One Approach to the Problem Solution of Specialized Software Development for Subject Search

Dmitriy Grinchenkov, Darya Kushchiy, Anastasia Kolomiets
Platov South-Russian State Polytechnic University (NPI)
Prosveshchenie Str. 132, 346400, Novocherkassk, Russia
E-mail: grindv@yandex.ru, dkushchiy@rambler.ru, anastasia.srstu@gmail.com

***Abstract*— in the article relevance of system development for subject search using computational linguistics is considered. The basic principles of system functioning are defined. The principle of grammar development for information retrieval from the partially structured text in a natural language is considered. The ranging principle of results of information search is defined.**

***Keywords:* information retrieval, subject search, semantic web, grammar, natural languages, formal languages.**

I. Introduction

Nowadays Internet is one of the most important sources of information in all fields of knowledge. However constantly growing volume of various structure text data makes more difficult process of subject search and determination of result practical importance for the end user. This circumstance negatively affects quality of information support for research of any sort. Further, as an example, we will consider subject search of electronic educational resources [1] in texts of working programs.

The main modern search services on the Internet such as subject directories and search services by keywords show a number of shortcomings when scientific information search. Firstly, correlation of the document with one or another category is not fully automatized, and secondly search results are influenced by limitation of demand assignment language. Thus, development of the specialized software for subject search is rather actual nowadays. [2]

II. Structure of Search System

In very general case the system of subject search can be presented in the form of the information and operating structural elements. Their interrelation is carried out by some mechanism L (Fig. 1).

A set of internal states of system is formed by a set of analyzed internet resources IR. Their processing is defined by transition function H from one resource to another (by change of internal states) (1):

$$H:(Q \bullet IR)^h \Rightarrow IR\,. \qquad (1)$$

There are not only a set of entrance values Q (system request) and output values R (search results), but also operators of transformation its forms of representation for the user and the system F_Q and F_R in the structure.

Functioning of subject search system is time-spaced process T of information transformation from entrance value Q to output value R (2):

$$T:\{T \rightarrow Q; T \rightarrow R\}\,. \qquad (2)$$

Generation a set of results is realized by algorithm called for system exit function *E* (3):

$$E:\{(Q \bullet IR) \rightarrow R\}\,. \qquad (3)$$

Feedback mechanism for system regulation is denoted by parameter *F*. Thus system can be represented as an ordered set of elements (4):

$$S =< Q, R, IR, F_Q, F_R, L, H, T, F >. \qquad (4)$$

The main distinctive feature of analytical search system from simple search system is the way of demand assignment and rating of found resources considering context of chosen subject field. View of documents search form can be fully or partially taken from the traditional search system [3].

III. Form of Query View

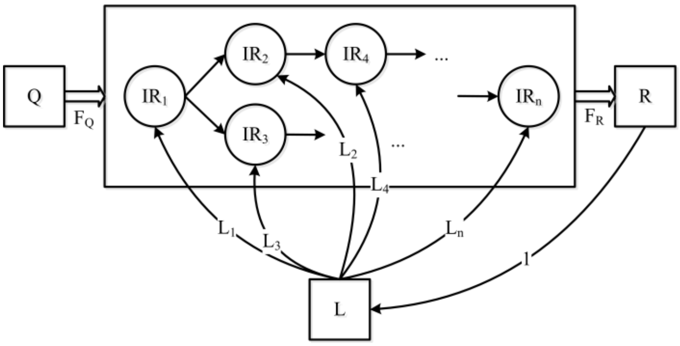

Fig. 1. Generalized structure of a search system.

In subject search inquire represents the model of subject field, where key words and relationship is determined. It is practical to use for it methods of computing linguistics [4].

As it was mentioned above, the working program of educational module serve as a source for search field in our case.

Formation of model sampling for creation of inquiry manually demands certain knowledge in subject domain and time. Those are the reasons for need in automation of this process.

Working programs of the Russian higher education institutions represent partially structured text which reflects subject domain and contains additional information on educational process with the repeating blocks. This fact causes low efficiency application of statistical methods for allocation of keywords by Tsipf method and the subsequent formation of terminological builds by t-score and MI-test [5][6][7]. In other words, for different disciplines will be formed almost identical retrieval of valid and syntactic words without display of a subject context For example, discipline, subject, section, literature, way, definition, target.

This problem is seen to be overcome by use of linguistic methods. In such documents linguistic analysis could be applied correctly only within the isolated text blocks. However despite some structuring, this text remains written in a natural language, and complexity of its analysis is caused by lack of the formal representation.

As it has been noted in some works [8][9], not all natural languages can be distinguished by regular grammars. Moreover, the existing languages may contain structures are recognizable only by context-dependent grammars [10].

However, use of context-free grammars at rules modeling provides good approximation to the truth and allows solving the majority of applied problems [11]. The semantic analysis of the text of the working program requires creation of the formal unified text structure, allowing fully displaying contents of structural blocks.

At the same time types and fields layout in the document from which text information is taken, define sense and communications of this information with other information in the document.

One more complexity consists of slippery character of information of a set of structural blocks. One reason lay in distinctive types of an academic load for different disciplines. For example, for one discipline the curriculum has only lectures and laboratory researches, and for another – lectures, laboratory researches, a practical training and a term paper. Other reason is covered in alternativeness and mobility of language norm borders and in statistical nature of separate information types [12].

It should be noted separately reasons for linguistic incomplete at formalization:

1. Continuous development of a natural language. It includes appearance of new language units, character change of the existing units and rules of their compatibility. Especially it is noticeable in sublanguages of new subject domains with not well-established terminology.
2. Language features of separate native speakers which could not be described and formalized today.

In Russian language there is much tension around this problem due to lack of rigidly regulated sentences construction.

Carrying out the text analysis of working programs the following stylistic features has been found: the text doesn't contain figural expressions, estimative adjectives, almost no adverbs; the natural language polysemy is minimized by use of in advance defined terms. The main language construction of text blocks is the grammatical basis with a number of additions.

The developing grammar of the working program leads to right-linear context-free grammar because of the choice of highly specialized area of a natural language and existence of attributes in grammar.

Actually this grammar is used for splitting a source text of the document for sections and processing most important of them for our task. For this purpose accurate observance of structure of the document is required, the working program consists of in advance defined sequence of sections.

Top level production rules serve for analysis of top level sections. Rules for analysis of sections consist of two parts: the first part serves for analysis of the section name, the second – for analysis of the section text content. Symbols of such grammar can have syntactic attributes. Names of semantic attributes are specified in attributes of nonterminal symbols. In attributes of terminal symbols syntactic text attributes can be specified in addition. Comparison of words at analysis is made taking into account their morphology.

We will consider a fragment of the developed grammar, provided by xml-format:

```
…<global-rule id="Section4" comment =
"4.SODERZHANIE DISCIPLINI ">
<rule><rulerefuri="#Section4Name"/>
<rulerefuri="#Section4x"/></rule>
  </global-rule>
  <global-rule id="Section4Name" sectionPart ="Name"
comment="Заголовок раздела 4"><rule><clauseType=
"NEOPRED"/></rule></global-rule><global-rule
id="Section4x" frame= "SubTitle" frameSlot="Title"
comment="4.1 KONTAKTANAYA AUDITORNAYA
RABOTA""or"comment="4.2 SAMOSTAYATELNAYA
RABOTA " "or"comment="4.3 KONTAKTNAYA
VNEAUDITORNAYA RABOTA"
</rule><ruleref uri="#Section4xContent" /></rule>
  </global-rule>
  <global-rule id="Section4xContent" section-
Part="Content"
comment=""><rule><rulerefuri="#Section42xInputs"
minOccurs=""/><rulerefuri="#Section42xOutputs"minOccu
rs=""/></rule>
  </global-rule>
  <global-rule id="Section4xInputs" comment="TEMA">
<rule><sentence/><clause/><rulerefuri="#Input"maxOccurs
="unbounded"/>
  </rule></global-rule> …
```

In structure of the working program keywords are highlighted. Those keywords determine ownership of section to the certain nonterminal.

In rules of grammar there are syntactic attributes and attributes which specify degree of the rule implementation:

1. Name – the text contain the name of the section.
2. Content – the text describe section contents;
3. Clause – clause;
4. Clause NEOPRED – the clause which does not have sense for the description of system structure;

5. Clause TIRE – a fragment with a dash;
6. GENIT_IG group – the nominative group connected by a genitive case.

The nominations accepted in the Penn Treebank project are used in names of nonterminal symbols. [13]

Sematic text representation consist of semantic representation of separate sentences, which elements are definition retrieved from the analyzed text and its semantic relations (Fig 2) [14][15].

Semantic representation of separate sentences is described by algebraic system, similar to the graph with definitions as a vertexes, any edge is marked to the semantic relations and connects those vertex-definitions which are with each other in this relation.

IV. Assessment of Search Results Relevance

The selection constructed according to the working program by search is considered as the reference text. Text representation of inquiry is formed using linguistic analysis methods at a search query. Data selection is carried out using normal form of words from special data structure containing information on word usage in texts of documents collection in which search is carried out. This structure of data contains information on documents texts according to text representation. To fill this structure texts of documents were preliminary analyzed by linguistic method [16][17].

Unlike of exact-match search function [18][19] offered approach will solve problem of sentential search by finding sentences corresponding to inquiry lexically and syntactic, but differ in a form and order of word usage (Fig 3).

By sentential search the sentences in the found documents are compared with the inquiry sentence so that at least one phrase will match to one in inquiry.

Under the match of phrases we mean presence of all syntactic links at the corresponding word usage as a part of phrases in reference sentence and found sentence.

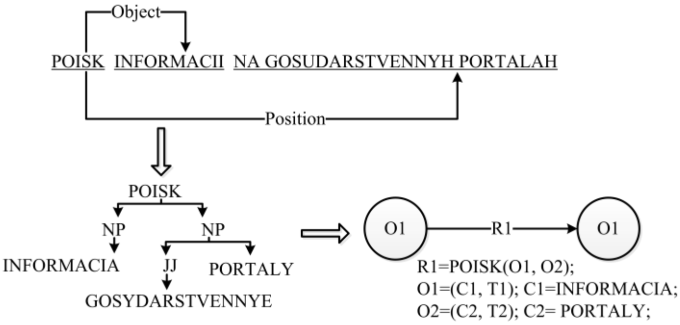
Object
POISK INFORMACII NA GOSUDARSTVENNYH PORTALAH
Position
POISK
NP
NP
INFORMACIA
JJ
PORTALY
GOSYDARSTVENNYE
O1
R1
O1
R1=POISK(O1, O2);
O1=(C1, T1); C1=INFORMACIA;
O2=(C2, T2); C2= PORTALY;

Fig. 2. Semantic image of offers.

The principle of results ranging of information search is defined (5) where the following notations have been accepted:

ref(r_i) – the document weight given by basic algorithm of ranging BM25 [20];

List – amount of the given results on the search page;

AllList – amount of all search results satisfying to inquiry *q*;

col_key(K_i,C_q) – the weight coefficient characterizing collocations among matches of a set of keywords to the text of the document.

$$F(r_i) = ref(r_i) + col\_key(K_i, C_i);$$

$$yes\_col(r_i) = ref(r_i)\cdot\left(1-\left(\max ref - ref(r_i)\right)\right);$$

$$not\_col(r_i) = \left(\max ref - ref(r_i)\right)\cdot\left(1 - ref(r_i)\right);$$

$$col\_key(K_i, C_i) = \begin{cases} \left(2^{yes\_col(r_i)} - 1\right)\Big/ norma, K_i \in C_i \\ \left(-2^{yes\_col(r_i)} - 1\right)\Big/ norma, K_i \notin C_i \end{cases};$$

$$norma = 2^{\max ref}\sum_{j=1}^{List}\log_2(1+j);$$

$$\sum_{j=1}^{AllList} ref(r_i) = 1; \max ref = \max_R\left(ref(R)\right).$$

(5) Ranging function has to satisfy orderliness property, which means that for couple of documents (r_i, r_j) holds $F(r_i) \geq F(r_j)$, if the document *rr*$_i$ more corresponds to inquiry q, than the document *rr*$_j$.

The complex algorithm of relevant information finding is presented with the help of the flowchart (Fig 4).

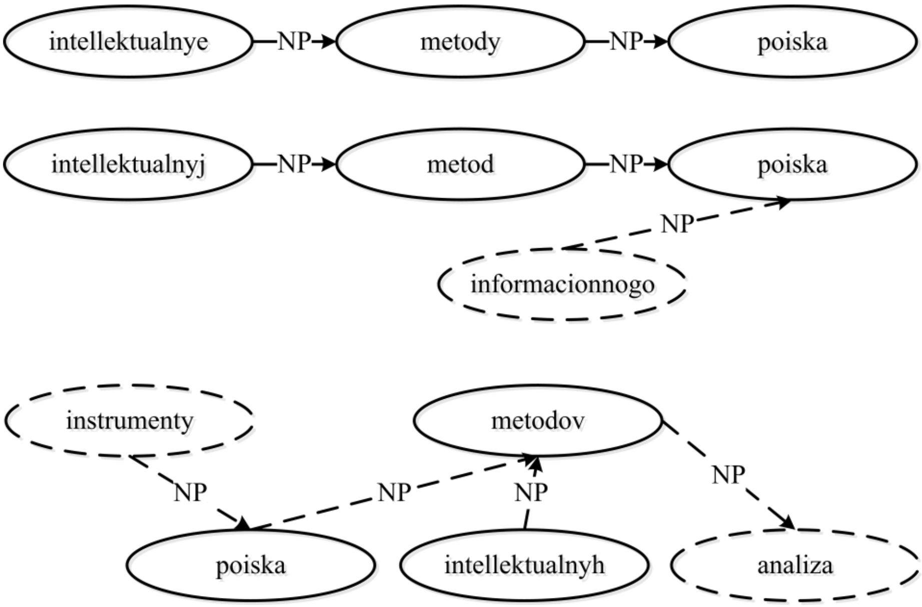

Fig. 3. Options of relevant search results.

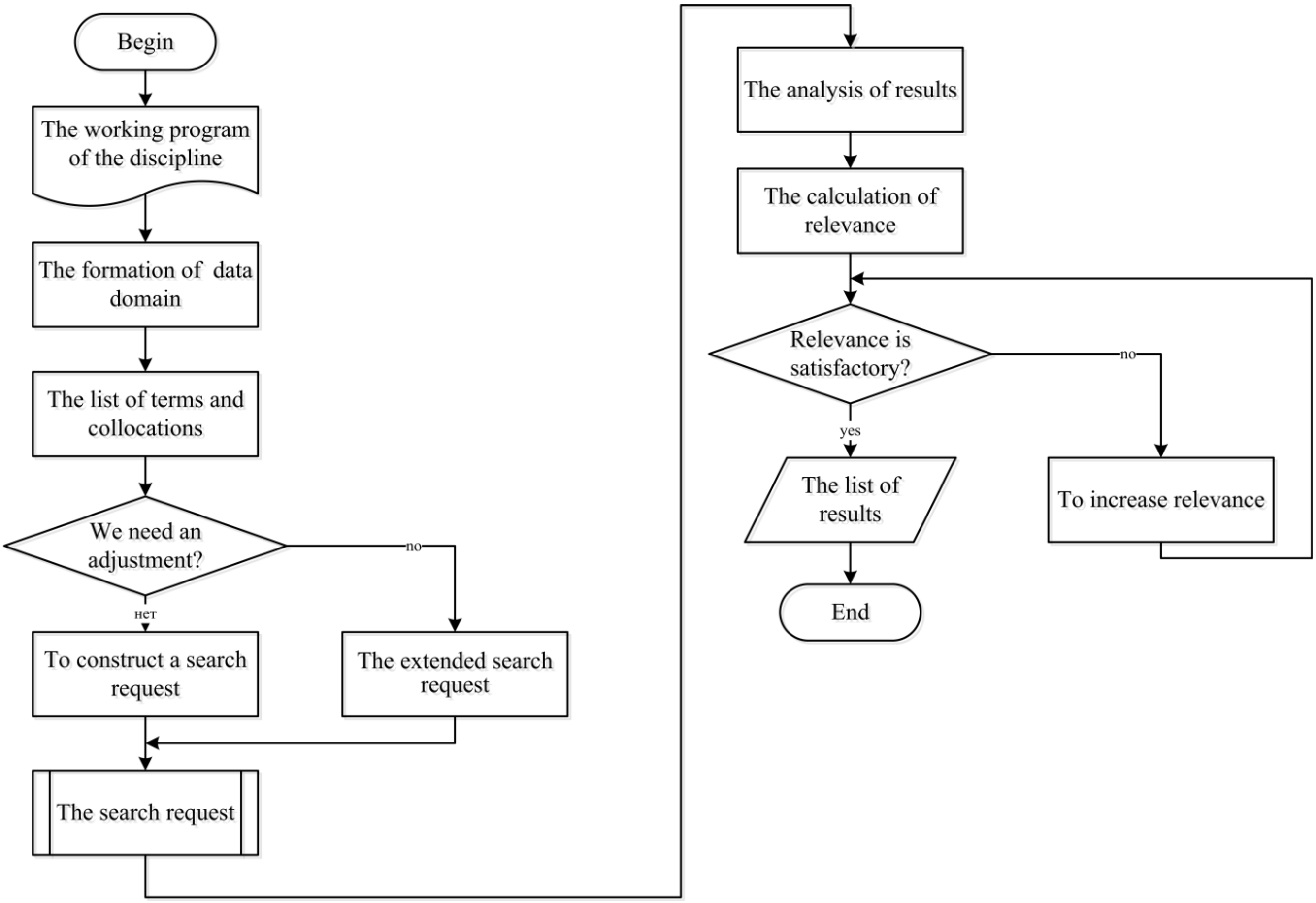
Begin
The working program of the discipline
The formation of data domain
The list of terms and collocations
We need an adjustment?
no
нет
To construct a search request
The extended search request
The search request
The analysis of results
The calculation of relevance
Relevance is satisfactory?
no
yes
The list of results
To increase relevance
End

Fig. 4. Complex algorithm of relevant information finding.

V. Conclusion

Comparison of efficiency of simulated system and traditional information retrieval systems has been carried out with the aim of efficiency assessment of the developed models and algorithms [21]. Results were compared to a reference system – the hypothetical system finding all available relevant to this inquiry documents. Comparison was carried out by the number of the relevant documents issued by systems. (Fig 5)

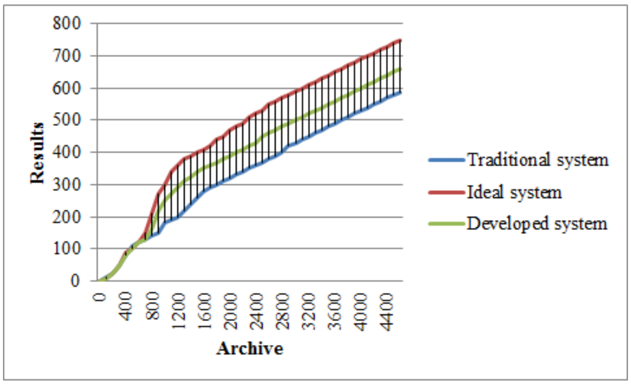

Fig. 5. The number of the issued relevant documents

It should be noted that distinctions in efficiency are seen in process of increase in volume of the worked out array of data. At the small sizes of archives (up to 800 documents) distinctions in results is becoming more and more obvious. The collection of pages of Wikipedia (about 2500 documents) was used for simulation. At such size of archive the difference in number of the issued relevant documents makes about 15-20%.

The further plan of work includes working out increase in efficiency of the developed algorithms on large volumes of text collections.

References

[1] D.V. Grinchenkov, D.N. Kushchiy, “The methodological, technological and legal aspects of using the e-learning resources,” J. University News. North-Caucasian Region. Technical Sciences Series, no.2, pp.118-123, 2013.

[2] D.V. Grinchenkov, D.N. Kushchiy, “Topicality and principles of construction of the intellectual information system for forming methodical support of educational disciplines on the basis of internet resources,” J. University News. North-Caucasian Region. Technical Sciences Series, no.3, pp.114-119, 2014.

[3] D.V. Landeh, A.A. Snarskij, I.V. Bezsudnov, Internetika: Navigation in difficult networks: models and algorithms, Moscow: LIBROKOM (Editorial URSS) Publ., 2009, pp. 142-153. (rus)

[4] I. A. Tikhomirov, I.V. Smirnov, “Applying linguistic semantics and machine learning methods to search precision improvement in search engine “Exactus”,” Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue-2009”(Bekasovo, Russia, May 2009), issue 8(15), 2009, pp. 483-487.

[5] D. Gildea, D. Jurafsky, “Automatic labeling of semantic roles,” J. Computational Linguistics, vol. 28, no. 3, pp. 245-288, 2002.

[6] D.V. Grinchenkov, D.N. Kushchiy, “On selection of keywords in the contents of the framework curriculum of academic disciplines,” Traditions of the Russian engineering school: yesterday, today, tomorrow: collection of scientific articles on problems of higher school, Novocherkassk, Platov South-Russian State Polytechnic University (NPI) Publ., Nov. 2015, pp. 114-117. (rus)

[7] D.V. Grinchenkov, D.N. Kushchiy, “The formation of fuzzy collocation on the basis of semantic analysis of the framework curriculum of academic disciplines,” Problems of modernization of engineering education in Russia: collection of scientific articles on problems of higher school, Novocherkassk, Platov South-Russian State Polytechnic University (NPI) Publ., Oct. 2014, pp. 298-300. (rus)

[8] N. Chomsky, Syntactic Structures, 2nd ed., Berlin: De Gruyter Mouton, 2002, pp. 112-115.

[9] R. Huybregts, “The weak inadequacy of context-free phrase structure grammars,” In Proc. Van Periferie naar Kern, Dordrecht, The Netherlands, 1984, pp. 81-99.

[10] M. Shieber, “Evidence against the context-freeness of natural language,” J. Linguistics and Philosophy, vol. 8(3), pp. 333-343, Aug. 1985.

[11] M.M. Gavrikov, A.N. Ivanchenko, D.V. Grinchenkov, Theoretical bases of development and realization of programming languages. The manual for students of the higher education institutions which are trained in “The software of computer facilities and the automatized systems” direction of training of certified specialists “The information scientist and computer facilities”, Moskow: KNORUS, 2010, pp. 134-149. (rus)

[12] M. Pennacchiotti, P. Pantel, “Entity Extraction via Ensemble Semantics,” Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), Singapore, Aug. 2009, pp. 238-247.

[13] M.P. Marcus, M.A. Marcinkiewicz, B. Santorini, “Building a large annotated corpus of English the Penn treebank,” J. Computational Linguistics, vol. 19, no. 2, pp. 313-330,1993.

[14] R. McDonald, F. Pereira, K. Ribarov, J. Hajic, “Non-projective dependency parsing using spanning tree algorithms,” Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05), Stroudsburg, PA, USA, 2005, pp. 523-530.

[15] R. Mcdonald, J. Nivre, Y. Quirmbach-Brundage et al., “Universal dependency annotation for multilingual parsing,” Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, vol. 2, Aug. 2013, pp.92-97.

[16] A.V. Sokirko, “ Bystroslovar: morphological prediction new Russian words using very large corpora,” Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue-2010”(Bekasovo, Russia, May 2010), issue 9(16), pp. 450-456, 2010.

[17] Leontyeva N. N. About theory of automatic understanding of natural texts. P.3: Semantic component. Local semantic analysis, Moscow: MSU Publ., 2002, pp. 27-32. (rus)

[18] D.V. Grinchenkov, D.N. Kushchiy, “Principles of software development for support of decision-making based on integreted expert estimates ,” J. University News. Electromechanics, no. 5, pp. 69-73, 2012.

[19] R. Johansson, P. Nugues, “Dependency-based syntactic-semantic analysis with PropBank and NomBank,” Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL '08), Stroudsburg, PA, USA, pp. 183-187, 2008.

[20] K. Sparck Jones, S. Walker, S. E. Robertson, “A probabilistic model of information retrieval: development and comparative experiments (part 1),” J. Information Processing and Management, vol. 36(6), pp. 779-808, 2000.

[21] I. Luzyanin and A. Petrochenkov "Regarding Information Systems Dependability Analysis" Proceedings of the 3rd International Conference on Applied Innovations in IT (2015). Jg. III. Koethen : Hochschule Anhalt, 2015, pp. 7-11 (DOI: 10.13142/kt10003.02, Anhalt University of Applied Sciences Digital library).