ontology, knowledge base, knowledge engineering

Vasyl LYTVYN[*]*, Mariya HOPYAK*[**]*, Oksana OBORSKA*[***]

METHOD OF AUTOMATED DEVELOPMENT AND EVALUATION OF ONTOLOGIES' QUALITIES OF KNOWLEDGE BASES

Abstract

The process of automated development of base ontology is considered. It has been offered to consider the concepts and elements of ontologies for increasing the effectiveness of knowledge bases, the core of which is the ontology. Methods of specifying the weights of the relevant elements and optimization the structure of knowledge base of ontologies has been elaborated. It has been offered to evaluate the quality of the ontologies based on ISO 9126.

1. INTRODUCTION

Knowledge base (KB) is the main component of intelligent systems, which is formed according to the subject area on which the functionality of operation system is oriented. Traditional knowledge of engineering (receiving knowledge from expert, data analysis, machine learning, etc.) are not based on a system of common and verified standards, that is why knowledge bases, built on this basis, eventually lose their functionality due to the low efficiency of its operation. Ontological engineering is used as the standard of knowledge engineering, applying of which results in receiving the ontology of knowledge base. Ontology – is a detailed formalization of a certain area of knowledge presented by means of a conceptual scheme. This scheme consists of a hierarchical structure of concepts, relationships between them, theorems and constraints, that are accepted in a particular subject area [1].

[*] Lviv Polytechnic National University, Ukraine, 79013, Lviv, Bandera str., 28a, vasyl17.lytvyn@gmail.com
[**] Lviv Polytechnic National University, Ukraine, 79013, Lviv, Bandera str., 28a, mariya.hopyak@gmail.com
[***] Lviv Polytechnic National University, Ukraine, 79013, Lviv, Bandera str., 28a, oksana949@gmail.com

Considering the foregoing, the *formal model of ontology O* determines the following:

$$O = \langle C, R, F \rangle,$$

where C – finite set of concepts notions (concepts, terms) of subject area, which ontology defines O; $R: C \to C$ – finite set of relations between concepts notions (terms, concepts) in a given subject area; F – finite set of interpretation functions (axiomatization, constraints) specified on the concepts or relations of ontology O.

Using of ontologies as a part of KB helps to solve a number of methodological and technological types of problems that arise during the development of such systems. In particular, for Ukraine, the distinctive problems consist in the lack of conceptual integrity and consistency of certain techniques and methods of knowledge engineering; the lack of qualified professionals in this field; in the stiffness of the developed software tools and their low adaptive capacity; in the complexity of intelligent systems implementation, caused by psychological aspects. All these thesis indicate and confirm the relevance of research problems of using ontologies in the process of intelligent systems development [2-4].

2. THE PROBLEM FORMULATION

In order to manually design a complete related ontology for a specific subject area it is necessary to spend a lot of time and resources. It is explained that applied ontologies must contain tens of thousands of items to be suitable for solving a wide range of problems that arise in this subject areas. Manual ontology designing – is a long routine process, which also requires a thorough knowledge of a subject area and understanding of the principles of building ontologies. Therefore, methods and algorithms of automated ontology designing are actively developing. The mathematical software implementation of the automation process of designing ontology will be suggested, or rather, its development, as it is accepted that the human expert introduced the basic terms and relations between them in the ontology manually. Such initial ontology will be called base and will be denoted as $O_{base} = \langle C_b, R_b, F_b \rangle$. That is, ontology designing starts from the moment when it has already had some data. Therefore, this process is called base designing ontology. Formally, we will write:

$$\chi : O_{base} \to O.$$

Ontology – is the language of science. The language of science, as structured scientific knowledge, sets a hierarchical multilayer formation, in which the following components are distinquished: terminological system; nomenclature; tools and rules for forming conceptual apparatus and terms. So, for designing an ontology, it is necessary to build the terminological system O_T and the nomenclature O_N. Basic ontology necessarily contains some part of terminological system, that is $O_{base} \cap O_T \neq \varnothing$. Encyclopedias, terminological and explanatory dictionaries on the basis of which terminological system of subject area is build, usually have a clear structure and consist of dictionary articles. The process of building a nomenclature is more complicated. When in dictionaries terms are singled, then in scientific texts (books, monographs, etc.) they must be allocated and the properties of concepts and relations between them should be searched. Thus, the natural language technology of processing scientific text are required.

The purpose of this article is to develop the method of automated designing base ontology and evaluating its quality [5-7].

3. MAIN PART

3.1. Structural model of ontology concepts and relations

Let the given set of names of relations $V = v_1, v_2, ..., v_s$ be suggested. Then the relation in the ontology is given as a reflection from *C* to *C*, using the element of set *V*: $R : C \xrightarrow{V} C$. That is, relation r_i – triplet form:

$$r_i = \left\langle C_{i_1}, v_{i_j}, C_{i_2} \right\rangle$$

As the ontology forms the taxonomy concepts, then, using the object-oriented approach terminology, each concept represents a class. Let's define the concept as class with this structure:

$$C = \left\langle N, R^X, R^Y, S, D, A, Ob \right\rangle, \quad (1)$$

where N – the name of the concept; R^X - set of relations in which the class *C* is domain (area of definition); R^Y – set of relations in which class *C* is the set of values; *S* – superclass *C*; *D* – subclasses *C*; *A* – axioms definition *C*, *Ob* – instances *C*.

Consequently, designing the base ontology O_{base}, it is necessary to build triplets r_i and new concepts *C*, which are suggested by the structure (4). This structure includes a set of axioms *A*, but to build such a set of axioms in an automated way is very difficult (at least author doesn't know any of such attempt). Therefore at present this process is performed manually.

Such ready parsers as Link Grammar Parser [8] for the natural language texts processing with the purpose to build an ontology have been used. Six groups of relation patterns have been developed: 1) hierarchy, 2) aggregation, 3) functional, 4) semiotic, 5) identity, 6) correlation, just as in the work [9]. The search of appropriate relations in the text is performed on the basis of these patterns [10].

3.2. Algorithm development of base ontology

The idea that underlies in the automated development of ontology is that the processed texts with the knowledge of subject area are used to obtain data to complete the existing ontology. At the same time, the intermediate ontology is used for text processing of subject area. The result is a recursive process that can be considered as self-education of the system (Fig.1). Learning can be both automated and semi-automated with the help of a teacher. In process of education of the system, the need for a teacher will disappear and the process will be completely automated. The initial ontology with the basic concepts of subject area and commonly used terms should be defined a priori.

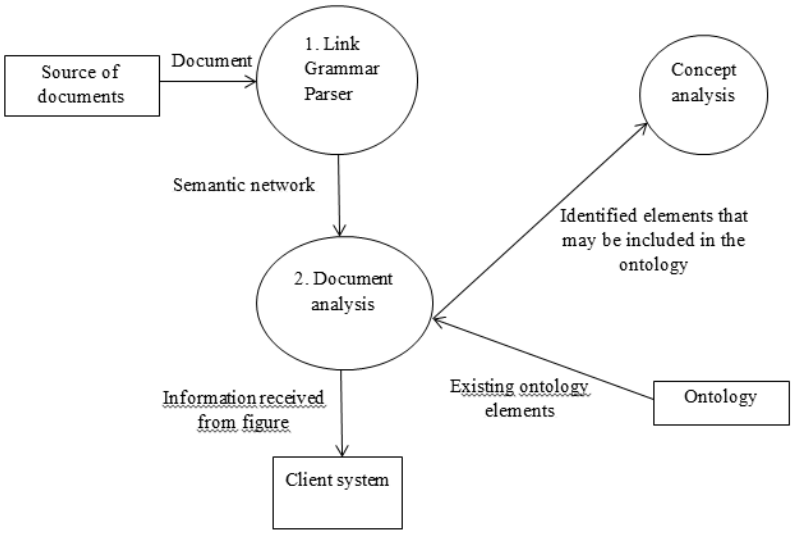

Fig. 1. Data flow diagram of automated ontology building [source: own study]

The designing of ontologies in the form of its learning based on scientific texts that are given by subject area, which are arranged in ascending order of difficulty processing will be organized. The degree of difficulty of processing text can be based on various criteria, such as the number of unknown terms that occur in the text, or the usage of the topological order of the tree of scientific papers that refer to each other.

To determine the new items that can be added to the ontology, a variety of methods that are mainly based on heuristics which take into account the existing elements in the ontology can be used. Using different methods, we obtain a set of possible modifications of ontologies among which we should choose the right ones. The choice is made by a teacher or it happens in an automated way, according to the previous studies.

Heuristics, involved in defining new elements, can be presented both as production rules, and can be based on pattern recognition algorithms, trying to supplement the areas of ontologies by the skipped items on the basis of existing templates.

The algorithm of developing base ontology on the basis of analysis of natural language text is as follows:

1) from the text the semantic units are singled out with reference to the corresponding elements in the ontology;
2) among interconnected semantic units the subset of which can form certain semantic templates which can create new elements for ontology are allocated;
3) semantic templates are added to the array in which after processing a text document a series of passes is carried out. During each pass a template can be seen to be possibly added into the ontology. If this template is allowed to be added to the policy of ontology designing, it is placed in the queue for a review performed by the administrator or added in an automated way depending on the degree of confidence of the template's type that is established by the policy of designing. Passages are carried until the new elements will be ceased to be added or a fixed number of times set by the designing policy;
4) the queue of templates is an oriented acyclic graph of proposals of the insertion of new elements in the ontology. Administrator will consider proposals from the upper level, if the proposal is declined, all proposals of lower levels will be refused in an automated way that became possible by the addition of the abolished in the queue. If the administrator has accepted a proposal, he takes the following proposals of the current level for consideration, if such proposals don't remain, it jumps to the next level. The role of administrator can run as heuristic algorithm of addition, depending on the policy of designing ontology;

5) any actions in the ontology logging in database, transactions and possibility of a change rejection are maintained, beginning from a certain point.

This process we described in details in [11, 12]. It should be noted that the line of research of automated designing of ontologies, using data bases of natural languages texts and systems based on them is actively developing. In particular, yearly The European Conference on Artificial Intelligence organizes individual sections of learning ontologies, which examines advances in the area of its automated formation.

The algorithm works with a semantic network obtained after using the Link Grammar Parser. The example of such a semantic network is shown in Fig. 2.

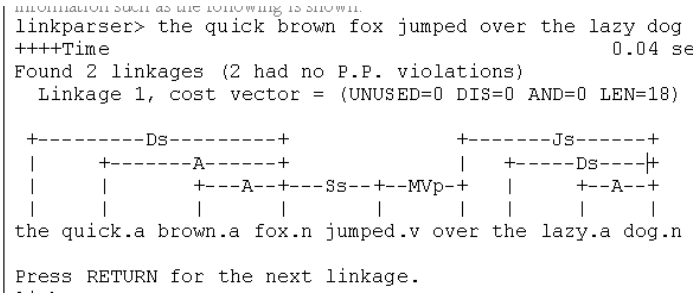

Fig. 2. A text document after using the Link Grammar Parser [source: own study]

Let the following sentence be suggested: "The quick brown fox jumped over the lazy dog". The verb "jumped" refers to the functional relations *Fun(v,x,y,z)* (Fig. 3). Next step is to investigate whether there exist the found concepts in the ontology. Five cases and appropriate actions can be possible (see. Table. 1). In the cases 2 and 5, new concepts are added in the ontology, in other cases either ignored or put in a database for further analysis.

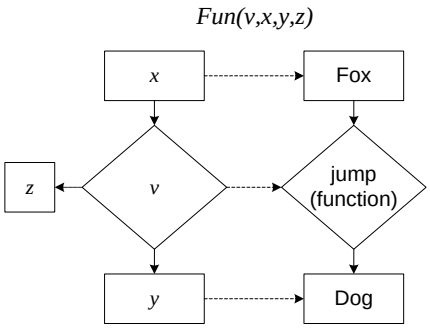

Fig. 3. Template matching [source: own study]

Table 1. Possible steps during automatic building of ontologies

№	Connection	Subject	Object	Possible action
1	+	+	+	Add into databases relations
2	+	+	–	Add an unknown concept to the ontology
3	+	–	–	Analyze
4	–	–	–	Ignore
5	+	–	+	Add an unknown concept to the ontology

Let the following sentence be suggested: "A steel has internal structure", as a result of processing of this sentence LinkParser obtains the relation "has", which belongs to the relations of hierarchical type, and it's pattern is Hier(a,x,y).

4. ADAPTATION ONTOLOGY

The effectiveness of adaptation ontologies of data bases to the subject area features determines the elements of its structure and mechanisms of its adaptation through learning during operation. One approach of the implementtation of such mechanisms is an automated weighing of data base concepts and semantic relations between them during learning. This is the role of weight's importance of concepts and relations. Weight's importance of the concept (communication) – is a numerical measure which characterizes the importance of certain concepts (communication) in specific subject area and dynamically changes according to certain rules during system operation. It is suggested by entering into its formal description the weight's importance of concepts and relations [13, 14]. Ontology like this one is defined as:

$$\hat{O} = \left\langle \hat{C}, \hat{R}, F \right\rangle,$$

where $\hat{C} = \langle C, W \rangle$, $\hat{R} = \langle R, L \rangle$, in turn W – weight of the importance of concepts C, L – weight of the importance of relations R.

Ontology defined in this way is called adaptation ontology, that is, one that adapts to the subject area using modification of default weights importance of concepts and relations between them.

To specify the weights' importance of relations for semantic tasks the research of the Danish scientists Knappe, Bulskov and Andreasen has been used [15]. They identified the following values of weights relations: $L_1 = 0,9$; $L_2 = 0,8$; $L_3 = 0,3$; $L_4 = 0,2$. For the relation of identities it is accepted that $L_5 = 1$. The relation of the correlation occurs only in attribute problems. For attribute tasks, leave (according to the heuristics), the concepts that are lower in the hierarchy are more important because they take specific values. Based on this, we consider that $L_1 = L_2 = 1,1$, such weights are similar as for semantic tasks. The relation of correlation is a two-way communication. Its weight is the module correlation between the features: $L_6 = |r_{ij}|$.

Methods of setting the weights of importance of concepts are:

- expert evaluations;
- frequency of use of concepts in scientific texts;
- using data mining, in which intelligent decision support system operates.

The method of calculating the weights of concepts is presented in [16, 17].

The defined ontology model of data base that allows to calculate the weight of its elements in the process of its adding, removal and usage during the operation of the system, thereby implements an adaptation mechanism to a given user subject area.

Obtained weight called weights of basic concepts, the set of such weights is indicated W_B.

These weights were developed for all ontology of subject area using taxonomy of ontology concepts, relations between concepts and their interpretation. The development of weights to the whole ontology depends on the definition (axiomatization) classes, their hierarchy (vertical connections) and horizontal connections. It is suggested to use a decision tree for setting the initial weights of concepts.

Vertices (signs) of a separate branch of the decision tree are placed into k levels. Obviously, the higher the level, the more significant feature that is included on this level. In addition, (it is suggested that) these weights to be normalized in order to their sum for each class (branch of the decision tree) be equal to 1.

Weights are defined as relation of the difference (k+1) level of tree and level that contains the sign to the sum of all levels of branches, that is:

$$w_i = \frac{k+1-i}{\sum_{j=1}^{k} j} = \frac{2\ \ k+1-i}{1+k\ \ k}.$$

We propose a method for determining the weights of all concepts in the ontology. First, the weight of all signs is equal 0. For the features that take part in the decision tree for the respective class to the primary weight we add the weight derived from the tree. All others should be calculated for the ontology of the corresponding task according to the formula:

$$W_j = \sum_{R^x} L_{ij} \cdot W_i + \sum_{R^y} W_k \Big/ L_{jk} \ . \qquad (2)$$

In the general case (2) is a system of the linear algebraic equations. However, in some cases, (2) is a sequence of the linear relations.

5. OPTIMIZATION OF ONTOLOGY AND QUALITY CRITERIA OF EVALUATION

Automated ontology development leads to the appearance of some weaknesses in its structure and content, the discrepancy of its content information filling according to the needs of the user. Therefore, such systems must be «complete» by the set of optimization procedures ontology.

Criteria optimization is formed according to the quality standard ISO 9126 [18]. According to this standard the quality characteristics are:

1. *Functionality* depends on the completeness and proper construction of ontology, how accurately it describes the specifics of the subject area and the problems that occur in it. In turn, the completeness of the ontology depends on the ability to give the correct answers to its queries, and it depends on whether the system is able to evaluate the novelty of knowledge offered to be added to the ontology. A measure of the quality of functional suitability is the percentage of non-trivial (non-zero) correct answers to queries to the ontology that is $\chi_1 = \frac{N_q^p}{N_q} \cdot 100\%$. Determination of functional suitability is one of the basic characteristics of ontologies.
2. *Reliability* (or correctness) functioning of data base – is the percentage of reliably solved tasks. This is the main quality characteristic of data base. Therefore $\chi_2 = \frac{N_z^p}{N_z} \cdot 100\%$.
3. *Usability* of resources (or resource efficiency) in standards is reflected as employment of the resourses of central processor, operative, external and virtual memory, input-output channels, terminals and communication channels. For improving those characteristics the optimization problem

is considered, a criterion which is to minimize the physical memory that takes ontology. On the other hand, it is obvious that ontology takes the least amount of memory, if there is no concept. So a threshold value established on the amount of memory occupied by the ontology.

4. *Efficiency* – difficult formalized concept that defines the functional suitability and effectiveness applications for certain users. This group of indicators includes subcharacteristics that reflect different aspects of functional clarity, ease of development, system efficiency and ease of use of ontologies. This suitability is based on the integrity of the ontology, that is the absence in its core mutually denied statements and duplication, as well as on balancing subject area, which consists in the full representation of its individual units in the ontology.
5. *Maintainability is* displayed by the convenience and effectiveness of corrections, improvements or adaptation of the structure and content of ontology database depending on changes in the external environment applications, and also in requirements and functional specifications of the customer.
6. *Portability* is characterized by long and laborious installation database, adaptation and replacement in case of transfering to other hardware and operating platform. The criterion of portability is the speed which is expressed in response time to external address (reaction time to change the parameters of the environment, to which system is sensitive).

Taking into account the above-mentioned criteria, optimization method ontology provides optimization problem of its structure and contents: 1) removing parallel edges, duplication tops with the same parameters and other features of the structure of the ontology graph that may impair its integrity and reduce the effectiveness of the functioning of intelligent systems (structure optimization problem ontology); 2) optimization of semantic part of ontology in order to increase its speed and information saturation of the given limits on the system's physical memory amount. Solving these problems is spaced in time, and to preserve the integrity of ontology, at first its structural inspection should be performed, and then – semantic optimization of content part of sequential reduction of the graph to the implementation of requirements of the selected criteria through maximizing the sum of weights of tops and edges of the graph [19-20].

Into the core of the minimization of graph structure of ontology problem a typical optimization problem of graph theory for finding the minimal base, which consists in finding the base of minimal weight in a weighted graph is entrusted. The task of ensuring consistency in the structure of the graph efficiently is solved by resolutions. The problem of content optimization is reduced to the inverse problem of the backpack. Let ontology consist of n elements with a general capacity of memory M. In the role of "backpack" acts

a certain given fraction of volume, for example $N = 0{,}1 \cdot M$, which should include the least valuable elements (the concepts with minimal weight of importance and maximal capacity) for their subsequent removal. Then it is necessary to maximize: $\sum_{i=1}^{n} \frac{1}{W_i} x_i \to \max$, such elements, for which $\sum_{i=1}^{n} m_i x_i \le N$, where $\begin{cases} x_i = 0, \text{ if the concept } C_i \text{ remains,} \\ x_i = 1, \text{ if the concept } C_i \text{ removes,} \end{cases}$, m_i – capacity of memory that holds the element C_i. Greedy algorithm was used for the solution of this problem. More optimization tasks of ontology are given in [7].

6. CONCLUSIONS

A method of automated development of a basic ontology using software Link Grammar Parser is considered. Classification of types of relationships is exercised. Mathematical software and algorithm for determining the type of relationships that occur in scientific texts has been elaborated. For the adaptation of the ontology knowledge base for the tasks that it can be solved by it, it is proposed to weight the ontologies' elements. The method of setting the corresponding weights of elements, which in turn makes it possible to optimize the content and structure of the ontology has been developed. It has been offered to use the standard ISO 9126 for evaluation of ontologies quality. The next step in the research will be the task of evaluation of knowledge innovation which are offered to be added to the ontology.

REFERENCES

[1] GRUBER T. A.: *Translation approach to portable ontologies.* Knowledge Acquisition, 1993, No. 5 (2), 1993, pp. 199-220.

[2] EUZENAT J.: *An API for Ontology Alignment*. [In:] Sheila A. McIlraith, Dimitris Plexousakis, and Frank van Harmelen, editors, Proceedings of the 3rd International Semantic Web Conference, vol. 3298 of LNCS, Berlin, Springer, 2004, pp. 698-712.

[3] BIAO QIN et al.: *Graph-based Query Rewriting for Knowledge Sharing between Peer Ontologies*. Information Sciences, 178(18), 2008, pp. 3525-3542.

[4] BAADER F., CALVANESE D., MCGUINNESS D., NARDI D., PATEL-SCHNEIDER P.: *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[5] BONTCHEVA K., SABOU M.: *Learning Ontologies from Software Artifacts: Exploring and Combining Multiple Sources.* [In:] Workshop on Semantic Web Enabled Software Engineering (SWESE), Athens, G.A., USA, 2006.

[6] JUNG J. J.: *Reusing Ontology Mappings for Query Routing in Semantic Peer-to-Peer Environment*. Information Sciences, In Press, Uncorrected Proof, 2010.

[7] DONINI F. M., NARDI D., ROSATI R.: *Description Logics of Minimal Knowledge and Negation as Failure*. ACM Transactions on Computational Logic, 3(2), 2002, pp. 177-225.
[8] LINK GRAMMAR HOMEPAGE – http://bobo.link.cs.cmu.edu/ link.
[9] НАЙХАНОВА Л.В.: *Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования*. Улан-Удэ: Изд-во БНЦ СО РАН, 2008. – 244 с.
[10] CHRISTOPH MEINEL SERGE LINCKELS: *Semantic interpretation of natural language user input to improve search in multimedia knowledge base*. Information Technologies, 49(1), pp. 40-48.
[11] LYTVYN V. SHAKHOVSKA N., PASICHNYK V., DOSYN D.: *Searching the Relevant Precedents in Dataspaces Based on Adaptive Ontology*, Computational Problems of Electrical Engineering: International journal, Lviv, Vol.2, Num.1, 2012, pp. 75-81.
[12] DOSYN D., LYTVYN V., YATSENKO A.: *DP-optimization of steel corrosion protection techniques in the intelligent diagnostic system.* Physicochemical Mechanics of Materials, No. 9, Lviv, 2012, pp. 329-333.
[13] LYTVYN V., MEDYKOVSKYJ M., SHAKHOVSKA N., DOSYN D.: *Intelligent Agent on the Basis of Adaptive Ontologies*. JOURNAL OF APPLIED COMPUTER SCIENCE, Vol. 20, No. 2, 2012, pp. 71-77.
[14] LYTVYN V.: *Design of intelligent decision support systems using ontological approach*. An international quarterly journal on economics in technology, new technologies and modelling processes, Lublin-Lviv, Vol. II, No. 1, 2004, pp. 31-38.
[15] KNAPPE R., BULSKOV H., ANDREASEN T.: *Perspectives on Ontology-based Querying.* International Journal of Intelligent Systems, http://akira.ruc.dk/~knappe/publications/ ijis2004.pdf
[16] LYTVYN V., DOSYN D., SMOLARZ A.: *An ontology based intelligent diagnostic systems of steel corrosion protection*. Elektronika, No. 8, 2013, pp. 22-24.
[17] LYTVYN V., SEMOTUYK O., MOROZ O.: *Definition of the semantic metrics on the basis of thesaurus of subject area*. An international quarterly journal on economics in technology, new technologies and modelling processes, Lublin-Lviv, Vol. II, No. 4, 2013, pp. 47-51.
[18] ISO/IEC 9126:1991. *Information technology – Software product evaluation* – Quality characteristics and guidelines for their use, 1991, p. 39.
[19] STOILOS G., STAMOU G., KOLLIAS S.: *A String Metric For Ontology Alignment* // In Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen, editors, Proceedings of the 4rd International Semantic Web Conference (ISWC), volume 3729 of LNCS, Springer, 2005, pp. 624-637.
[20] QIU JI, HAASE P., GUILIN QI: *Combination of Similarity Measures in Ontology Matching using the OWA Operator*. [In:] Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Base Systems (IPMU'08), 2008.