a finite set of possible facet types.

The frame-class structure is defined as

$$c =< Name_C, (isa \ \ c_{parent}), (s_1, s_2, ...s_{n(c)}) > ,$$

where $c, c_{parent} \in C$ is a parent frame-class, which is in R_{ISA} relation with the current class; $s_i \in S$ are frame slots; $Name_c \in B$ is the name of a class, which is also the base term from the controlled vocabulary T. The hierarchy of frames is formed by specifying relationships and the involved classes, such as c_{parent} for "is-a" relation.

All the classes are divided into two sets, $C = C_{abstract} \cup C_{concrete}$. For the set of concrete classes, $C_{concrete}$, it is possible to define instances (concrete objects), $e \in E$. The structure of the frame-instance is similar to the structure of class C, from which it is constructed, for example:

$$e(c) =< Name_e, (s_1^e, s_2^e, ..., s_{n(c)}^e) > ,$$

where $s_1^e, s_2^e, ..., s_{n(c)}^e$ are instances of the class c slots, filled with concrete values. The structure of the frame-slot is the following:

$$s_C =< Name_{S,C}, (gs_1, gs_2, ...gs_{k(S,C)}) > ,$$

where $s_c \in S$ is the slot of class c, $gs_i \in G$ is the slot facets, $Name_{s,c}$ is the slot name.

Associative relationships (from R_{ASS}) are created by explicitly setting the value of a frame slot as another frame, and describing the type of the relationship between the frames. To implement the associative relationships, D_S includes not only simple types D_{SS} (*symbol*, *string*, *float*, etc.), but also D_{class} (the allowed value is frame-class) and $D_{instance}$ (the allowed value is frame-instance):

$$D_S = D_{SS} \cup \{D_{class}\} \cup \{D_{instance}\} .$$

The relationships of $D_{instance}$ type are common in data modelling, but using frame-classes (D_{class}) as values opens new possibilities together with the vocabulary of terms. Since every frame-class corresponds to a base term, their set can be used to index textual or other multi-dimensional domain entities. For example, if a guideline g_i stored in the system's knowledge base has the following text: "The website logo must have hyperlink to the homepage, except on the homepage itself", the corresponding frame-classes constituting the set $T(gi) \subseteq B$ would be *Logo*, *Homepage*, *Hyperlink*.

The context of a particular interaction problem is described in the same way, but a set of classes ($Pr \in B$) is formed automatically by the system from target user attributes and requirements. By comparing these two sets, the system can define the degree of each guideline's similarity (relevance) for a concrete web interaction, by marking out the common terms $B(gi) \subseteq B$:

$$\mathrm{Pr} \cap T(g_i) = B(g_i), \left| B(g_i) \right| = k_i \ . \qquad (2)$$

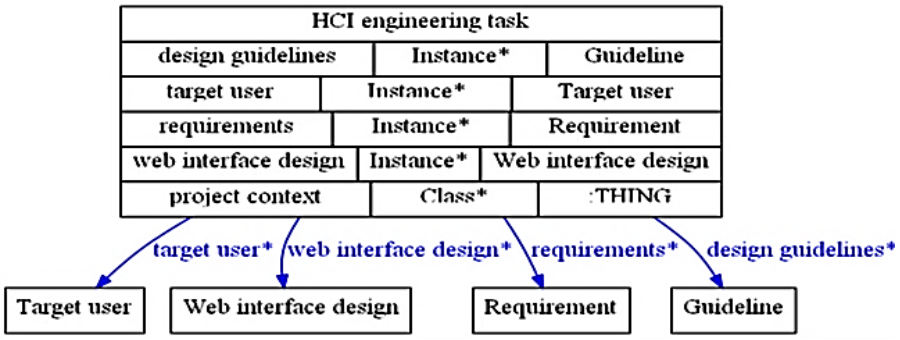

Fig. 1. HCI engineering task class structure and relationships

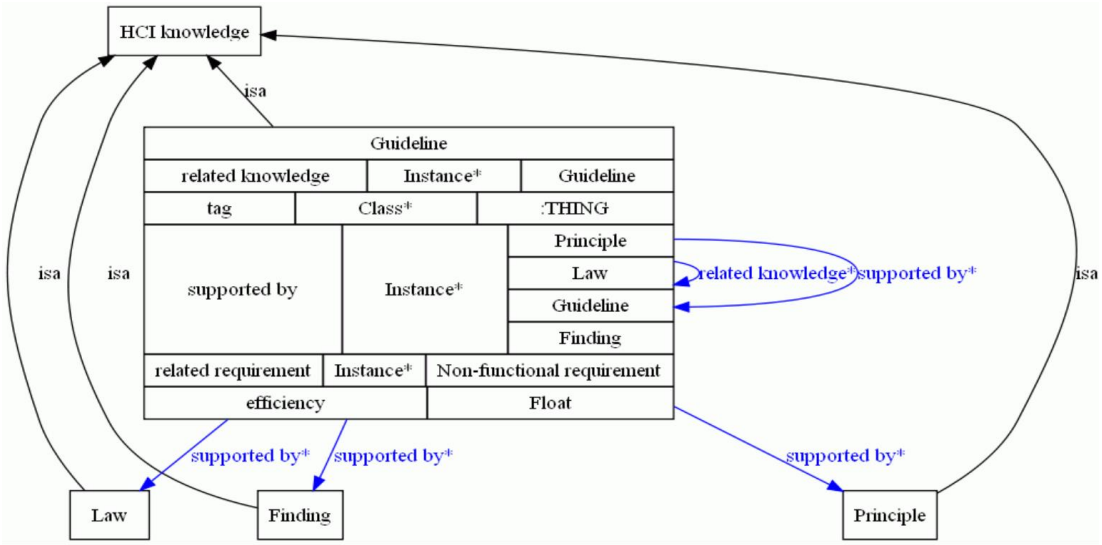

Fig. 2. Guideline class structure and relationships

Correspondingly, for the system output, only guidelines with $k_i>0$ are selected.

3 Application

Based on the above model, we constructed an HCI engineering support domain ontology, which included more than 150 classes overall. The first class was *HCI engineering task* with relationships to classes that represent input data: *Target user* and *Requirement*, and output data: *Guideline* and *Web interface design*. The possible values of the slot *project context* are any class of the ontology, thus any concrete HCI engineering problem can be described (indexed) with a set of selected base terms ($Pr \in B$). The structure of an *HCI engineering task* class and its relationships in ontology are shown in Fig. 1.

To represent practical knowledge existing in a HCI engineering domain, we used the class *Guideline* which has relationships *supported by* with classes that correspond to other tiers of domain knowledge: *Law*, *Principle* and *Finding*. The values of the tag slot of the class *Guideline* are

any class of the ontology, thus any guideline g_i can also be indexed with a set of selected base terms ($T(gi) \subseteq B$), see Fig. 2.

In the system, the instances of *Guideline* class are recommendations in HCI engineering, for the following types of web applications: 1) Web applications in general, 2) e-commerce, 3) e-government sites, and 4) web sites of educational institutions. The structure of the knowledge base was derived from the ontology, while its content includes instances and production model rules (in the form "*if*... *then*..."). The knowledge base is the main component of the recommendation system developed and put online, together with additional web interface component (at http://clips.vgroup.su).

The developed system was applied for solving a number of real domain problems, in particular, for designing the official website for the *People's Faculty* of *Novosibirsk State Technical University*. The Faculty provides "computer literacy" courses for senior citizens, so target users of the website were defined as older people with low experience in IT, seeking to obtain information about the courses. The website functional and non-functional requirements in natural language were analyzed by the system which extracted the ontology base terms from texts. After the system applied the corresponding production model rules, the project context was formed as the set of the following terms:

Website, Web page, Website element, Website service/section, Target user, Target emotion requirement, Older user, Accessibility requirement, Error rate, Size, Color, Aesthetic impression, Trust impression, IT experience, Website experience, Contacts, Services, Web page element, Requirement, Functional requirement, Non-functional requirement, Interface quality requirement, Design requirement, Interface design property, Content design property, Guideline, Interface element, Web form element, Interface design, Web interface design, Interface quality metric, Forum, CMS / back-office, User registration, Search, Success rate, Homepage, News events, About us, Sitemap, Help, Departments facilities, Faculty / staff, Alumni employment, Majors / programs, Admission, Curriculum, Tutorials.

The system then automatically generated the web interface wireframe and the ordered set of guidelines relevant to the project context, some of them are presented in Table 1 as an example (the base terms that matched in comparison are in bold).

Table 1. A subset of the system output: selected relevant design guidelines

The guideline text	Index (base terms)
Visited hyperlinks must have different color, otherwise older people can easily forget, which section of the website they've already visited.	**Color**; Hyperlink; Navigation; **Older user**
If graphics and video are used, the size and quality must be good enough for older people's perception. The color combinations used in images must ensure sufficient contrast.	Imagery; Media object; **Interface element; Older user; Accessibility requirement**
Interface elements size must be no less than 8 px (at 1024x768, 17''), otherwise the error rate increases dramatically for older people.	**Size**; **Error rate**; **Older user**; **Interface element**
Web forms must avoid using standard html elements in a non-standard way. Older users highly rely on their previous experience and have hard time recovering from errors.	Web form; **Web form element**; **Error rate**; **Older user**

4 Experiments

To assess the effectiveness of the KBS application for web interface design, we fulfilled an empirical research exploring whether the website developed with the system's support yields better usability for its target user group, senior users.

The above mentioned website, designed and implemented for the *People's Faculty* of *NSTU*, was used in our experiments, together with 5

other websites. In total, six websites were used in the experiments. *Websites # 1, # 2, # 3, and # 4 were third-party sites* selected on the basis of being representative of small or medium e-commerce sites, or possibly having seniors as one of target user groups. Website #5 was also a real company's website, but created about 6 months before our experiments by a development team affiliated with the authors of the current paper. The team included an experienced web designer and usability specialist. Finally, Website # 6 in our experiments was the *People's Faculty* website, designed based on the guidelines provided by the KBS. More detailed information about the experimental websites is provided in Table 2.

Table 2. Websites used in the experiment

Website ID	Website URL	Description
#1	http://pensionerki.ru	Web forum for pensioners
#2	http://npfraiffeisen.ru	Non-state pension fund
#3	http://euro-kurses.ru	Business education center
#4	http://moscow.apteka.ru	Online medical shop
#5	http://vgroup.ru	Web development company
#6	http://nf.assoc.nstu.ru	The *People's Faculty* of *Novosibirsk State Technical University*.

For each of the websites, one to four tasks were developed, up to total amount of twelve (12) experimental tasks. The tasks were of two types typical for e-commerce: 6 "search" tasks, in which subjects had to look for certain information, and 6 "input" tasks that involved filling-in web forms. As the experiment participants were inexperienced senior internet users, the developed tasks were relatively simple, so that the total time of the experimental session would be no more than 90 minutes.

Among the recent graduates of the computer literacy courses for seniors provided by the *People's Faculty*, 11 subjects (2 male, 9 female) were recruited for our experiments. The sampling was not random, as higher priority was given to graduates with more intense online experience, which was deemed necessary to better simulate the current and future senior population online, or even the alumni themselves in a few months from the graduation date. The subjects' age ranged from 58 to 71 years, with the mean of 62.5 and standard deviation of 4.1. The mean self-reported time spent online by the participants was about 9 hours a week. All the subjects took part in the experiment voluntarily and provided their informed consent after reading through the tasks and learning the instructions.

The experimental design and settings were quite typical for a user testing session, but relatively high number of websites was involved and the subjects' online experience often differed considerably. Thus, some of the participants did not attempt to perform all the tasks, presented to them in a random order, during the experimental session time. All the subjects used the same software environment, 1024×768 screen resolution, and accessed the experimental websites with Mozilla Firefox 3.6.3 browser.

For each of the attempted tasks, the success rate was measured by the instructor: 0 was assigned for completely failed tasks, 0.3 for tasks involving major errors possibly requiring support from the instructor, 0.7 for tasks involving minor errors possibly requiring encouragement from the instructor, and 1 for successfully completed tasks. After completing all the tasks with all the websites, the subjects were also asked to evaluate their overall impression of the websites by ranking them on a scale from 1 (worst) to 5 (best). Thus, independent variables in the experiments were website ID and group (third-party website, expert designer's website or KBS-developed website), as well as task ID.

Dependent variables were task success rates and user subjective evaluations of the websites.

4.1 Experimental Results

In total, 106 tasks were performed by the participants and the overall mean success rate was 63.4% (see Table 3). The mean success rate for search-related tasks was 67.2%, while for input-related tasks, 59.4%. The mean success rate for the control group of websites (#1, #2, #3 and #4) was 40.8%, while for the website developed with the KBS (#6) it ran up to 85.9%.

For the website developed by the team including an expert designer and usability specialist (#5), the success rate was even higher, 86.4%.

Some of the senior participants refused to rank their overall impression of the experimental websites, blaming the lack of experience in judging websites, or assigned all-positive grades, which were not included in the analysis: all in all, the subjective evaluations were gathered from 7 subjects. Table 4 shows mean values and standard deviations for the evaluations included in the analysis.

Table 3. Success rates for tasks in our experiments

Website ID	Task ID	Task type	Task attempts	Success rate
#1	1	input	10	28.0%
	2	input	9	17.8%
#2	3	search	10	20.0%
#3	4	search	9	82.2%
#4	5	search	8	27.5%
	6	input	7	80.0%
#5	7	search	7	77.1%
	8	input	7	95.7%
#6	9	search	10	93.0%
	10	search	10	100.0%
	11	input	10	78.0%
	12	input	9	71.1%
			106	63.4%

ANOVA was used to test whether the subjective evaluations were significantly different for different websites. The difference was found to be not significant ($F_{5,32}=1.96$; $p=.111$); however, a *post-hoc* analysis showed that the evaluations for the Website #2 significantly ($p=.03$) differed from the evaluations obtained for Websites #3, #5, and #6.

5 Conclusions and Future Research

The paper proposes an application of indexing—the technology widely used in processing and comparing textual information—to multi-dimensional domain entities. The recommender system was built based on the frame ontology developed according to the proposed knowledge representation model and combining the domain conceptual model with the controlled vocabulary of base terms used for indexing.

Table 4. Subjects' subjective evaluations of websites

Website ID	Mean evaluation (standard deviation)	
#1	3.86 (0.69)	3.88 (0.91)
#2	3.29 (1.25)	
#3	4.29 (0.76)	
#4	4.20 (0.45)	
#5	4.33 (0.52)	
#6	4.50 (0.84)	

One of the major tasks was organization of design guidelines, so that they could be retrieved against the interface design task context, which is the "query" for the recommender system. Though the *bag of words* indexing is used for multi-dimensional objects in most existing solutions, we proposed to use base terms that correspond to the classes of the frame-based ontology describing the HCI (web applications) domain. This more sophisticated approach allows utilization of more complex algorithms for comparing indexed design guidelines, which are the main knowledge content of the system, with design context that is indexed using the same base terms.

The overall success rate in the testing, 63.4%, is in line with the data obtained for senior users in

other experiments. For example, Jacob Nielsen reported a corresponding success rate of 52.9% in 2002 [7], but web usability has obviously manifested certain improvement since then. Input-related tasks were predictably harder for seniors than search-related ones, producing success rates of 59.4% and 67.2% respectively. The developed website attained the success rate of 85.9%, higher than the double of the success rate for the control group's websites (40.8%) and almost catching up with the success rate for the website developed by human experts (86.4%). Although the users' subjective impression of the developed website was not significantly better than of the others, the results suggest reasonable feasibility of the developed knowledge-based system to support web design activities.

Further research prospects include introduction of "weights" for the terms, as measures of their "specificity" and significance in the index. Correspondingly, more complex comparison algorithms for defining the entities relevancy may be proposed, e.g., chosen among existing methods for measuring similarity of texts or other multi-dimensional objects (such as in [4]).

References

1. **Kosinov, D.I. (2007).** Использование статистической информации при выявлении схожих документов. *(Ispolzovanie statisticheskoy informatsii pri viyavlenii shozhih dokumentov) "Using statistical information to mark out similar documents". Internet-matematika,* Ekaterinburg, Russia, 84–91. Retrieved from http://elar.urfu.ru/handle/10995/1336.
2. **Mihalcea, R., Corley, C., & Strapparava, C. (2006).** Corpus-based and knowledge-based measures of text semantic similarity. *21st National Conference on Artificial Intelligence*, Boston, USA, 1, 775–780.
3. **Nguyen, B.N. (2012).** Модели и методы поиска информационных ресурсов с использованием семантических технологий. *(Modeli i metodi poiska informatsionnih resursov s ispolzovaniem semanticheskih tehnologiy) "Models and methods for performing search in information resources using semantic technologies".* Doctoral dissertation, Tomsk Polytechnic University, Tomsk, Russia. Retrieved from http://www.sscc.ru/Diss_sov/D02_2012.11.27.html
4. **Lv, Q., Josephson, W., Wang, Z., Charikar, M., & Li, K. (2007).** Multi-probe LSH: efficient indexing for high-dimensional similarity search. *33rd International Conference on Very Large Data Bases (VLDB'07)*, Vienna, Austria, 950–961.
5. **Chevalier, A., Fouquereau, N., & Vanderdonckt, J. (2009).** The Influence of a Knowledge-Based System on Designers' Cognitive Activities: a study involving Professional Web Designers. *Behaviour & Information Technology,* 28(1), 45–62.
6. **Henninger, S. & Ashokkumar, P. (2005).** An Ontology-Based Infrastructure for Usability Design Patterns. *Semantic Web Enabled Software Engineering (SWESE'05),* Galway, Ireland, 41–55.
7. **Nielsen, J. (*2002*).** Usability for senior citizens. Retrieved from http:www.nngroup.com/articles/usability-for-senior-citizens/.

Maxim Bakaev is Senior Assistant Professor of Economic Informatics Department of Novosibirsk State Technical University (Russia) and the Vice-Dean of Business Faculty. He has Master Degrees in Information Systems and Digital Design, and his research interests include Human-Computer Interaction, Interface Design and Usability, Knowledge Engineering, Universal Accessibility. He is a member of International Association of Computer Science and Information Technology (IACSIT) and reviewer for various international conferences, including KEER and NordiCHI.

Tatiana Avdeenko is Full Professor and the Head of Economic Informatics Department of Novosibirsk State Technical University (Russia). She holds the Doctoral degree in Mathematics Modelling, Calculus of Approximations and Software Systems, and her research interests include Knowledge Engineering and Management, Artificial Intelligence, Information and Decision-Support Systems. She is a Certified Expert of the National Agency for Accreditation in Education, and reviewer for various international conferences, as well as the journal Inverse Problems in Science and Engineering.

Article received on 01/10/2012; accepted on 15/12/2012.