EURASIAN JOURNAL OF MATHEMATICAL AND COMPUTER APPLICATIONS
ISSN 2306–6172
Volume 5, Issue 1 (2017) 74 – 86

THEMATIC CLASSIFICATION OF THE THESIS ABSTRACTS

Leonova Yu.V., Fedotov A.M., Fedotova O.A.

Abstract In this paper the method of thematic classification of the thesis abstracts. It uses specially constructed proximity measure documents, taking into account the specificity of the subject area. The values of the weighting coefficients in the formula for calculating the proximity of the proposed measures are defined a posteriori reliability of the corresponding scale data.

Key words: scientific thematic, thesis, classification, proximity measure

AMS Mathematics Subject Classification: 68T50, 91F20

1 Introduction

In this work we consider the method of thematic classification of the thesis abstracts. For this is used a specially built proximity measure of the documents, taking into accounts the specification of the subject. As a scale for determination of the measure me suggest to take the characteristics of the structural attributes of the abstracts description (scientific newness, principle thesis, etc). The values of the weighs coefficients in the formula for the calculation of the proximity measure are determined by the supposed a posteriori reliability of the date of corresponding scale.

2 The subject area

Scientific thematic of the thesis is divided in classes called «specialties». In the Russian classification of professions of higher scientific qualification specialties use 3-level classification [1] of the dissertations including the next levels:

- Branch of the science,
- Group of specialties,
- Specialty.

This way all the system of knowledge is divided in 25 classes and reflects the common differentiation of sciences in physics-mathematical, biological, chemical, technical, agricultural, medical, economical, etc.

Every class of the linguistic level is characterized by document called passport of the specialty. The passport of the specialty maintains the formal criteria of the correspondence of thematic of the thesis of the definite specialty, also the determination of area of investigation and the list of the points to which the dissertation must corresponds.

However, the formulations of the same results with small differences can correspond to different specialties. Unformal criteria of the correspondance of subject of the thesis and specialty are determined by the dissertation Council who give different demands to the thesis within the same specialty.

Specialty may be characterized by the set of the key terms with bond of the type «parent-descendant» represented as thesaurus. For every specialty the common technical thesaurus built in the system must be added by its specific thesaurus corresponding to the specialty. Classificational features of the thesis abstract, based on the key terms are described bellow.

For the classification it's necessary to fulfill the comparison of classification signs of the thesis abstracts, represented by the key terms, introducing the passports of specialties, to decide to what class belongs the thesis abstract.

3 Feature area

Feature area for classification of the thesis abstracts is represented by the totality of the keywords, terms of natural language and the set of relations between them. All the chosen relations, getting the qualities of weight, are described in the systematic dictionary of terms — thesaurus [2]. Every document is described by the set of its characteristics, named features. Determination of the features — formation of the vector of typical signs of the document, further used for taking the decisions in the work with a document. Feature area represents a totality of metadata.

The formation of the feature area is an important step in the problem of classification. It is necessary to select the most significant features, containing the largest information of the classified documents. The usage of too big set of features makes worth the exactness of classification as the features contain too much extra information. Feature area is formed on the basis of analysis of the text of thesis abstracts and metadata containing in bibliographic description — name, UDC, opponent information, scientific curators, etc. Features space is divided into 5 types. Each type is associated with features of a certain weight, indicating it's importance [3].

1) Code of UDC and dissertation specialty. The weight of this features is not high because here the correctness of the correspondence of the UDC code and specialty code is checked.

2) Subjects reflected in the thesis abstracts — persons: curators, opponents, organizations, specialty, dissertation Council.

   Subjects are determined by sets of features — key terms having weights.

   Specialty is determined by the code of specialty and is additionally characterized by key terms from the passport of specialty.

   Scientific curator is characterized by the set of specialties with weights determined by priorities of scientific curator. For the opponent the list of specialties he can represent is equivalent. To organization and dissertation commission also corresponds the list of equivalent specialties.

3) References on the close works. The person (subjects) are chosen from reference of the author. The persons get the list of key terms.

4) Key terms got from text structural parts of the thesis abstracts: scientific newness; actuality; aims and problems; theses of the dissertation; object and subject of investigation; theoretical and practical importance of the work; methodology and of investigation; degree of reliability and approbation of the results; representation of the work.

   Key terms from different parts of thesis abstracts have different weight. For example, practical use has low weight as in this part the thematic direction is not reflected, and theses of the dissertation — highest.

5) Keywords mentioned by author and expert. Author keywords are not always mentioned correct, so have a low weight.

   Thus the feature area is represented as the lists of key terms and importance. Every key term gets the weight vector (importance) of the parts, where this term is met. In our case we use the following order of document treatment. There is a set of categories (themes) and a new document comes in for which we must determine the list of corresponding categories. If the document does not correspond to any category it is thrown away. On the multitude of categories the theoretic-multiple relations may be established. For example, the multitudes of documents forming the categories, can cross or not, thus the same document can belong to some categories. Searching of thesis abstracts, corresponding to the definite specialty occurs in direction of decreasing importance of classification features. The use of classification features with weights makes the exactness of classification higher.

4 Proximity measure

The most popular variant of document classification is bezel classification formed by Indian bibliographer Sh.P. Ranganatan [4]. The objects are classified due to some independent features (bezels) at the same time. Applying to figure documents as the bezels the elements of metadata are used, also the key terms. This suggested approach to the classification is based on the conception of analogy of the document, determined in the work [5]. We limit the work with consideration only of key terms aggregated according to types of features. Quantity characteristic of proximity measure is determined on the multitude of documents D in the next way [6]:

$$m : D \times D \to [0, 1]$$

in the case of complete analogy function m gets value 1, complete difference — value 0.

Consider two documents d_1 and d_2. Let $T = \{t_i\}_{i=1}^{M}$ the list of key terms regulated in some way (for example, lexicographically) present in both documents, considering repetitions (M — common quantity of key terms). Calculation of proximity measure

occurs according to the following formula:

$$m(d_1, d_2) = \sum_{i=1}^{M} \alpha_i m_i(d_1, d_2),$$

where i — number of element of metadata (key terms), $m_i(d_1, d_2)$ — proximity measure by i-element (by i-scale), α_i — weight coefficients. As in described situation practically all the scales are nominal (contain of discrete text values), the degree of analogy by i-scale is determined the next way. If the values of i-elements of documents coincide, then the proximity measure is equal 1, otherwise — 0.

Weight coefficients must satisfy the following conditions $\sum_{i=1}^{M} \alpha_i = 1$, $\alpha_i = \alpha_j$, if value of the term t_i coincides with the value of the term t_j.

Let $P = \{p_k\}_{k=1}^{N}$ — the list of unique key terms be inside both documents, M_k — number of repetitions of term p_k. Then proximity measure may be rewritten in this way:

$$m(d_1, d_2) = \sum_{k=1}^{N} (\alpha_k * M_k)(m_k / M_k),$$

α_k — weight coefficient, corresponding to the value of the term p_k,
m_k — number of coinciding of the term p_k in documents d_1 and d_2.

We get new weight coefficients $\beta_k = \alpha_k * M_k$, which already characterize the concrete key term. It is easy to see that

$$\sum_{k=1}^{N} \beta_k = 1.$$

Note that here we automatically get that weight coefficient is proportional to frequency of term meeting. Except this while setting the measure the fact must be taken into account that the values of weight coefficients β_k are determined by the supposed a posteriori reliability of the data of the corresponding scale and in definite cases one of the coefficients can be increased with proportional decreasing of the rest. For example, complete (or even «almost full») coincide of the value of some attribute of the document and document can have more weight in case when the amount of values of this attribute in the document is big enough (as compared with the case when document has only one).

5 Description of algorithm of classification

Algorithm of determination of thematic belonging of text of thesis abstracts is based on four procedures: *LemmatizeText, ParseText, FindThemeCover* and *CalculationRelevance*. See in details.

1) *LemmatizeText* — simplified morphological analysis or lemmatization — searches lexeme from multitude W of dictionary lexemes (see Table 1) by prefix of lexical form.
Procedure: LemmatizeText (word).
Input: word.
Output: term in normal form $L \subseteq W$ oneword terms of dictionary in normal form, corresponding to word.

Table 1: List of terms thesaurus for 4 themes

Theme1	**Theme2**	**Theme3**	**Theme4**
Recognition of images	Speech command	Cartography	Domain analysis
Treatment of figure images	Recognition of the speech	Spatial-time modeling	Ontological modeling
System of treatment of images	Dictionary of commands	Spatial structure	Fuzzy integrals
Automatic method of treatment and recognition of images	Recognition of the voice commands	GIS-technology	Making decisions
Iconic	Speech signal	Landscape analysis	Fuzzy measures
Algorithm of analysis of images	Algorithm of recognition of speech	Landscape geophysic method	Objective decision making
Searching images	Noise composition	Spatial modeling	Language knowledge
Computer seeing	Recognition of the fused speech	Spatial-time structure	Treatment of expert knowledge
Treatment of graphic information	Recognition of keywords	Geoinformation analysis	Expert information
Pattern recognition	Method of recognition of speech	Vector format	Methods of presentation and processing of knowledge
Machine recognition of handwritten symbols	Speech automatic recognition	Landscape differentiation	Method of modeling of objects
Searching image by content	System of voice direction	Spatial distribution	Theory of fuzzy measures
Finding image	Recognized words	Array of geoinformation	Knowledge representation
Medical image	Speech expression	Geoinformation zoning	Fuzzy set theory
Treatment of image	Speech transcription	Geoinformation control system	Thesaurus
Searching the graphic file by content	Speech automatic treatment	Methods of interpolation of geospace data	Concept
Digital image	Modeling of speech signal	Spatial-distributed information	Ontological approach
Compression of graphic data	Model of speech signal	Geoinformatics	Domain concepts
Extraction of image from database	Sound speech interpretation	Geoinformation projects	Attribute concepts
Pixel of image	Interpretations of sounds	Spatial data	Domain ontology

Discrete mapping	Acoustic model of speech signal	Geoinformation technologies	Inference machine
Chaotic shuffle of pixels	Interpretation of sounds	Earth remote sensing	Isolation concepts
Chaotic dispersed pixel	Acoustic sign of sound	Geostatistical analysis	A set of relations between terms
Process of image recognition	Segmentation of speech stream	Spatial organization	Set of attributes of terms
Technical vision	Apportionment of acoustic correlates	Manipulation map information	Domain knowledge
Three stage	Acoustic different sign	Raster data	Model driven architecture
Processing of three-dimensional image	Phonetic characteristic of sound of speech	Graticules	Development of domain ontology
Block of images	Automatic transcription of Russian oral speech	Chart projection	Metaontologies
Image search process	Acoustic correlate of resonance	Projection images	Structure of subject
Image presentation	Location of formant	Spatial Reference	Information modeling

2) *ParseText* — searching multiword terms.
Procedure: ParseText(words).
Input: words — sequence of words, corresponding to the summarized template describing word combinations of thesaurus. The summarized template connects groups of words by community of morphological signs, for example: [Adv] + [Subject].
Output: found term of thesaurus.
First two steps form terminological cover of the text — ordered multitude of the terms of thesaurus, found in concrete text.

3) *FindThemeCover* — builds the thematic cover of the text, forms the model of the text and data for classification.
Procedure: FindThemeCover(t).
Input: $t \in T$ — terminological cover of the text with weights corresponding to the terms in thesaurus.
Output: thematic cover $Y \subseteq t \times T$ — relation which connects themes from T with terms from t.

4) *CalculationRelevance* calculates the relevancies of subjects and selects the most fitting subjects.
Input: Y — thematic cover of request.
Output: multitude $\{< \tau_1, v_1 >, \ldots, < \tau_n, v_n >\}$ of subjects $\tau_i \in T$ with corresponfing proximity measures $v_i \in [0, 1]$.

6 Testing of the algorithm of classification. Methodic

From teaching set we remove all the documents of rubric which are present in the test, but do not participate in teaching. Variants of outcomes for the document:

1) *Correct*: the document fits the rubric
2) *Stranger*: the document is determined as strange
3) *Mistake*: the document does not fit the rubric
4) *Correct_strange*: the correct document was determined as strange by mistake
5) *Strange_correct*: the strange document was determined as correct by mistake

Outcomes 1 and 2 correspond to the correct work of the algorithm, the rest — mistaken.

Evalution:

$$Accuracy = \frac{Correct}{Correct + Mistake + Strange\_correct}$$

$$Completness = \frac{Correct}{Correct + Mistake + Correct\_strange}$$

7 Practical results

As initial data for the test of the algorithm were used the thesis abstracts of dissertation in 4 themes (see Table 1): «Recognition of the images» (Theme1), «Recognition of the speech» (Theme2), «Geoinformation systems» (Theme3), «Ontologies, description of the object area» (Theme4). Every reference set for every subject consisted of 30 thesis abstracts. Forming of the list of key terms (dictionary) is the apart problem [7].

For example, the dictionary of key terms can be formed by expert on the basis of his knowledge of the object area. In our case the list was formed on the basis of the texts of standard thesis abstracts, it's volume was 192 keywords.

Classification was carried out due to the following algorithm: Initially for every theme on the basis of reference set was found a centroid — typical set of keywords with weights, which was later used for comparison. Further was calculated the proximity measure of the tested thesis abstracts to centroid to of the class (theme).

8 Results of the test

On the input system got 4000 unfamiliar texts of thesis abstracts. For classification all text of thesis abstract was used, from which the most important keywords were selected. Proximity measure was calculated from selected key terms in dictionary for every theme. In classification information was selected from the following sections of thesis abstracts:

- Actuality of the theme of investigation.
- Aim and problems; scientific newness.
- Object and subject of investigation.

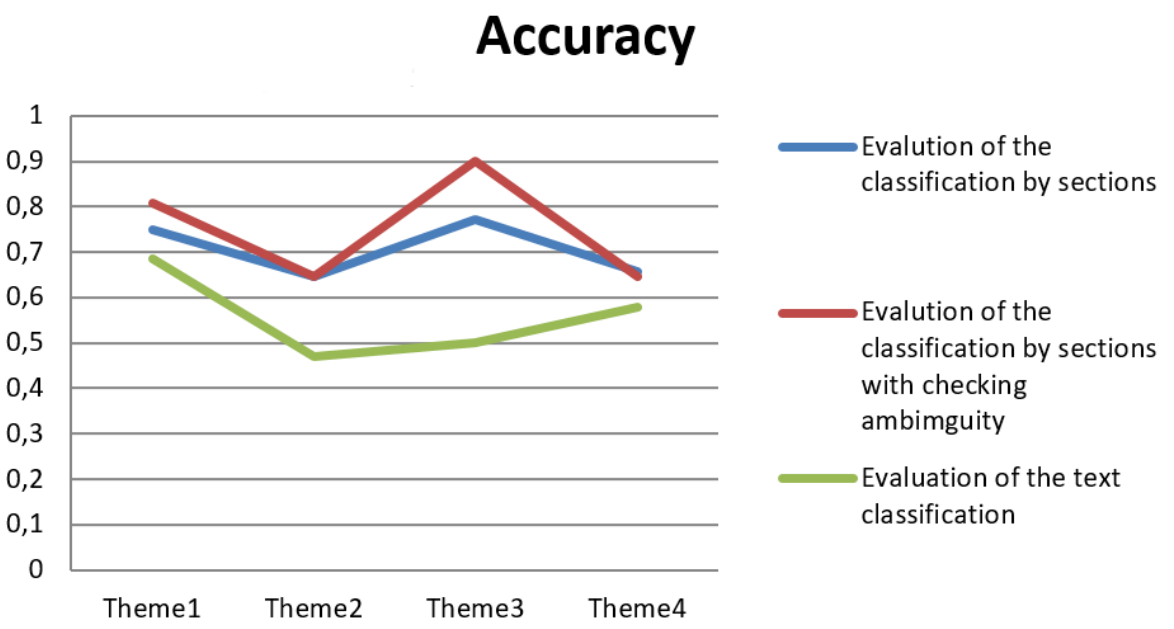

Figure 1: Accuracy of finding of the *Correct* documents.

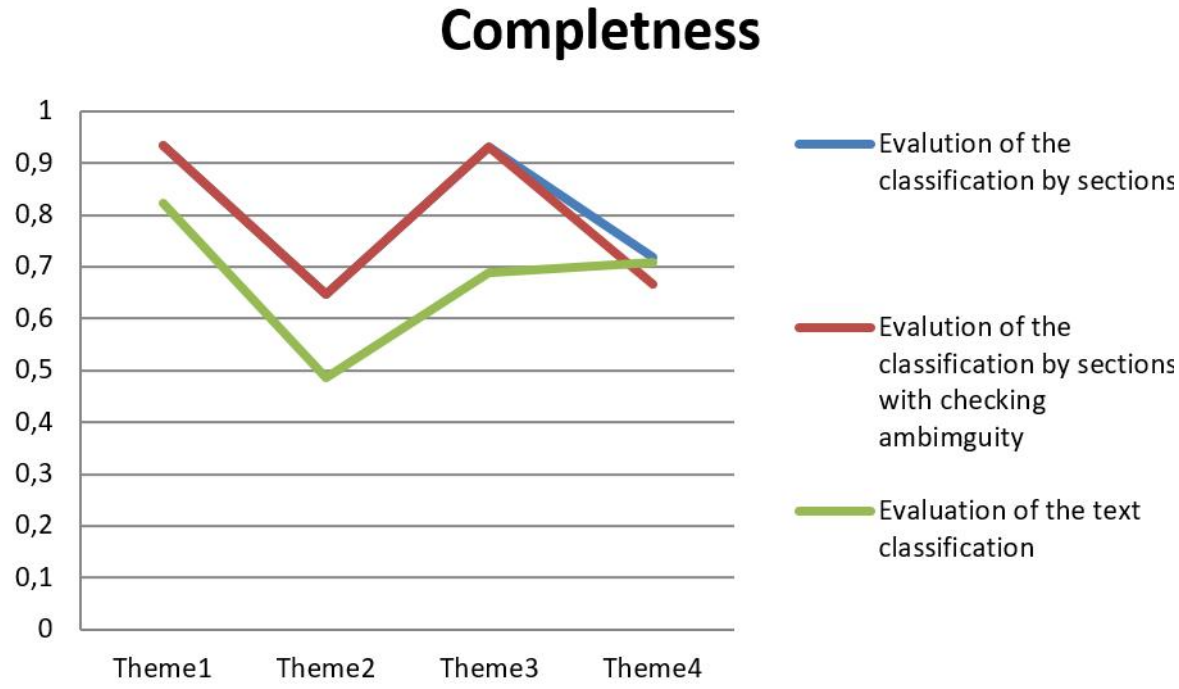

Figure 2: Completness of finding of the *Correct* documents.

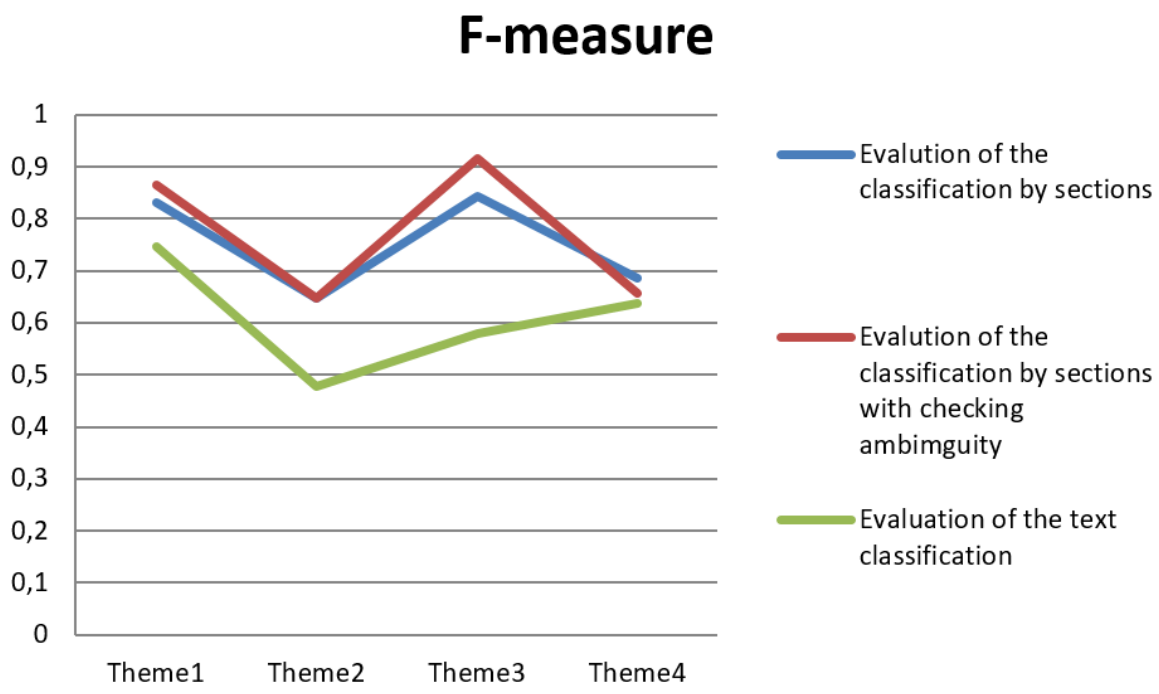

Figure 3: F-measure of finding of the *Correct* documents.

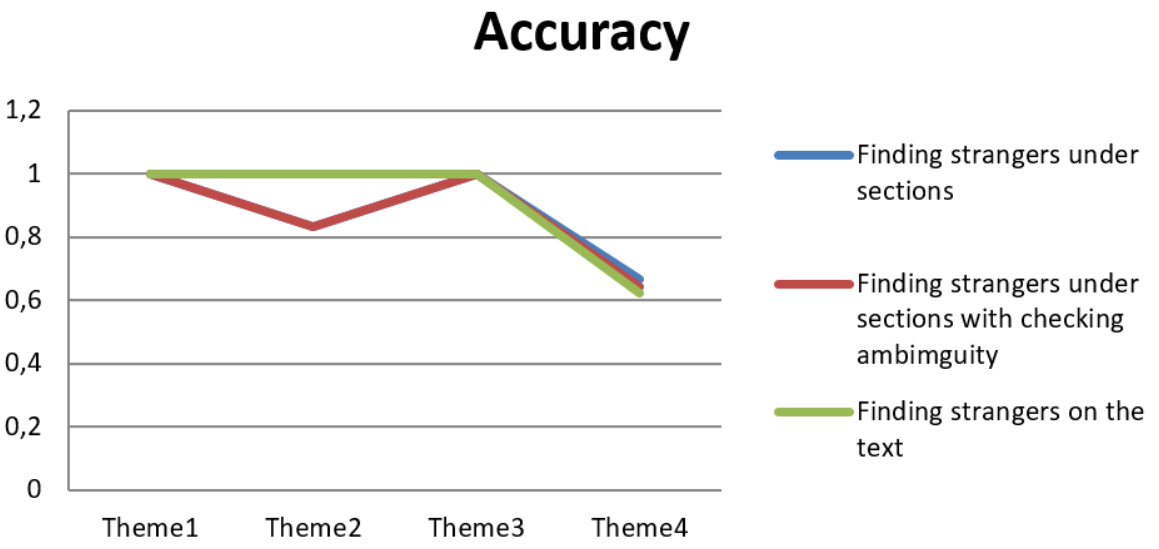

Figure 4: Accuracy of finding of the *Stranger* documents.

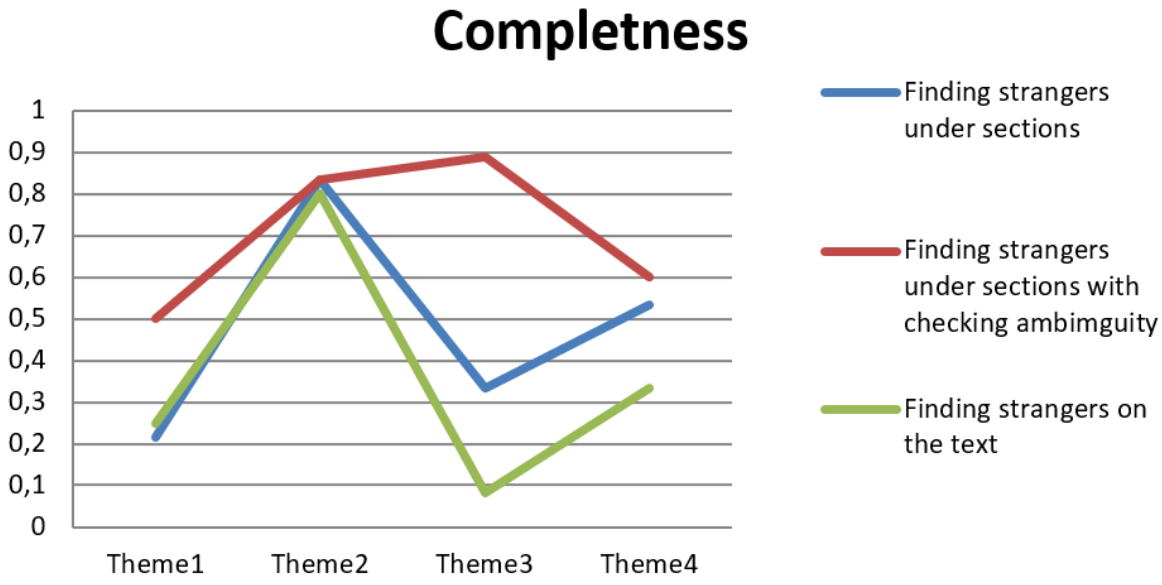

Figure 5: Completness of finding of the *Stranger* documents.

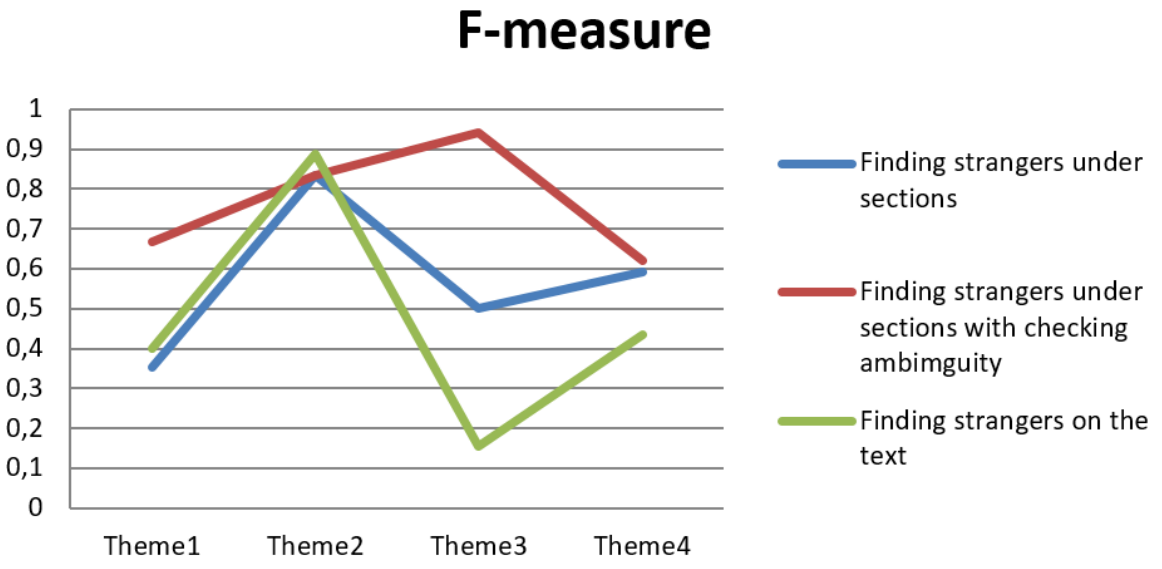

Figure 6: F-measure of finding of the *Stranger* documents.

- Theoretical and practical importance of the work
- Methodology and methods of investigation
- Theses of dissertation.
- Degree of reliability and approbation of the results.

For every part a proximity measure was calculated, the resulting proximity measure was calculated as middle value. The test of the algorithm was carried out in 3 regimes:
1) classification of section of thesis abstracts;
2) classification of section with checking ambiguity of terms — if the term belongs to some themes, the thematic belonging of the neighboring is checked;
3) classification of all the text of thesis abstracts without selection of sections.

Belonging of thesis abstract to the theme is determined by exceeding of proximity threshold between the tested thesis abstract and theme centroid.

It is established experimentally that is threshold value is more then 0.83, then the selected thesis abstracts relate to this theme, with the number not preceded thesis abstracts devoted to this theme as not more then 5%. If the threshold value of proximity measure does not exceed 0.17, the thesis abstract does not relate to this theme.

The Figures 1-6 represents the result of accuracy, completeness and F-measure obtain from test of the algorithm. As seen, the method of searching by text of thesis abstract has the best accuracy in searching the strange documents, but worst searching completeness and F-measure. The method of searching of sections with checking ambiguity of terms has the best characteristics.

The worst accuracy parameter corresponds to Theme3 — geoinformation systems what is caused by the presence of some terms, like «spatial distribution», «spatial structure», etc in the texts of chemical branch (see Figure 7). The supplement of dictionary with chemical terms makes the accuracy higher.

- **Petukhov Alexander Sergeevich. Syntheses, spatial structure and properties of 7-nomial acetates of pyridoxine: abstract dis. ... Cand.Chem.Sci: 02.00.03 / Kasan state Univ. - Kazan 2004**
- **Veselovskyi Alexander Vladimirovich. Computing modeling of active centers of monoaminoaxides and creation of inhibitors with definite selectivity: Dis. ... Dr.Biol.Sci: 03.00.04: M., 2004**

Figure 7: Examples of thesis abstracts creating noise for Theme3.

9 Conclusion

On the basis of obtained data we can make the following conclusions. Algorithm of classification of the whole text of thesis abstracts gives not bad results in case when the *Strange* documents must be selected. On practice just this is necessary in most

cases. However in case when it is known that every document has a theme, it looses to two other algorithms carrying out the classification by sections of thesis abstract. Algorithm of classification of sections with checking ambiguity of the terms represents itself not worse in search of strange documents both as algorithm of classification of text and algorithm of classification of sections. In tests algorithm of classification of whole text of thesis abstracts in great extent prevails with the algorithm of classification of sections with checking ambiguity of the terms is more prominent in comparison with other algorithms in searching the *Correct* documents.

Acknowledgement

The work is supported by a Russian grant of Leading Scientific Schools ScS-7214.2016.9.

References

[1] *Russian classification of professions of higher scientific qualification specialties (OK 017-2013),* 2013. — p. 34.

[2] Gilyarevsky R. S. *Categories of information as a tool for navigation* / R. S. Gilyarevsky, A. V. Shapkin, V.N. Beloozerov. — SPb.: Profession, 2008. — 352 p.

[3] Novikov A. M., Novikov D. A. *Methodology of scientific research.* — M .: LIBROKOM, 2010 — 280 p.

[4] Ranganathan Sh. R. *Classification of the colon. The main classification* / per. from English. M.: SPSTL USSR, 1970.

[5] Fedotov A. M., Barakhnin V. B., Zhizhimov O. L., Fedotova O. A. *Information System Model to support the scientific and pedagogical activity* // Vestnik of the Novosibirsk State University. Series: Information Technology. — 2014 — vol.12. — в„– 1. — P. 89–101. — ISSN 1818-7900.

[6] Voronin Yu. A. *Start similarity theory.* Novosibirsk: Nauka. Sib. Department, 1991. 128 p.

[7] Leonova Yu. V., Fedotov A. M. *Extracting knowledge from texts and facts abstract* // V International Conference "System Analysis and Information Technologies" (SAIT 2013): Conference Proceedings (Krasnoyarsk, Russia, September 19–25, 2013). — Krasnoyarsk: ICM SB RAS. — T. 1. — 2013. — P. 232–242.

Yu.V. Leonova,
Institute of computation technologies of SB RAS,
Novosibirsk, 630090, Russia,
Email: `juli@ict.nsc.ru`,

A.M. Fedotov,
Institute of computation technologies of SB RAS,
Novosibirsk state university,
Novosibirsk, 630090, Russia,
Email: `fedotov@sbras.ru`,

O.A. Fedotova,
State Scientific and Technical Library of SB RAS,

Novosibirsk state university,
Novosibirsk, 630090, Russia,
Email: ofedotova@ict.nsc.ru.

Received 13.12.2016, Accepted 21.01.2017