Expertise Search in Unstructured Data in ECM using S-BPM Approach

Alexander Gromoff, Julia Stavenko, Kristina Evina
and Nikolay Kazantsev

Institute National Research University, Higher School of Economics,
Faculty of Business Informatics, BPM Department,
Science & Education Center of Information Control Technologies, Moscow, Russia
alexander.gromoff@me.com, kevina@convera.ru,
{ystavenko, nkazantsev}@hse.ru

Abstract. This article describes the application of currently most promising methods of graph theory, content analysis and (3) subject-oriented approach to business process modelling for creating and automation of innovative process and therefore for maximization of ROI (return on investments) in intellectual and social capital of enterprises. In a course of development, instant full-text indexation takes place and taxonomic picture of different branches for such community is formed. In due course system gathers the statistics and builds-up maps of intercommunication with priority allocation of most discussed topics. A group of predetermined experts begins discussion on development prospects of this or that subject afterwards. The strategic map of investments into innovative development that can be offered to group of investors for competitive investments eventually turns out. In this process all steps except final (gathering of experts) are human non-dependant, what increase efficiency of the process in general.

1 Introduction

The most important property and feature of any information system is knowledge management, its allocation, processing and transformation, production and reproduction, transfer, storage and codification. Knowledge carriers are workers and external advisors. Therefore, while discussing innovative processes it is necessary to point out that not only 'who knows what' is important, but also 'who knows who and how comes that' is an essential part of knowledge exchange between members of social network. It is possible to talk about effective integration of employees into the added value chain of knowledge exchange process only having realized a nature of information as kernel element of the "doing by learning, learning by doing" approach to the company's activity.

How to create the corresponding information environment? Due to the researchers position, the answer to this question lies in an integrated management of process resources such as intellectual (people, information, and knowledge), and quality, and

corresponding processes is increased this in conjunction (K+Q) reduce the risk (general and operational), and as a result entropy of that system either stabilized or even reduced, because of the positive knowledge accumulation. And reversely, as soon knowledge is stabilized (stagnate) or even starts to decrease, because of internal or external processes destructing intellectual capital or potential of the system, at that particular moment quality starts to fall down letting operational risk to raise and all that in combination led immediately to increase of entropy of the system in general.

These relations are not un envying at all. In a big system with serious delay in reaction on internal or external changes it often can lead to total system destruction or dissimilation on subsystems till the level when each newly organized smaller system will obtain its level of entropy stability or manageability. That management is possible only by merging social and intellectual capital for obtaining the maximum efficiency. *In medias res,* 'social capital' as net substance connects an intellectual capital; these are interaction patterns, which create advantages to one social group, and, perhaps, barriers to another one.

2 Main Problems in Enterprise Knowledge Management

2.1 Problem 1

The main problem, in this case, is that not all types of structures are suitable for knowledge assignment. Everything rides on category of transferred knowledge. Explicit, easily codified knowledge can be hanged over by means of an e-mail, FTP, Internet or fixed in documents. Implicit knowledge, on the other hand, demands direct interaction and experience exchange between two and more employees. For example, presentations exchange, which are shown to employees, is usual practice, and here an exchange of context and expertize, necessary condition for creating such presentation (that has much higher intellectual value), not so simply occurs in companies because of existing organizational and social barriers. For transferring implicit knowledge direct connections with source of this knowledge, based on mutual understanding and trust between the recipient and the sender (mentoring and training condition) has to be established.

2.2 Problem 2

Next problem in network knowledge exchange is that the required knowledge is not situated often in a zone of employee's visibility, for example, in different clusters of employees in social structure. Social networks have so-called horizon, which is characterized by the degree of nods distance (managers, employees) from each other. It was shown repeatedly that such horizon in social networks represents two distance degrees – direct contacts of nod and their direct contacts of contacts [1]. On the third degree of a distance both the manager and the employee don't understand any more what is going on and this knowledge isn't available for them, except obvious well-

Thus, the popular theory that all of us are living in a small world and are connected by "six handshakes» is illusion. Six degrees of a distance are actually a really «big world» for organization and our possibility to find knowledge inside it, is very limited, therefore, in such a case knowledge is considered as inaccessible and often it is necessary «to invent a bicycle» for the solution of the task.

Without deepen discussion the root of this problems is obtained in human being or natural language uncertainty and redundancy, which accumulated with each transfer and finally exceed a thread hold of content identification. That is why hierarchical management structures with more than 3 levels are ineffective and extremely slow reactable.

In our case neither manager nor employee can receive the necessary expertize due to badly designed communications inside the company, mainly because of social and cultural barriers between them. In one's turn, it leads to occurrence of pseudo-scientific discussions in blogs, and then transfer of these imaginations in information flows and processes. Bert calls this situation «structural holes» in a network, meaning the existence of communicative spaces that are not connected among themselves [2].

2.3 Ex-ante Conclusion

All this leads to the problem of social capital management and merger of semantically close spheres of competences, which are burning issues in knowledge management as a whole. Scientific community is seriously bothered about its own ideas and opinions distribution and more specifically, implementation of these ideas in innovations that should convert future from Value added to Quality added paradigm. The true professionals, gained through the years of experience their unique knowledge, are sure that this knowledge would never be reduced to the elementary business implementation since it simply change a way of 'traditionalism' in business, and very often reduce significantly cost adding chain, what became 'immoral' in modern business world. Otherwise it's impossible to explain the facts of decades delay in implementation of evidently socially beneficial results of investigation in different areas and countries, but this question exceed a frame of the current work.

3 Enterprises as Closed Systems

Before shifting to description of the above-mentioned problems, it is necessary to formulate understanding of modern company as a system. Respectively, at first it is necessary to consider it as a part of system of higher order (for example, knowledge management system) and to allocate properties of this system and its subsystems from the governmental and social points of view.

The most important property and feature of this system is knowledge management, its allocation, processing and transformation, production and reproduction, transfer, storage and codification, as been mentioned above.

Among set of definitions of "knowledge": «Knowledge is information that is

process dedicated to obtaining measurable result. If to consider any knowledge one could have, it becomes evident that only application of the particular knowledge is a prove of its existence, otherwise it's just 'a manifestation of a will without potention'.

Still predominates the understanding of knowledge as certain number of definite information, instead of qualitative essence, which develops similar to live organism self-organizes, cooperates with an environment, breeds and modifies: «Practically any activity represents interaction of data carriers and knowledge. Optimal control of thi system consists of integrated process management of resources (people, information knowledge), quality and risks». It follows thence knowledge exists only in processe which have certain input and target characteristics according to the purposes. Thes processes have certain nods and problem points. The management integrity i knowledge exchange process consists in necessity for search of continuous balanc between volumes of information transferred for understanding, quality of resultan knowledge and risks of updating, commercialization, a prioritization, socializatio and environmental friendliness (respect for the environment) of knowledge.

3.1 Research Timeliness

Recently there was a set of researches on the topic of analysis of social networks which are growing out of their universal exploitation, both in corporate and in socia environment. The greatest distribution was received by the egocentric approach base on behaviour analysis of certain person (network nod), explaining social reality a derivative of its intelligent actions and its interaction with other people (nods). I egocentric approach it is possible to allocate the direction of social network analysi from status and role position of network nod point of view and from communicativ position (social capital) point of view. For instance, the carried-out researches showe that the degree of social capital of employee influences his promotion [3, 4], increas of salary [5], encouragement from friends and colleagues [6].

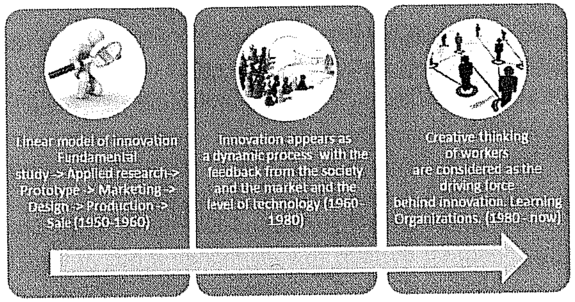
Linear model of innovation
Fundamental
study-> Applied research->
Prototype -> Marketing->
Design -> Production->
Sale (1950-1960)
Innovation appears as
a dynamic process with the
feedback from the society
and the market and the
level of technology (1960-
1980)
Creative thinking
of workers
are considered as the
driving force
behind innovation. Learning
Organizations. (1980- now)

Fig. 1. Current trend in knowledge management.

Existing researches on expertise search in a social network contain man achievements in this question. For example, some researchers use text classificatio for definition of expertise profile of each nod in a network by means of citatio analysis [7]. In other research an approach to expert-oriented search in social netwo

3.2 Motivation, Methodology, Research Questions

3.2.1 Research Aim

In this current research the main objective is the creation of innovative process in companies which is based on automation of knowledge exchange between employees. The approach is based on egocentric analysis of a social network and detection of manager and employee competences, in other words, definition of their expertize level.

Accordingly, some kind of virtual community communicating among itself on different topics is created; it is adhered to knowledge exchange process, for example, webinars on various subjects. These workshops are formed at once with a sight on receiving innovative and significant result, at its final stage the mechanism of investors competition takes place.

3.2.2 Methodology

On first stage the service of expertize search is developed by means of search inquiry creation on corporative portal or in e-mail, as a result the relevant list of experts (the indicator of intellectual capital of the company) is shown. Besides the list of relevance it is necessary to understand, whether this or that expert (employee/manager) is available to communication and adjustment of communication (the indicator of social capital of the organization). Merger of these two functionalities allows to create the two-factor indicator of expertize based on measurement of its intellectual and personal contribution.

On second stage the elimination of communication gaps between experts and efficient knowledge exchange by means of free ideas circulation in company on basis of internal communications. In other words the creation of convenient information environment in order to receive return from employee in form of a relevant independent expert appraisal of problem area. Such environment can be created through the virtue of subject-oriented approach application to automation of innovative processes in organization, where the main emphasis is placed on employees (subjects) reflexivity, in other words on ability to creative potential and self-analysis activization.

3.2.3 Research Questions

Research Question 1
Expert's identification on user demand (according to concepts in inquiry).

Research Question 2
Expert's social capital measurement in intra corporate expert network for employees ranking.

Research Question 3
Innovative process management of expertize transfer from one employee to another.

4 Research Description

Empirical base of this research is based on real correspondence data of managers and employees for the chosen period of time and unique communications base in the real processes, kindly provided by IT Co. for this particular work. E-mail can give a real backbone for semantic information observation and information on real social network. Implicit expert knowledge contains in text documents which employees exchange and describes their competences.

Essentially, any message in a network, in process of communications, directly or indirectly relating to business activity or business process execution, possesses the value for the analysis. This value can have various aspects as from point of view of solved in this work task (allocation of expert community), and from sociology, psychology and psychoanalysis point of view, besides, certain interest to results of this survey inevitably arises at enterprises security departments (including information security). Unavoidably, there should raise a question of private life rights protection of analyzed community, however, authors of this work would like to evade from discussion on "private life" existence within official duties or business processes.

However, it should be noted that the received results can be used by wide range of experts who are engaged in researches of organized communities for concrete result achievement.

4.1 Research Question 1

The first task is expert's identification on user demand (according to concepts in inquiry). Experts search (people possessing high qualification (competence) of subject domain uses the proved hypothesis that person's qualification strongly correlates with set of characteristic concepts which he uses; these terms are specific to concrete area. In this respect, subject domain for each set of terms can be different and is not connected with cognitive subject domains that is realized by person and is not allocated as separate essence. Now therefore, expert is the person who with high probability understands questions mentioned in the text. In order to separate significant terms from common ones, how it is described above, it is possible to formulate hypothesis that characteristics of rank distribution possess not only dependence on rank from frequency of word usage in the text (Tsipf's law), but also dependence on rank from relative frequency of term usage by the author. For this purpose it is necessary to count up statistics of relative frequency of term t_i usage for all texts written by the specific employee p_j

$$TF(t_i, p_j) = \frac{m(t_i p_j)}{\sum_k m_k}, \tag{1}$$

where $m(t_i, p_j)$ is number of utilization of term t_i by person p_j; total number yof utilization of all terms by person p_j in denominator

It is possible to assume that significant terms should have strong non-uniform dependences in double logarithmic scale for various terms (see fig. 3) chosen from the employees texts[1]:

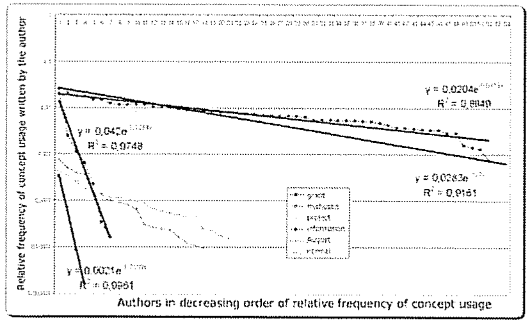

Fig. 2. Dependence of relative frequency of term usage written by the author from rank in double logarithmic scale.

It is intuitively clear that such terms as "grant" and «mishustin» are significant so "grant" is the specific term of subject domain, «mishustin» is the proper noun which are notional in narrow range of experts. Terms "project", "information" are common in the chosen subject domain because probability of its utilization by each enterprise employee is approximately the same. Analyzing the diagram it is visible that for significant words and general meaning words the distribution character differs, namely dispersion in observable distribution of relative frequencies of terms usage.

In consequence of the experiment, it is possible to draw conclusion that exactly higher values of dispersion of relative frequency of term usage is that criterion which allows distinguishing significant terms from the common ones. It has to be mentioned that on the basis of dispersion calculation of relative frequency distribution of terms occurrence algorithm it is possible to reveal experts for program service of intra corporate experts search which is based on texts analysis. This analysis is used for formation of limited list of employees (experts) working at the enterprise who are most competent in this or that question. On the basis of expert's lists the intra corporate expert network is formed.

However, receiving one assessed value on dispersion might not be enough in order to get stable result. On this figure the dispersion value for words "august" and "normal" are approximately equal but are they are displaced towards "significant" ones. Whether this means, that they should be defined as concepts for the narrow range of experts, categorically "no" as these concepts were allocated at the expense of discussion of the holidays season and conditions of its carrying out (so-called social factor).

The following algorithm allows to essentially extend the definition of expert distribution and to clarify its specifics. Using expert-determined taxonomies for concrete "closeness" processes definition of each specific employee to these distributions by calculating percent of words belonging to taxonomy, from total number of used words employed by the employee, finally, we will receive the distribution vector of personal "closeness" of each employee concerning all (or allocated) processes in corporation. It is note-worthy, that this representation won't be static or constant, moreover, it is possible to judge adaptability of the specific employee or change of his personal ambitions, interests etc. from speed of its change.

The following chart (Fig.3) depicts change of accents in activity of the specific employee within a year.

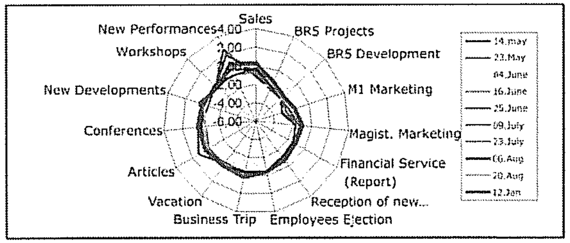

Fig. 3. Changes in activity of concrete employee within a year.

The analysis of changing interests or priorities for this distribution remains behind this framework. It is possible to note only that the distribution "anomaly" falling on 4th of June corresponds to the first return week from fortnight business trip of this employee.

Such distributions are unique and can be in many respects comparable with "intellectual prints". Distributions of new subjects, new concepts, and discussion circles are developed analogically; their basic difference from significant or traditional is that till some particular period of time they just haven't occurred. It is unessential that emergence of similar innovation gives the evidence that there is a potential innovative process. Emergence of new words, jargons, new political subjects inevitably will be fixed but this will be question of an expert assessment what has a potential for development and what is language evolution.

4.2 Research Question 2

The second research topic is expert's social capital measurement in intra corporate expert network for employees ranking. Methodological base of research are provisions of modern theory of networks, mathematical linguistics, sociology and psychology. Communications can be presented in form of social system described by means of graph G = (V, E) where set of V-tops represent employees, and set of E-edges represent connections between employees, expressed in their communications with each other.

Messages exchange among employees can be presented as network structure of interaction – as bigraph. Positioning isn't satisfied by spatial distance. Any position of individual in network is defined by its relations to other positions.

In order to measure social capital it is offered to calculate indicators of nods centrality. Centrality – is indication of how high is the employees social capital, it is based on number of its communications with other network nods. There are three main indicators of centrality:

ways of interaction of other nods except through this nod, it will have the maximum influence. Removal of nod which has big betweenness indicator will cause break of information flow and will lead to network [10] fragmentation. Such nods act as brokers or doorkeepers as they supervise information flows [11].

- Closeness shows possibility of fast access to information; it is inversion of sum of the shortest distances between each nod and each different nod in network. The fewer the intermediary nods between the current nod and other nods, the lower is the closeness indicator and the higher is the closeness degree[12]. This position is quite advantageous at communications implementation.Than less than intermediary nods between the flowing nod and other nods, the indicator of closeness and subjects is lower than subjects degree of closeness [12] is higher. This position is very advantageous at communication implementation.
- Centrality degree – this characteristic shows who the most active nod in network is. In compliance with networks theory a large number of interactions of nod can not only change nod position in network, but also change positions of other nods. The individual indicator of centrality shows, in what degree the nod is connected by other knots, that is how closely it is connected with group [13].
- Centrality as indicator of centrality of own vector (eigenvector centrality) — nod importance in network [14]. The indicator estimates relative measures for all nods inside network based on to whom nod neighbors have connection.
- Clustering coefficient [15] - degree of nods connectivity in network. This cofficient characterizes tendency to formation of groups of interconnected nods, so-called cliques.

4.3 Research Question 3

The third task is innovative process management of expertize transfer from one employee to another.

Thesis 1: The innovative system should possess ability to support interaction between innovators and experts for carrying out expertize of an innovation.

Thesis 2: The most expedient way of creation and automation of innovative process is application of subject-oriented approach to innovative process management. In such case there are all necessary conditions for realization process and network communities ad hoc and also for brightest development of reflection while creating new knowledge.

For specification of above-mentioned theses let's consider how S-BPM realization in tool system Metasonic (former jCOM1) S-BPM Suite looks.

The model of innovative process in «Process Manager» is designed in such a manner that the subject "Initiator" (the founder of innovation) sends the message "Innovation" to the subject «Experts Search Service» (it not the person but the element of system which is processing information). «Experts Search Service» possesses profiles of enterprise staff, in reply to the demand sends the message to the initiator with candidates of potential investors of intellectual capital and their profiles.

approbation «Confirmation of Accedence to Community» or denial. The new community for innovation development is automatically created where all experts who have accepted the inquiry take part. Afterwards the innovation discussion process will occur in expert's community. The potential investor and experts turn into participants of innovative process. After accumulation of intellectual investments of all community participants development of innovation takes place (Fig. 4).

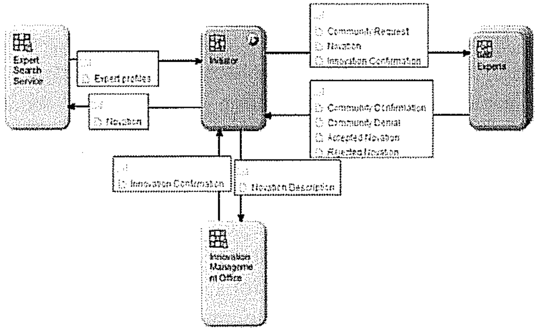

Fig. 4. Example of describing an interaction of actors in the subject-oriented model of the Innovation Process Management (S-BPM point of view).

4.4 Research Conclusion

In course of innovative creativity the most essential risk factor is uncertainty and irrelevance of utilized information. Continuous updating of analyzed information and information selection, which is relevant to solving task, is required. Therefore, except workflow system the connection with information services providing reliable, adequate access to relevant information from unstructured sources and services of coding, preservation and access to problem-structured information are necessary.

The described approach on the S-BPM platform allows realization of operative connection to both: multiple services of information access to unstructured information and various DBMS with access to data at level of fields. Thereby, there is an innovative process management system architecture autogeneration which in more mature phases of self-realization can be corrected and analyzed by the experts groups who are responsible for innovative development processes.

5 Conclusions

The carried-out research solves three problems:

1) Intellectual capital assessment: experts search (people registered in system as texts authors and possessing high qualification (competence) of subject domain) by concepts, which relative frequency of mention allows to connect it with specific subject domain. Characteristic terms can be in relation of ontological proximity with

2) Social capital assessment: metrics calculation allowing estimation of communication efficiency among experts.

3) Innovative process management: innovations creation and distribution process automation in companies at the expense of independent expertize founded by experts search service and assessment of employees information interaction.

At the moment there already exist solutions of automatic experts search. The HelpNet tool uses information generated by users for creation of an expert profile [16]. One more similar tool is Expert Locator which uses a representative collection of technical documentation written by the employee for creation of expert indexes [17]. Recently NASA's agency presented the Expert Finder [18] solution which uses substances (keywords) allocation for publications submission summaries of user and then builds experts rating in process of inquiries relevance. One more similar system is I-Help [19] – agent system which models user's characteristics for the other employee search who might help him. Vector model is used for the most suitable employee selection from the stated inquiry of allocated information.

The main problem of above-listed options is that they represent the subjective value of employees competence and assume structured information utilization which is created, collected, classifed and exchanged by employees. However, the majority of organizations do not structure information but it actually contains answer to the question who is the expert finally.

Novelty of the current research lie in creating means for automatic detection of person who is an expert in what area, effective search of such expertize, informing who from the identified experts is in network and means of communication with him. All this will allow strengthening and accelerating innovative processes in organizations at the expense of favorable information environment creation which simplifies information exchange between employees and allows accumulating, generalization and classification of advanced knowledge.

Quantity of efforts, which have to be made in order to organize innovative process in organization depends on number of factors, such as organization size, automation level, force of social communications in organization and etc. The developed service can be applied as the tool to the solution of experts search problems on corporate portals, in ECM – system and in any other IS accumulating publications data of employees.

The further approaches and results of these studies may be used afterwards for improvement of the incumbent companies as well as for processing and transferring of the complicated unstructured information content within the Enterprise 2.0, joined ventures or modern vertical integrated organization.

Acknowledgements

The given research was held in a frame of the contract № 13.G25.31.0096 with the Ministry for Education and Science of Russian Federation «Creation of hi-tech manufacture of unstructured information processing in cross-platform system on the

References

1. Noah E., Friedkin & Eugene C. Johnsen: Social Influence Network Theory: A Sociological Examination of Small Group Dynamics (Structural Analysis in the Social Sciences), Cambridge University Press; 1 edition edition (2011)
2. Burt R., 2001. The Social Capital of structural holes // Guillen M.F., Collins R., England P., Meyer M. (eds.). New Directions in Economic Sociology. N.Y.: Russel Sage Foundation: 201-246.
3. Fernandez, R. M., Castilla, E. J., & Moore, P. 2000. Social capital at work: Networks and employment at a phone center. American Journal of Sociology, 105: 1288-1356.
4. Granovetter, M. [1974] 1995. Getting a Job: A Study of Contacts and Careers. Chicago, IL: The University of Chicago Press.
5. Seidel, M. D. L., Polzer, J. T. & Stewart, K. J., 2000. Friends in high places: The effects of social networks on discrimination in salary negotiations. Administrative Science Quarterly, 45:1-24.
6. Adler, P. S. & Kwon, S. W., 2002. Social capital: Prospects for a new concept. Academy of Management Review, 27: 17-40.
7. Xiaodan Song, Belle L. Tseng, Ching-Yung Lin and Ming-Ting Sun, "ExpertiseNet: Relational and Evolutionary Expert Modeling", Intl. Conf. on User Modeling, Edinburgh, UK, July 2005.
8. Jing Zhang, Jie Tang, Juan-Zi Li: Expert Finding in a Social Network. In Proceedings of DASFAA'2007. pp.1066~1069.
9. Yupeng Fu, Rongjing Xiang, Yiqun Liu, Min Zhang, Shaoping Ma: Finding Experts Using Social Network Analysis 2007 IEEE/WIC/ACM International Conference on Web Intelligence.
10. Borgatti, S. P. & Everett, M. G., (2006). A Graph-Theoretic Perspective on Centrality. Social Networks, 28(4), 466-484.
11. Wang, J., & Chen, C., (2004). An Automated Tool for Managing Interactions in Virtual Communities - Using Social Netwrok Analysis Approach. Journal of Organizational Computing and Electronic Commerce, 14(1), 1-26.
12. Wasserman, S. & Faust, K., (1994). Social Network Analysis: Method and Applications.Cambridge, UK: Cambridge University Press.
13. Ahuja, M., Galletta, D. & Carley, K., (2003). Individual Centrality and Performance in Virtual R&D Groups: An Empirical Examination. *Management Science, 49*(1).
14. P. Bonacich, Factoring and weighting approaches to status scores and clique identification, Journal of Mathematical Sociology, 2 (1972).
15. Watts, D. J.: Collective Dynamics of «Small-world» Networks // Nature / Ed. by S.H.Strogatz. 1998. Vol. 393.
16. Maron, M. E., Curry, S., Thompson, P.: An Inductive Search System: Theory, Design and Implementation. IEEE Transaction on Systems, Man and Cybernetics, 16(1) (1986), 21–28.
17. Steeter, L. A., Lochbaum, K. E.: An Expert/Expert Locating System based on Automatic Representation of Semantic Structure. In Proc. of the Fourth IEEE Conference on Artificial Intelligence Applications: San Diego, CA, (1988), 345–349.
18. Staab, S.: Human language technologies for knowledge management. Intelligent Systems, 16(6): (2001), 84–94.
19. Bull, S., Greer, J., McCalla, G., Kettel, L., Bowes, J.: User Modelling in I-Help: What, Why, When and How. In Proc. of the 8th International Conference on User Modeling, Sonthofen, Germany, July, (2001), 117-126.
20. Alexander Gromoff, Valery Chebotarev, Kristin Evina and Yulia Stavenko: An Approach to Agility in Enterprise Innovation S-BPM One Learning by Doing - Doing by Learning