University's Scientific Resources Processing in Knowledge Management Systems

Zhomartkyzy Gulnaz[1], Milosz Marek[2] and Balova Tatiana[1]
[1]*Department of Information System, D. Serikbayev East Kazakhstan State Technical University, 69 A.K. Protozanov, Ust-Kamenogorsk, Kazakhstan*
[2]*Institute of Computer Science, Lublin University of Technology, 36b Nadbystrzycka, Lublin, Poland*

Keywords: Knowledge Management System, Ontology, Semantic Search, Terminological Collocations, Collocations.

Abstract: This article deals with some issues of modern approaches to word processing in knowledge management systems. The method of documents' profiles formation based on scientific knowledge ontology model which provides the semantic processing and retrieval of information is proposed. The article describes the main stages of the university's information resources word processing to form a semantic document profile: the extraction of terminological collocations, the automatic classification of texts on scientific topics, the formation of a document's semantic profile.

1 INTRODUCTION

The development of information technologies has enhanced the dissemination and the use of knowledge in all spheres of public life. The informatization of the society has served the establishment of knowledge-based economy; therefore, the research in the field of knowledge formalization and smart information processing is of particular importance and practical necessity.

The scientific knowledge of a university is important for both the university's innovative development and its educational process improvement. The task of preserving and sharing the knowledge accumulated by the university has been brought into the forefront, and the research related to the formation and the completion of the university's scientific knowledge base has become topical.

The members of the scientific community need to interact with each other and to gain access to information in order to create new knowledge. The need for information is growing exponentially; the existing information systems do not provide the search, development, exchange and management of information to the full extent. So, the need to develop knowledge management systems (Maier, 2007) has emerged.

Scientific cooperation is done through paper and electronic information flows and is based on direct (personal) communication. Scientific on-line journals, electronic catalogs and university libraries databases, the Internet, collaboration tools, and e-mail facilitate the access to the required information and accelerate the traditional processes of knowledge dissemination, but they do not support the semantic search. Modern smart information technologies require new scientific approaches and fundamental changes rather than the automation and acceleration of traditional information exchange processes.

The Scientific Knowledge Management System (SKMS) presented in this paper serves as the technological component of the university management; it ensures the creation, organization and dissemination of scientific knowledge among the university staff. The ontology is used as the information model in the SKMS. An ontology is a conceptual domain model which is the system of concepts, their properties and relations understandable to both humans and computers.

The developed ontology of scientific activities describes the structure of scientific knowledge base, provides access to its content and allows forming of employees' scientific profiles. As a result, when there is a need in knowledge, a user doesn't have to search through different sources. The ontology allows access to all the necessary knowledge through a single (user's) interface regardless of the knowledge representation form.

Gulnaz Z., Marek M. and Tatiana B..
University's Scientific Resources Processing in Knowledge Management Systems.
DOI: 10.5220/0004886802250232
In *Proceedings of the 16th International Conference on Enterprise Information Systems* (ICEIS-2014), pages 225-232
ISBN: 978-989-758-028-4

The accumulation of scientific knowledge is provided from different sources of structured information (digital library databases) and unstructured information (research articles, conference proceedings). The unstructured scientific resources are classified by means of classifiers, then they are recorded in the ontology as "Information resources" class instances and form the document's semantic profile.

The aim of this work is to create documents' profiles in ontology model for effective knowledge management in scientific knowledge management system SKMS.

The proposed semantic technologies form the basis of the university's scientific knowledge semantic portal.

2 LITERATURE REVIEW

The ontological approach for annotating documents for automatic identification of employees' competences is considered in the papers (Allemang and Hendler, 2011); (Altınçay and Erenel, 2010).

The paper (Ma et al., 2012) shows the use of ontology-oriented hybrid methods of TextMining for processing text applications.

The methods of users' profiles formation based on their personal documents and computing the semantic similarity of users' profiles with spreading activation networks derived from ontologies is described in the paper (Thiagarajan et al., 2008).

The technology for automatic text categorization for automatic conceptual indexing is described in literature (Lukashevich, 2011), as well.

This paper describes the methods and the problems of automatic categorization, the scheme of heading description, various machine learning techniques for text categorization. The problems of document classification and clustering were described in detail in papers (Manning et al., 2009); (Du and Chen, 2013); (Shengyi et al., 2012); (Jiang et al., 2012) in which it was proposed to use the ontological vocabulary domain. The methods of terminological collocation automatic extraction in order to form a domain subset of terms were outlined in (Pivovarova and Yagunova, 2010); (Sedova and Kvyatkovskaya, 2011). The topic-subtopic relations are used in the domain ontology.

The analysis of the available literature sources confirms the scientific and practical significance of this document profile formation method based on the scientific knowledge ontology model, which provides the semantic processing and retrieval of information.

3 THE INFORMATION MODEL OF THE UNIVERSITY'S SCIENTIFIC RESOURCES

The ontology-oriented approach allows organizing and structuring of the university's research related information resources and developing the methods of search for knowledge.

The information model of the university's knowledge can be described as the ontology which includes the basic concepts of the university's scientific activities, such as organizational structure, subjects, the objects of scientific schools and research, information resources, other subdisciplines, etc.

Subclasses correspond to major research areas in the ontology of scientific activity in the "Research directions" class. The subclasses consist of topics which correspond to the headings of the knowledge area classifier.

OWL DL (*Web Ontology Language*) is used as the ontology description language (Allemang and Hendler, 2011). The developed hierarchical taxonomy of classes and their determined roles cover the basic elements of the university's scientific activities management. The possibilities of DL (*Description Logics*) provide the ability to define cardinality, hierarchy of roles, inverse and transitive roles.

OWL axioms of classes and relations were compiled and attribute constraints were set when developing ontology O_A for descriptions for different characteristics classes and properties, i.e. (Guarino, 2009):

A. Classes:
- RC_U – university research centers
- RC_D – department research centers
- RC_I – interuniversity research centers
- $D(x)$ – departments
- SM_i – members of scientific schools
- O – organization
- SS – scientific schools
- $IRSS$ – information resources of research schools
- SD – subdiscipline

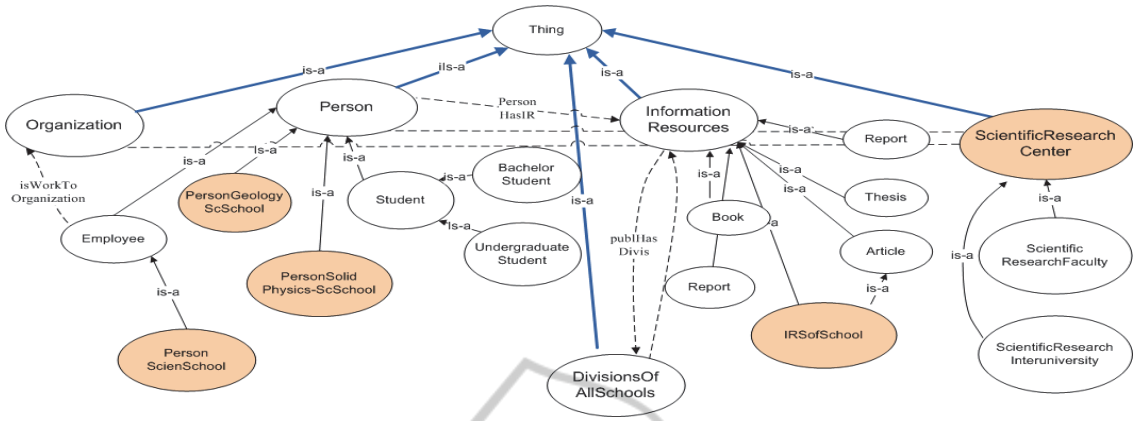

Figure 1: A fragment of the scientific activities ontology.

P – persons
PSS – persons at research schools
IR – information resources

B. Roles:
isWorkOrganization – works for the organization
workSchool – in research school
personHasIR – has publications
publHasDivis – refers to a subdiscipline

C. Axioms of classes and roles:

$$RC_U \equiv RC_D \cup RC_I$$

$$D(x) \equiv \{x_1\} \cup \ldots \cup .\{x_i\}$$

$$SM_i \equiv P \cap \exists isWorkOrganization.O \cap \exists WorkSchool.SS$$

$$SS_i \equiv SS$$

$$PSS_i \equiv P \cap \exists personHasIR.IR$$

$$IRSS_i \equiv IR \cap \exists personHasIR.IR$$

Figure 1 shows a part of scientific activities ontology as an example.

4 THE PROCESSING OF THE UNIVERSITY'S INFORMATION RESOURCES

Figure 2 shows the procedure of the university's information resources processing with the purpose to form the university's scientific schools and research areas. It includes the following stages:

- Extraction of terminological collocations. C-value method is used as the method for detecting collocations (Braslavsky and Sokolov, 2008); (Min et al., 2012); (Sedova and Kvyatkovskaya, 2011);
- Feature selection. test was chosen to assess the importance of terms (Manning et al., 2009);
- Classification of texts according to scientific areas. The method of k nearest neighbour (kNN) is used for text classification (Braslavsky and Sokolov, 2008); (Du and Chen, 2013); (Shengyi et al., 2012); (Jiang et al., 2012).

A detailed description of processing stages is given in the following sections.

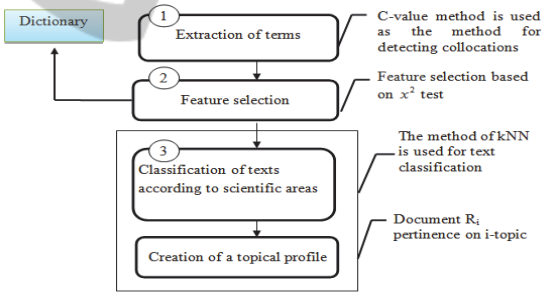

Figure 2: Stages of the university's scientific resources processing.

4.1 Automatic Extraction of Terminological Collocation from Scientific Texts

Collocation is regarded as a non-random combination of two or more lexical items common to most scientific texts. The set of terminological collocations generated by the specified collection of scientific texts describes a narrow subject area (topics and subtopics) of this collection.

For automatic extraction of terminology collocations from scientific texts that form ontological vocabulary such methods as mutual

information (MI) method and the C-value method (Pivovarova and Yagunova, 2010); (Sedova and Kvyatkovskaya, 2011); (Braslavsky and Sokolov, 2008) were examined C-value method ranks the terms based on their frequency and nesting:

$$C - value\ (a) = \begin{cases} \log_2 |a| \times f(a) \\ \log_2 |a| \times \left(f(a) - \frac{1}{P(T_a)} \times \sum_{b \in T_a} f(T_a) \right) \end{cases} \quad (1)$$

where:

$\log_2 |a| \times f(a)$ – if a is not into other substrings;

a – is a candidate term;

$|a|$ – is the length of a collocation;

$f(a)$ – is the frequency of a collocation;

T_a – is the set of phrases containing a;

$P(T_a)$ – is the number of longer collocations containing a;

$f(T_a)$ – is the sum frequency $P(T_a)$.

The main modification of the method based on the static approach includes the preliminary use of morphological templates of filters similar to the following (used abbreviations: n. – noun, adj. – adjective, g.c. – genitive case, participle - part.) (Braslavsky and Sokolov, 2008):

[adj. + adj. (genitive case) + n. (genitive case)]

[adj. + adj. + n.]

[adj. + adj. + n. (genitive case)]

[adj. + n. (genitive case) + n. (genitive case)]

[adj. + n.]

[part. + n.]

[n. + n. (genitive case)]

[n. + n.)]

To obtain the list of dominant terms from a document it is necessary to solve the following problems:

- to get a list of all the terms used in the document;
- to select the terms which are dominant in the given document from the list.;

Phrases similar to the terms are extracted from the text by C-value method:

- the text is divided into strings taking into account the punctuation, any sequence of words in the text which are not separated by punctuation marks is considered;
- phrases that meet the following conditions are extracted from the text:
  - morphological analysis of the text (pos tagging) is performed, each word's part of speech and morphological characteristics are determined. The maximal sequence of words is a trigram;
  - only those collocations which match the template are selected, stop words which do not have a semantic load are removed.
- entries are created in the database for all candidate terms which have the value greater than 1 by C-value method.

Restricting the C-value will only allow considering the terms longer than one word, because the C-value of a one-word term is always zero.

Candidate terms obtained in this way form a list of n-grams (bigram, trigram).

The corpus of documents for processing was compiled from articles in various fields published in "Physics of the Solid State" journal founded by the Russian Academy of Sciences, the Department of General Physics and Astronomy, Ioffe Physical Technical Institute, RAS (Science journal "Solid State Physics, 2013).

Resource: the author's own development with the corpus of documents from "Solid State Physics" journal.

For each candidate term the C-value is calculated according to formula (1). Table 1 shows the results of terminological collocations extraction module work for the "Solid State Physics" area of science which is a part of created ontology.

4.2 Feature Selection for Classification

Methods of feature selection are used to reduce the dimension feature space T to form the most informative space (Manning et al, 2009); (Altınçay and Erenel, 2010). Feature selection promotes:

- learning efficiency by reducing the size of the vocabulary;
- classification accuracy by eliminating noise characteristics.

Utility measure $A(t,c)$ of each term in the vocabulary is calculated for each class, N terms having the greatest value $A(t,c)$ are selected. All other terms are discarded and they do not participate in the classification. In this paper three utility measures are considered: Mutual Information, criterion X^2, Expected Mutual Information.

Table 1: Terminology n-grams (bigrams, trigrams) for Solid State Physics.

Trigrams	
Terms	C-value
partial pressure of oxygen	42,79
interstitial silicon atom	38,04
intrinsic interstitial atom	33,28
formation of vacancy micropore	33,28
intrinsic point defect	31,70
strain-hardening coefficient	23,77
amplitude of oscillatory strain	23,77
dislocation-free silicon single crystal	23,77
speed of free surface	22,19
binary solid solution	22,19
dynamic point defect	22,19
transverse size of a crystal	22,19
degree of plastic deformation	19,02
elastic planar mesodefect	19,02
amorphous covalent material	17,43
factor of size mismatch	17,43
Bigrams	
plastic deformation	162,50
solid body	139,00
grain size	107,67
point defect	76,31
room temperature	73,00
grain boundary	70,60
oxygen atom	59,56
plastic flow	57,60
Burgers vector	56,09
dislocation density	54,20
Young 's modulus	52,83
deformation rate	49,83
strain hardening	48,56

To remove non-informative terms, i.e. to assess the importance of terms criterion X^2 was selected. Criterion X^2 is used to test the independence of two random events, where events are considered independent. When selecting features the two events are the appearance of the term and the emergence of a class.

Experimental evaluation of this method has shown that it extracts key terms with high accuracy and completeness (Manning et al., 2009). The ontological dictionary is formed from these terms, and it is then used for the classification of thematic text annotation.

4.3 Text Classification by Scientific Areas

For the classification of scientific resources *kNN*-classification is used. The classification task in machine learning is a task to assign an object to one of the predefined classes based on its formal characteristics.

kNN method (*k* method of nearest neighbor) is a vector classification model. *kNN* classifier assigns the document to the prevailing class of nearest neighbors, where k is the method parameter. The k parameter in *kNN* method is often selected on the basis of experience or knowledge about the classification task at hand.

In this paper, the *kNN* method is used for multilabel classification. Classification for classes that are not mutually exclusive is called multilabel classification (Malarvizhi and Ramachandra, 2013); (Ceci et al., 2012).

In multilabel classification j learning occurs for different classifiers γ_j. Each of the classifiers returns either class tag c_j, or class tag $\overline{c_j}$, i.e. $\gamma_j(d) \in \{c_j, \overline{c_j}\}$, where d is the tested document.

The document may belong to several classes at the same time, one class or not to belong to any class. Classes are conditionally independent of each other.

The solution of multilabel classification problem using linear classifiers can be described as follows:

- a classifier for each class is created, at that a training set consists of a set of documents belonging to the class;
- each classifier is applied individually to the document, the decision of one of the classifiers does not affect the decision of the other one.

Each document d in this problem is represented as a vector $v(d) = (w_1, w_2, ..., w_N)$ in N - dimensional space, where each dimension is a description of one of the document features (Manning et al., 2009); (Bolshakov et al, 2011); (Liu et al., 2009). The weight of each feature (term) is calculated as follows:

$$w_{t,d} = tf_{t,d} \times \log \frac{|D|}{df_d} \quad (2)$$

where:

D – is a set of documents of class c_j;

$|D|$ – is the total number of documents in the

corpus;

$tf_{t,d}$ – is a term frequency in the document;

df_d – is the document frequency which is the number of documents in the collection containing the t term.

To calculate the k nearest neighbors the cosine measure was selected as the primary measure of affinity (Manning et al., 2009); (Du and Chen, 2013); (Shengyi et al., 2012):

$$\cos(v(d'), v(d)) = \frac{\sum_{i=1}^{n} v_i(d') \times v_i(d)}{\sqrt{\sum_{i=1}^{n} (v_i(d'))^2} \times \sqrt{\sum_{i=1}^{n} (v_i(d))^2}} \quad (3)$$

where:

$v(d')$ – is the vector space of the document d' of a training set of documents.

$v(d)$ – is the vector space of the tested document d.

In the multilabel kNN classifier class ranks of are calculated as follows:

$$rank(c,d) = \sum_{d' = S_k} I_c(d') \times \cos(v(d'), v(d)) \quad (4)$$

where:

S_k – is the set of k nearest neighbors of document d'.

$I_c(d') = 1$ – If (and only if) document d' belongs to class c. Otherwise, $I_c(d') = 0$.

Thus, the assigned document classes are ranked according to formula (4).

The classification of scientific resources consists of several steps:

- habituation of the classifier (normalization of terms of documents' vector space);
- selection of a document;
- classification of the document.

Classification of the document consists of the following steps: a linguistic analysis, extraction of terms and the document's vector space formation, calculation of k nearest neighbours of ranking classes.

As a result of the university's scientific resources processing the documents' profiles are formed. A document profile is defined as the vector of all relevant topics of its ontology:

$$PD(d) = (R_1^d, ..., R_c^d) \quad (5)$$

where:

R_c^d – are relevant topics c of document d.

Accordingly, the academic profile of a staff member is defined as the profile of all his publications:

$$PD(a) = (R_1^{da}, ..., R_i^{da}) \quad (6)$$

where:

R_i^{da} – are all the documents of the author.

The final step of the text classification is the formation of the document's semantic profile by creating the individuals of "Information resources" class in the ontology of scientific research activity. Figure 1 shows the composition of the university's scientific research ontology.

5 EXPERIMENTS AND RESULTS

The semantic portal implemented the possibility of search for any object of the ontology in the following classes: researchers, the university's research areas, events, key terms, organizations, departments, sub-departments, the university's publications.

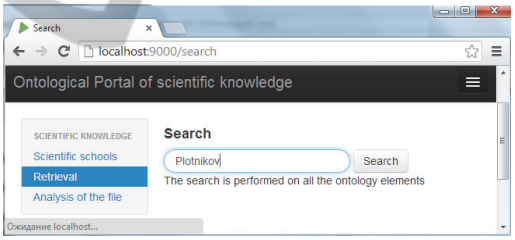

Figure 3: A screenshot of the result of the object search in the knowledge base.

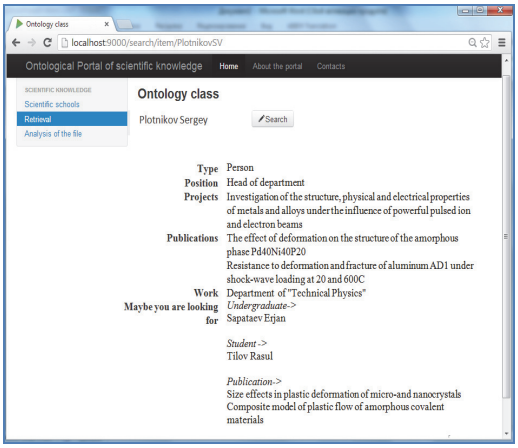

Figure 4: A screenshot of the dynamic page of the Researcher class instance.

The semantic portal search module supports the input of inaccurate user requests. Figures 3 and 4 show the search function result at the portal.

6 CONCLUSION AND PROSPECTS FOR FUTURE DEVELOPMENTS

The pilot project of the university's scientific activity semantic portal has allowed us to generate a fragment of the university's scientific knowledge base and to explore the functionality of the developed models and methods of text document processing.

"Solid State Physics" vocabularies have been formed by the developed methods of automatic extraction and key terms list selection from the body of scientific papers.

The topical classification of documents has allowed the researchers to create the university researchers' profiles and to implement a personalized search engine at the university.

The next stage of the research is the development of the semantic portal functionality and its implementation as a part of the university's scientific knowledge management system.

ACKNOWLEDGEMENTS

The work was performed under grant "The development of an e-university's ontological knowledge base", state registration number 0213RK00305.

REFERENCES

Allemang, D., Hendler, J., 2011. *Semantic Web for the Working Ontologist*. Morgan Kaufmann Publisher, Burlington, USA.

Altınçay, H., Erenel, Z., 2010. Analytical evaluation of term weighting schemes for text categorization. In *Proceedings of the Pattern Recognition Letters*, 1, pp. 1310–1323.

Bolshakov, E., Klyshinsky, E., Lande D., Noskov, A., Peskov, O., Yagunova, E., 2011. *Automatic processing of natural language text and computational linguistics.* MIEM Publishing House, Russia, 272 p.

Braslavsky, P., Sokolov, E., 2008. Comparison of five methods for extraction of terms of arbitrary length. In *Proceedings of International Conference "Dialogue" - Computational Linguistics and Intelligent Technologies*, vol. 7(14). Russia, pp. 67-74.

Ceci, F., Pietrobon, R., Gonçalves, A., 2012. Turning Text into Research Networks: Information Retrieval and Computational Ontologies in the Creation of Scientific Databases. *PLoS ONE*, vol. 7(1), pp. 1-9.

Cherman, E. A., Monard, M.C., Metz, J., 2011. Multi-label Problem Transformation Methods: a Case Study. *Electronic Journal CLEI*, vol. 14(1), pp. 4-13.

Du, M., Chen, X., 2013. Accelerated k-nearest neighbours algorithm based on principal component analysis for text categorization. *Journal of Zhejiang University-Science C-Computers & Electronics*, vol. 14(6), pp. 407-416.

Guarino, N., 2009. The Ontological Level: Revisiting 30 Years of Knowledge Representation. *Conceptual Modeling: Foundations and Applications*, pp. 52-67.

Jiang, J., Tsai, Sh., Lee, Sh., 2012. FSKNN: Multi-label text categorization based on fuzzy similarity and k nearest neighbors. In *Proceedings of the Expert Systems with Applications 39*, pp. 2813–2821.

Kryukov, K.V., Kuznetsov, O., Suhoverov, V., 2013. On the notion of a formal competency researchers. In *Proceedings of III International Scientific and Technical Conference – OSTIS-2013*, pp. 143-146.

Liu, Y., Loh Han, T., Sun, A., 2009. Imbalanced text classification: A term weighting approach. In *Proceedings of the Expert Systems with Applications*, vol. 36, pp. 690–701.

Lukashevich, N.V., 2011. *Thesauri in information retrieval tasks*. Moscow University Publishing House, Russia, 415 p.

Ma, J., Xu, W., Sun, Y., Turban, E., Wang, Sh., Liu, O., 2012. An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection. *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans*, vol. 42(3), pp.784-790.

Maier, R., 2007. *Knowledge Management Systems: Information and Communication Technologies for Knowledge Management*, Springer, 3rd edition.

Malarvizhi, P., Ramachandra, V.P., 2013. Multilabel classification of documents with MAPREDUCE. *International Journal of Engineering and Technology (IJET)*, pp.1260-1267.

Manning, Ch.D., Raghavan, P., Schütze, H., 2009. *Introduction to Information Retrieval*. Cambridge University Press.

Min, J., Josh, C.D., Buzhou, T., Hongxin, C., Hua, X., 2012. Extracting semantic lexicons from discharge summaries using machine learning and the C-Value method. *Proceeding of the AMIA Symposium*, pp. 409-416.

Pivovarova, L.M.,Yagunova, E.V., 2010. Extraction and classification of terminological collocations on the material of linguistic scientific texts (preliminary observations). In *Proceedings of Symposium: "Terminology and knowledge"* Russia, Moscow, [http://webground.su/data/lit/pivovarova_yagunova/Izvlechenie_i_klassifikatsiya_terminoligicheskih_kollokatsyi.pdf].

Science journal "*Solid State Physics*", [http://journals.ioffe.ru/ftt/].

Sedova, Y.A., Kvyatkovskaya, I.Y., 2011. Intelligent analysis of corps of scientific information. *Bulletin of the Astrakhan State Technical University. Series: Management, Computing and Informatics*, vol. 1, Russia, pp. 128-136.

Semantic search engines. [http://asknet.ru/Analytics/semantics.htm]

Shengyi, J., Guansong, P., Meiling, W., Limin, K., 2012. An improved K-nearest-neighbor algorithm for text categorization. In *Proceedings of the Expert Systems with Applications 39*, pp. 1503–1509.

State subject heading list of Scientific and Technical Information, [http://www2.viniti.ru/index.php?option=com_content&task=view&id=57&Itemid=6].

Thiagarajan, R., Manjunath, G., Stumptner, M., 2008. *Finding Experts By Semantic Matching of User Profiles*. HP Laboratories, [http://www.hpl.hp.com/techreports/2008/HPL-2008-172.pdf].