BUILDING THE LIBRARY CATALOG SEARCH MODEL BASED ON THE FUZZY SIMILARITY RELATION

Liliya Vershinina, Mikhail Vershinin, Andrej Masevich

Abstract: *We describe our approach to building the model of the search in libraries' catalogues based on the fuzzy similarity relation. To construct the model we carried out an experimental search to the variants of name in the catalogs of two libraries - the German National Library and the National Library of France. The model we constructed is based on the result of our experiment.*

Keywords*: fuzzy similarity relation, names transliteration, authority control, library catalogs*

ACM Classification Keywords*: H.3.6 Library Automation H.3.3, Information Search and Retrieval, 1.5.1 Pattern recognition. Fuzzy set*

Introduction

The implementation of fuzzy logic in the electronic catalogs is one of the requirements for the software of current library information systems.

Taking as examples the catalogues of two national libraries, we described from the users' point of view a search tool, which is embedded into the search system of these catalogues. It ensures taking into account during the search the spelling versions of the search terms and the elements of the records. The tool is likely based on the fuzzy sets theory.

Using the fuzzy sets theory, we built a general mathematical model on the base of which an instrument of this kind could be constructed.

For our research we used the catalogues of the German National Library (Deutsche Nationalbibliothek), of the National Library of France (Bibliothèque nationale de France), of the search portal of the European Library and of the Library of Congress Authorities. We selected just one type of element variation – the differences in the spellings of the transliteration of the name of the Russian composer Piotr Chajkovskij in the Latin alphabet. This name was chosen because it has many various forms in Latin scripts.

Transliteration of Cyrillic Script in the Latin alphabet

The transliteration of Russian texts in Latin characters has a long history and various traditions both in Russia and in other countries, where different languages and different systems of writing are used [Reformatskij A.A.1972].

The selection of the version of transliteration depends on many factors: in the first place, apparently, on the phonetic systems of the target languages. It is necessary to note that the application of various forms of the transliteration of names has a diachronic aspect, i.e. it changes with time.

These changes are well outlined on the graphs constructed with the aid of the Google n – gram viewer (http://books.google.com/ngrams), which makes it possible to determine the frequency of the occurrence of the form of words in texts of several millions of books in different languages and to build the graphs of the changes of this value with time.

For graphing, we selected five most common versions of transliteration (1.Tchaikovsky, 2.Chaikovsky, 3.Cajkovskij, 4.Čajkovskij, 5.Tchaikowsky). At least four regularities are distinctly visible on the graphs (Fig.1-3).

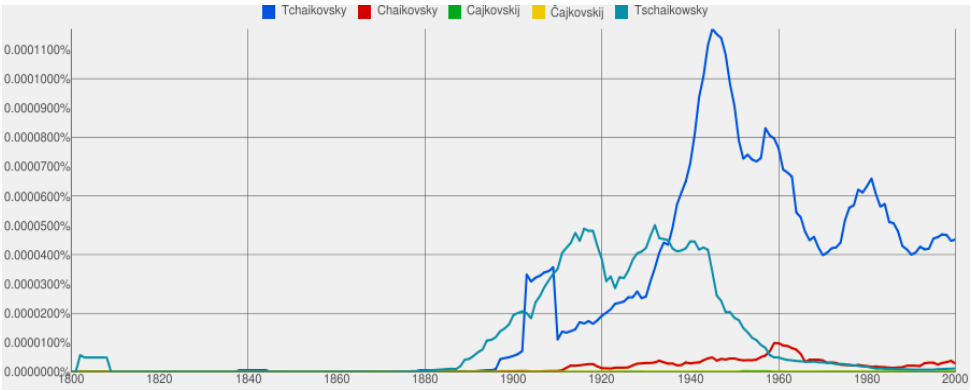

Fig.1 The frequency of the diverse variants of writing Chajkovskij in English 1800 -2000

First, in each of the three languages more than one version of the transliteration is present.

Secondly, different versions predominate in the different languages.

Thirdly, the frequency of the occurrence of each version changes in the course of time. The appearance of new versions is noted.

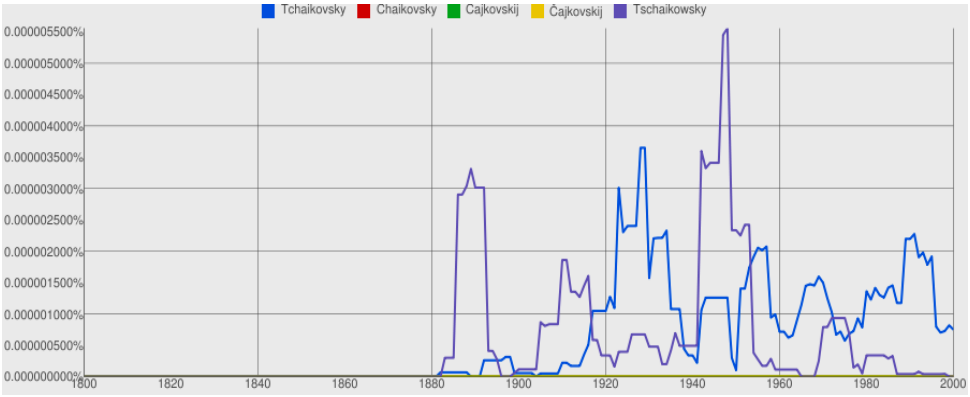

Fig.2 The frequency of the diverse variants of writing Chajkovskij in French 1800 -2000

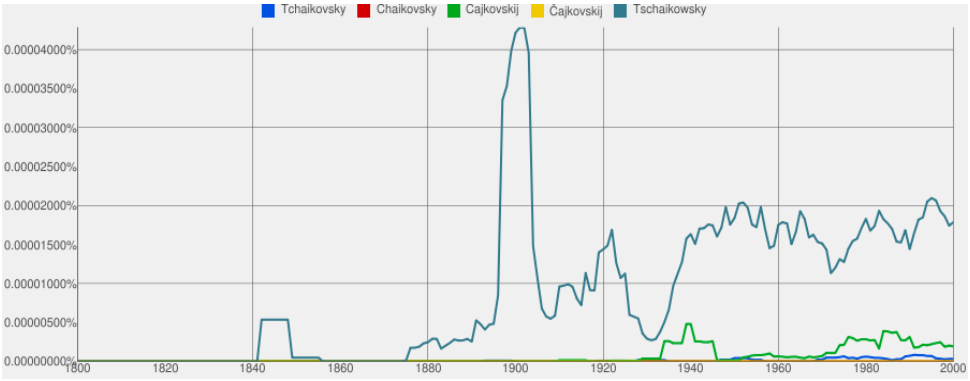

Fig.3 The frequency of the diverse variants of writing Chajkovskij in German 1800 -2000

Fourthly, in English and French in the different periods, versions 1 and 5 predominate, and in German, the steady predominance of version 5 is noted.

Currently in Russia the state standard GOST 7.79-2000 “Rules of transliteration of Cyrillic script in Latin alphabet” is accepted; it is the Russian version of the international standard ISO 9.95.

The standard proposes two versions of the transliteration – the strict one , where one character of the source alphabet is substituted by only one character of the target alphabet (system A, table 2 for non-Slavic languages) and the weakened one, where one character of the source language can be represented by more than one character of the target language [GOST 7.70-2000, 2001].

Thus, according to the standard, the surname “Chajkovskij” when transliterated in the Latin alphabet can appear in two versions:

Čajkovskij System A

Chajkovskij System B

Let us note that version 4, transliterated according to System A, which, therefore, corresponds to the international standard ISO 9.95, is encountered in none of the three natural languages (Fig. 1-3). We can easily explain this – the character Č is not used in these languages.

Records on the personal name "Chajkovskij" in the authority file of three national libraries

We compared authority records for the personal name Tschaikovsky P.I. from three sources: the authority files of the Library of Congress, of the national library of France and of the German national library.

The accepted transliteration form in the Library of Congress Authorities is Tchaikovsky, Peter Ilich, 1840-1893 (http://lccn.loc.gov/n79072979). This form, as we can see, does not conform to the standard ISO 9.95. The authority record presents 51 versions of the name spelling. However, it includes versions in Russian, Hebrew, Arabic and Chinese characters. The number of versions in Latin is thus 47.

The authority file of the German National Library (Gemeinsame Normdatei (GND)) accepted as heading the form Čajkovskij, Pëtr I. (http://d-nb.info/gnd/118638157), which fully conforms to the ISO standard; the record has 85 versions as references only in the Roman alphabet.

In the authority file of the national library of France (Autorités BnF) the chosen heading is Čajkovskij, Petr Il'ič (1840-1893) (http://catalogue.bnf.fr/ark:/12148/cb13900329p/PUBLIC). The transliteration conforms to ISO 9.95 with one exception: according to the standard the Russian character «ë» should be transliterated as «ë». The record contains 9 references.

Results of our experimental search in the catalogs of the German National Library and the National Library of France

From the authority files of three libraries, we selected ten versions from which 3 records are versions of the full name and 7 records only of the surname. Then we performed a search using each version as a search term in the catalogues of the libraries and the union catalogues which are accessible on Protocol Z39.50 via the portal of the European Library (TEL) (http://theeuropeanlibrary.org). The result is given in Table1.

Table 1. The result of the search through the European Library portal

The form of the name	Source (Authority file of the library)	Number of retrieved records	The codes of libraries where at least one record with this term was retrieved	Number of libraries where at least one record with this term was retrieved
Tchaikovsky	The National Library of France	16496	AL, AT, BE, CH, CY, CZ, **DE**, DK, EL, ES, FI, **FR**, IE, IS, IT, LU, LV, NL, NO, RS, SI, SK, TR, UK	24
Cajkovskij	The National Library of France	10799	AL, AT, BA, BE, CH, **DE**, DK, ES, FI, **FR**, HR, HU, IS, IT, LI, LU, LV, NL, NO, RS, SE, SI, UK	23
Tchaïkovsky	The National Library of France	20129	AT, BA, BE, CH, **DE**, DK, ES, FI, **FR**, HR, HU, IS, IT, LI, LU, LV, NL, NO, RS, SE, SI, TR, UK	23
Tschaikowsky	German National Library	12941	CH, CY, CZ, **DE**, DK, EL, ES, FI, **FR**, IE, IS, IT, LU, LV, NL, NO, RS, SI, TR, UK	20
Tchaikovsky, Peter Ilich	Library of Congress	10450	CH, CY, CZ, **DE**, DK, EL, ES, FI, **FR**, IE, IS, IT, LU, LV, NL, NO, RS, SI, UK	19
Čajkovskij	The National Library of France	11125	BE, CZ, DK ES, **FR**, HU, HR, IS, IT, LU, LV, NL, NO, RS, RU, SI, SK.TR, UK	17
Tchaikovsky, Pyotr Ilyich	Library of Congress	10262	BE, CZ, **DE**, ES, **FR**, HU, IS, IT, LU, LV, NL, NO, RS, RU, SI, SK.TR, UK	17
Chaikovsky	German National Library	7693	AT, CZ, **DE**, ES, FI, **FR**, IS, IT, LV, RS, RU, TR, UK	13
Tchaĭkovsky, Piotr Ilitch	Library of Congress	5	ES, IT	2

It is evident from Table 1 that the positive result of the search (at least one record with the term was retrieved) is obtained in the average in 16 libraries. In many libraries several forms are found. We selected two libraries.

German National Library (DNB) and National Library of France (BnF). In each of them a positive result was obtained with the search according to 8 versions. In the catalogues

of these two libraries, the search according to several versions of the name was carried out.

We presumed that if the difference between numbers of retrieved records was not significant, the catalogue had the program tool, based possibly on the fuzzy set logic, which takes in account a certain number of spelling versions. The search results are given in tables 2-3

Table 2. Results of our search according to the variants of the name in the catalog of the German national library

The spelling form of the name	Number of retrieved records
Tchaikovsky, Peter Ilich	6960
Tchaikovsky, Pyotr Ilyich	7122
Cajkovskij	7214
Tchaïkovsky	8503
Čajkovskij	7214
Tchaikovsky	8503
Tschaikowsky	10290
Chaikovsky	6991
Average number of record per a version	7849,63
Maximal (Tschaikowsky)	10290
Minimal (Tchaikovsky, Peter Ilich)	6960

It is evident from Table 2 that the numbers of records retrieved for each version are close. Differences in the number of obtained records with the search only on the surname can be caused by the presence in the catalog of the namesakes of the composer

It is evident from Table 3 that the results of our search in the catalogue of the National Library of France are different from the results in the German Library. Difference between the results of searches on various forms of the complete name between the numbers of records does not exceed 25. In five queries out of six the complete form of the name is used and, evidently, there is no factor of homonymy, i.e., the namesakes.

Table 3. Results of our search according to the variants of the name in the catalog of the national library of France

The spelling form of the name	Number of retrieved records
Tchaikovsky, Piotr Illitch	5360
Tchaïkowsky, Piotr Illitch	5365
Czajkowski, Piotr	5361
Tchaĭkovsky, Piotr Ilitch	5386
Čajkovskij, Petr Il'ič	5383
Tschaikowsky	424
Average number of record per a version	4546,5
Maximal (Tchaĭkovsky, Piotr Ilitch)	5386
Minimal (Tschaikowsky)	424

In the search using the version of the transliteration “Tschaikowsky”, a considerably smaller number of records was retrieved, which is apparently connected with the fact that this form is little used in French (see fig.3), whereas in German it has been most commonly used since 1920.

Data of our experimental search confirm, thus, the presence in the catalogues of these two libraries of the tool, which ensures the relative independence of the results of the search from the version of writing. Note that our research is a pilot one and our data have to be verified.

Some theoretical developments of the search in the library catalogue with the use of the fuzzy sets theory

In the process of designing, modifying and maintaining large libraries catalogues programmers, librarians and users deal with tasks in which one is supposed to operate with uncertain concepts and knowledge. The challenges of this kind are automation of documents indexing, multilingual search and the search by a term with more than one spelling.

To construct a relevant algorithm (of classification or record retrieval) it is necessary to formalize these concepts and knowledge. To carry out the description of uncertain concepts, and to operate many-valued incompletely specified lexical units and finally to build the models of the search, corresponding to the users’ information demand, one should apply the fuzzy sets theory.

The advantage of the fuzzy sets theory approach might be that within the framework of many-valued logic one can find solutions for a broader class of problems, than using clear logic.

We thus proposed [Vershinin M.I, 2000] to use a thesaurus with fuzzy relations between its elements (fuzzy thesaurus) for decreasing the expenditures for maintaining the catalogue and increasing the effectiveness in its use.

Further, we proposed [Vershinin M.I. et al, 2007] the algorithm of the automatic classification (indexing) of bibliographical records. The algorithm is based on the idea of the automatic fuzzy classification (indexing) of the records

The system attributes each bibliographic record to a definite class of terms marked with a subject heading. The assignment to the record of a subject heading index is determined by the comparison of the vocabulary of the record with the existing cluster of keywords, in this case the relation of similarity is uncertain. The classes do not have clear boundaries. That is why it cannot be determined unambiguously whether a record belongs to one class or another.

In the fuzzy classification, each document can be assigned to several classes with different degrees of membership. Several documents will be assigned to the same class if the degree of membership of the subject of each document to this class is maximal in comparison to the degrees of belonging to another class.

Thus, the possibility appears to perform a search by the uncertain attributes.

To deal with the problem of search and correction of errors of the system, a method of strings comparison based on the theory of fuzzy sets was developed [Vershinin M.I, 2001] The methodology of fuzzy logic permits to work under conditions when statistical data are lacking and to compare strings taking into account the possibility of errors without correcting the strings or without participation of an operator. The developed method takes in consideration both types of errors and their ranking according to the frequency of their occurrence and to other criteria.

Theoretical approach to building a search model based on the fuzzy sets theory

The formalization of uncertain concepts and relations is ensured by the introduction of linguistic and fuzzy variables, fuzzy set and fuzzy relation.

The fuzzy relations play in the theory of fuzzy sets and fuzzy logic an important role. Traditionally one applies fuzzy relations for modeling the structure of a complicated system, technological processes control and analysis of decision-making processes.

As far as maintaining the library catalogues is concerned, there is almost no practice of the use of fuzzy relations in that field. Below we shall try to show a perspective outlook of appliance of the method in order to ensure an effective search in the library catalogues.

The fuzzy relations theory is used as quality test for defining interrelations between the objects of the investigated system. Therewith the differences in the constraint force between objects are considered.

Commonly fuzzy n-ary relation is defined as a subset of Cartesian product of n sets [Pospelov D.A., ed., 1986]

$$R \subseteq X_1 \times X_2 \times \ldots \times X_n \tag{1}$$

and is specified with the membership function

$$\mu_R : \quad X_1 \times X_2 \times \ldots \times X_n \to L \tag{2}$$

As L one can take, for example, a set of real numbers, a segment of a real straight line, a set of linguistic variables, a set of m-dimensional vectors, pseudo Boolean algebra, completely distributive lattice etc.

This approach in the definition of L makes it possible to create various generalizations of the concept of relation, which one can apply to different fields. In addition, it allows using the well-developed set of devices of the relation theory, which results from the interpretation of a different function with the values from L.

As to the search in the library catalogues, the appliance of fuzzy relations gives the possibility to consider from a single point of view a variety of factors influencing the search quality, in particular to determine the links between query and catalogue records taking into account many factors.

To show the possibilities of solving several problems of search in library catalogues we confine ourselves to the consideration of binary fuzzy relations.

Generally one call fuzzy a binary relation between sets X and Y the function

$$R : \quad X \times Y \to L \tag{3}$$

Where L-is a completely distributive lattice, i.e. a partially ordered set, in which any non vacuous set has the greatest lower bound and least upper bound, and the operations of conjunction $\wedge$ and $\vee$ disjunction in L follow the distributive low. All the operations with fuzzy relations will be defined by these operations from L.

If we take a limited set of real numbers, then the operations of taking of greatest lower bound and the least upper bound will be correspondingly operations inf and sup, and the operations conjunction $\wedge$ and disjunction $\vee$ will be operations min and max. These operations will define also an operation with fuzzy relations.

In the case when L is a segment of a real straight line [0, 1], function R will be written as membership function

$$\mu_R : \quad X \times Y \to [0,1] \tag{4}$$

If the sets X and Y are bounded then the fuzzy relation between X and Y can be represented by its relation matrix, in which sets elements of X and Y correspond to lines and columns, and in the crossing of line x and column y the element R(x; y) is located.

In the case when sets X and Y coincide, the fuzzy relation R is called the fuzzy relation on the set X. To this relation one can assign a weighted graph in which each pair of knots (x;y) from X is connected with an arrow with weight R(x;y)

The model of search in the library catalogue based on the fuzzy similarity relation

Assuming that for the search in a catalogue the name of the Russian composer Chajkovskij Pyotr Il'ich is selected, the record for this name from the authority file of a certain library could be represented as a set of versions of its transliteration in Latin alphabet, which we designate by X.

Then the possible elements of set X are as follows:

x_1 – Tchaikovsky,

x_2 – Tchaïkovsky,

x_3 –Čajkovskij,

x_4 – Tschaikowsky,

x_5 – Chaikovsky etc.

Let us build a fuzzy similarity relation between versions taking in consideration diverse factors

(1) Account versions of transliteration (factor 1)

We present a fuzzy similarity relation as similarity matrix

$$M_1 = \{\mu_1(x_i; x_j)\}, i, j = 1,\ldots,m, \quad (5)$$

where $\mu_1(x_i; x_j)$ is the evaluation of similarity of the variants $\mu_1(x_i; x_j) \in [0,1]$.

The matrix of similarity can be obtained either by a quantitative evaluation of the certain parameter indicating the link between versions (number coincided characters, their sequence etc), or by questioning experts, who will indicate the degree of similarity for each pair of version from X in a certain scale of comparison which possibly consists of phrases like “very strong similarity”, “strong similarity”, “middle strong similarity”, “weak similarity”. It is obvious that in generally matrix M will be unsymmetrical.

(2) Account of occurrences of different versions of transliterations in diverse languages (factor 2)

To consider this factor we used the data of the diachronic search of the name Chajkovskij (fig1 – fig.3)

On the base of the data we built the matrix

$$A = \{a(x_i; t_j)\}, i = 1,\ldots, m, j = 1,\ldots, n \quad (6)$$

Where $a(x_i; t_j)$ – evaluation of the degree of occurrences of the name transliteration version x_i in the year t_j

The fuzzy similarity relation version R_2 can be represented as matrix of similarity

$$M_2 = A \bullet A^T \quad (7)$$

Where A^T is a transpose of matrix.

Note that the matrix M_2 has dimension $m \times m$ and is a symmetrical one.

Let us assume that the number of factors, which we have to take into account, is equal to k. After we defined matrixes of fuzzy relations M_1, M_2, …, M_k we build a matrix of fuzzy similarity relation M which takes into account all factors:

$$M = M_1 \wedge M_2 \wedge \ldots \wedge M_k \quad (8)$$

Now we use the fuzzy relation we built for the search in a library catalogue by the query z.

The algorithm looks as follows:

The similarity degree between query term z and versions of the set X is established. As a result we have a vector

$$\mu_z = \{\mu_z(x_1), \mu_z(x_2), \ldots, \mu_z(x_m)\}, \quad (9)$$

Where $\mu_z(x_i)$ – is evaluation of similarity degree of query z and record x_i (i=1,2,…, m).

To make evaluation of the similarity degree we can use the algorithm of fuzzy comparison of strings [3]. Note that the algorithm works even if the query formulation has errors.

(2) From the set X we select version x_{i0}, which

$$\mu_{я}(x_{i0}) = \max_{i} \mu(x_i), \quad i = 1,2,...,m \tag{10}$$

In the matrix M fuzzy relation we select the string with mark i_0 which correspond to the version x_{i0} This string provides a way to rank all the versions of set X. Essentially we got a corrected on the base of fuzzy relation vector μ_z. The records will be retrieved starting from the one, which contains version x_{i0} according to result of the ranking.

In the search technique one can also introduce a threshold α on the force of fuzzy similarity relation R, for example α = 0,5, and thus fix a selection of meaningful records.

The search starts from the record, which contains the term with the maximal degree of similarity. Note that the matrix of similarity M is generally asymmetrical and therefore when starting the search from different version, we may obtain different results.

Conclusion

It is evident from Tables 2-3 that the number of retrieved records varies depending on the form of name used in the query. The same result can be obtained by using the algorithm described above. We do not affirm that in the search systems of the catalogues of two libraries the fuzzy logic is realized. Possibly, in them probabilistic or any other model of the search is used

In our paper, we described a model of search where the fuzzy relation of similarity (matrices of similarity) is applied. For building the model, we use the result of our experimental research. The suggested model might be efficient in the design of search system.

Bibliography

[Vershinin M.I, 2000] Vershinin M.I. Sozdanie nechetkogo tezaurusa dlya e'lektronnogo kataloga in: Informacionny`e resursy` bibliotek i ix kadrovoe obespechenie: Mater. Mezhd. nauch. – prakt. konf., 23-26 maya 2000 g. / Belarus. un - t kul`tury`. - Minsk, 2000. - S. 91-96.

[Vershinin M.I. et al, 2007] Vershinin M.I., Vershinina L.P. Primenenie nechetkoj logiki v gumanitarny`x issledovaniyax in: Bibliosfera, 2007, №4. S.43-47.

[Vershinin M.I., 2001] Vershinin M.I. E'lektronny`j katalog: problemy` i resheniya. SPb.: Professiya, 2009.-232 s.

[GOST 7.70-2000, 2001] GOST 7.70-2000 Pravila transliteracii kirillovskogo pis`ma latinskim alfavitom. – M.: Gosstandart, 2001- 20 s.

[Pospelov D.A., ed., 1986] Nechetkie mnozhestva v modelyax upravleniya i iskusstvennogo intellekta / Pod red. D.A.Pospelova. M.: Nauka, 1986. 312 s.

[Reformatskij A.A., 1972] Reformatskij A.A. O standartizacii transliteracii latinskimi bukvami russkix tekstov in: Nauchno-texnicheskaya informaciya – 1972. -№ 10 S.32-36 [Zakharov V.P. et al, 1996]

[Zakharov V.P. et al] Zakharov V.P., Masevich A.C., Pimenov E.N. Authority control as a linguistic support element of an automated library system in: International Cataloguing and Bibliographic Control - 1996. - Vol 25, N4. - p.84-86.

Authors' Information

Liliya Vershinina *– Head of Department of Information Science and Mathematics, Saint Petersburg State University of Culture and Arts, e-mail: zk-inf@yandex.ru*

Major Fields of Scientific Research: General theoretical information research

Mikhail Vershinin*– Associate Professor. Department of Mechanics, National University of Mineral Mining "Gorny'j" ; e-mail:* stephen@smtp.ru

Major Fields of Scientific Research: Software technologies, General theoretical information research

Andrej Masevich *– Assistant professor. Saint-Petersburg State University of Culture and Arts; e-mail: andmasev@mail.ru*

Major Fields of Scientific Research: Computational Linguistics, Library information systems