# ВЫДЕЛЕНИЕ ЗНАНИЙ И ЯЗЫКОВЫХ ФОРМ ИХ ВЫРАЖЕНИЯ НА МНОЖЕСТВЕ ТЕМАТИЧЕСКИХ ТЕКСТОВ: ПОДХОД НА ОСНОВЕ МЕРЫ TF-IDF

*Д.В. Михайлов[1], А.П. Козлов[1], Г.М. Емельянов[1]*
*[1] Новгородский государственный университет имени Ярослава Мудрого, Новгород, Россия*

***Аннотация***

Статья посвящена проблеме выделения единиц знаний из множеств (корпусов) тематических текстов. Данная проблема актуальна для построения систем обработки, анализа, оценивания и понимания информации, в частности, изображений. Конечной практической целью здесь является поиск наиболее рационального варианта передачи смысла средствами заданного естественного языка (ЕЯ) для последующей фиксации фрагментов знаний в тезаурусе и онтологии предметной области (ПО). В настоящей статье разбиением слов исходной фразы на классы по значению меры TF-IDF относительно текстов корпуса решается задача поиска в корпусе фраз, максимально близких исходной по описываемому фрагменту фактического знания и формам его выражения в языке.

*Ключевые слова:* распознавание образов, интеллектуальный анализ данных, теория информации, тест открытой формы, языковое представление экспертных знаний.

*Цитирование:* **Михайлов, Д.В.** Выделение знаний и языковых форм их выражения на множестве тематических текстов: подход на основе меры TF-IDF / Д.В. Михайлов, А.П. Козлов, Г.М. Емельянов // Компьютерная оптика. – 2015. Т. 39, № 3. – С. 429-438.

## *Введение*

Разработка эффективных способов и средств описания решаемых задач – одно из ведущих направлений распознавания образов и интеллектуального анализа данных. Сказанное немаловажно в сфере обработки, анализа и понимания изображений [1]. Для многих приложений, в частности, учебных курсов с использованием открытых тестов, естественным источником знаний здесь будут публикации отечественных и зарубежных научных школ в виде монографий, обзорных статей, сборников трудов конференций и т.п. Наиболее актуальными при этом задачами являются тематическая рубрикация текстовых документов [2], а также представление предметных областей в виде тезаурусов и онтологий [3]. Основная проблема – поиск наиболее рационального варианта передачи смысла в единице знаний, определяемой семантически эквивалентными (СЭ) фразами предметно-ограниченного ЕЯ. Сам же смысл должен быть отражён в максимально компактном объёме текстовых данных. Решение данной проблемы выделением необходимого и достаточного набора минимальных семантико-синтаксических текстовых единиц и связей между ними на множестве СЭ-фраз обсуждалось авторами в [4]. При этом в круг задач эксперта, требующих автоматизации, входит:

- поиск СЭ-форм описания отдельного фрагмента знания (факта ПО) в заданном ЕЯ;
- сопоставление знаний эксперта с наиболее близкими фрагментами знаний других экспертов.

Следует отметить, что решение указанных задач не сводится к простому выделению из текста понятий и отношений между ними с подсчётом семантической близости пар и групп понятий [12]. Поиск и классификация языковых форм представления знаний здесь предполагает выявление в текстовом корпусе контекстов использования универсальной (общей) лексики, за счёт которой обеспечивается переход от исходной фразы к фразам, наиболее близким ей по смыслу (генерация синонимичных перифраз [5]). Близкую задачу, но принципиально обратного характера решает обучаемый детектор перифраз, предложенный в [6]: для исходной пары фраз определяется, есть ли одна синонимичная другой перифраза. Само детектирование осуществляется нейронной сетью, для обучения которой используются результаты синтаксического разбора пар фраз из обучающей выборки, формируемой экспертом. Последняя обязательно должна содержать примеры и контрпримеры перифраз, что не вполне соответствует требованию сопоставления различных фрагментов знаний: не учитываются смысловые связи фраз помимо синонимии. Кроме того, данный подход субъективен в плане представления о самой синонимии: не учитывается ПО каждой из фраз, а также степень их смысловой близости.

С учётом частоты встречаемости общей лексики в текстах разных ПО наиболее естественный путь решения вышеуказанных задач состоит в использовании известной статистической меры TF-IDF для выделения среди слов исходной фразы общей лексики и слов-терминов (в том числе в составе сочетаний). В настоящей работе рассматриваются возможности разбиения слов на классы по значению TF-IDF для поиска в текстовом корпусе описаний близких фрагментов знаний и языковых форм их выражения.

## *1. Мера TF-IDF и её интерпретация*

В задачах анализа текстов и информационного поиска TF-IDF есть статистическая мера, используемая для оценки важности слова в контексте документа, входящего в некоторый текстовый корпус.

Согласно определению, данная мера есть произведение TF-меры (отношения числа вхождений слова к общему числу слов документа) и инверсии частоты встречаемости слова в документах корпуса (IDF).

TF-мера оценивает важность слова $t_i$ в пределах отдельного документа $d$ и определяется как

$$tf(t_i,d)=\frac{n_i}{\sum_k n_k}, \quad (1)$$

где $n_i$ – число вхождений слова $t_i$ в документ, а в знаменателе – общее число слов в документе $d$.

Значение IDF (inverse document frequency – обратной частоты документа [7]) является мерой объёма полезной информации, которую может передать слово по корпусу в целом, и равно

$$idf(t_i,D)=\log\left(\frac{|D|}{|D_i|}\right), \quad (2)$$

где $|D_i \subset D|$ есть число документов корпуса $D$, в которых слово $t_i$ встретилось хотя бы один раз.

Следует отметить, что чем чаще слово встречается в документах корпуса, тем ближе к нулю будет значение меры (2). Это относится как к общей лексике (глаголы-связки, служебные части речи), так и к терминам, преобладающим в корпусе. В то же время, к примеру, слова общей лексики, обозначающие одни и те же ситуации либо действия с разных точек зрения и задающие так называемые конверсивные замены [5] («*приводить ⇔ являться следствием*»), будут иметь более высокие значения оценки (2).

Допустимо предполагать, что наиболее уникальные слова в документе (с наибольшими значениями произведения мер (1) и (2)) будут относиться к терминам ПО документа. Если же слово-термин имеет синонимы, встречающиеся в этом же документе, значение меры TF для него будет ниже. Как и в случае вышеупомянутых конверсивных замен, здесь мы имеем меньшую встречаемость в документах корпуса каждого слова из синонимического ряда и, как следствие, более высокие значения меры IDF по сравнению со случаем отсутствия синонимов у слова.

Возьмём приведённые выше рассуждения за основу требуемого нам разбиения слов исходной фразы на классы по значению произведения TF и IDF.

### *2. Классификация слов исходной фразы*

Пусть $X$ – упорядоченная по убыванию последовательность значений $tf(t,d)\cdot idf(t,D)$ для слов исходной фразы относительно документа $d$ в составе корпуса $D$. Выполним разбивку последовательности $X$ на кластеры с применением Алгоритма 1, содержательно близкого алгоритмам класса FOREL [8]. В качестве центра масс кластера $H_i$ здесь, как и в [4], берётся среднее арифметическое всех $x_j \in H_i$.

**Алгоритм 1.** Формирование кластера.
**Вход:** $X$; // упорядоченная по убыванию
// числовая последовательность
**Выход:** $H_i$, $X_p$, $X_s : X_p \bullet H_i \bullet X_s = X$;

**Начало**
1: $i := 1$;
2: $H_i := X$;
3: $X_p := \varnothing$;
4: $X_s := \varnothing$;
5: **если** $\text{good}(H_i)=true$ **или** $\text{diam}(H_i)=1$ **то**
вернуть $H_i$, $X_p$ и $X_s$;
6: **иначе если** $|\text{mc}(H_i)-\text{first}(H_i)| > |\text{mc}(H_i)-\text{last}(H_i)|$ **то**
7: $X_p := \{\text{first}(H_i)\}\bullet X_p$;
8: $H_i := \text{rest}(H_i)$;
9: перейти к шагу 5;
10: **иначе**
**если** $|\text{mc}(H_i)-\text{first}(H_i)| < |\text{mc}(H_i)-\text{last}(H_i)|$ **то**
11: $X_s := \{\text{last}(H_i)\}\bullet X_s$;
12: $H_i := \text{lrev}(H_i)$;
13: перейти к шагу 5;
14: **иначе**
15: $X_s := \{\text{last}(H_i)\}\bullet X_s$;
16: $X_p := \{\text{first}(H_i)\}\bullet X_p$;
17: $Tmp := \text{lrev}(H_i)$;
18: $H_i := \text{rest}(Tmp)$;
19: перейти к шагу 5;
**Конец** *{Алгоритм 1}.*

Здесь и далее «$\bullet$« – операция конкатенации.

*Табл. 1. Вспомогательные функции Алгоритма 1*

| Функция | Возвращаемое значение |
|---|---|
| First($X$) | первый элемент последовательности $X$ |
| last($X$) | последний элемент последовательности $X$ |
| lrev($X$) | исходная последовательность $X$ без последнего элемента |
| rest($X$) | исходная последовательность $X$ без первого элемента |
| good($X$) | **true** либо **false** в зависимости от выполнения условия (3) |
| mc($X$) | центр масс последовательности $X$ |
| Diam($H_i$) | диаметр кластера $H_i$ |

Будем считать, что элементы последовательности $X$ могут быть отнесены к одному кластеру, если

$$\begin{cases} |\text{mc}(X)-\text{first}(X)| < \dfrac{\text{mc}(X)}{4} \\ |\text{mc}(X)-\text{last}(X)| < \dfrac{\text{mc}(X)}{4} \end{cases}. \quad (3)$$

Выбор знаменателей правых частей неравенств в формуле (3) был основан на предположении, что элементы одного кластера всегда имеют больше сходств, чем различий.

Алгоритм 1 применяется к последовательностям $X_p$ и $X_s$ на его выходе, каждая из получившихся последовательностей, как и $X$, будет упорядоченной (доказательство очевидно). Данный процесс продолжается рекурсивно до тех пор, пока на очередном ша-

ге $X_p$ и $X_s$ не окажутся пустыми. В результате исходная последовательность $X$ разбивается на подпоследовательности (кластеры) $H_1,\ldots,H_r$, причём для $\forall i \neq j \;\; H_i \cap H_j = \varnothing$, а $H_1 \bullet H_2 \bullet \ldots \bullet H_r = X$. Обозначим далее $\{H_i | i = 1,\ldots,r\}$ как $F$.

Отметим, что классическая формулировка алгоритма FOREL предполагает минимизацию функционала качества, определяемого как сумма внутрикластерных расстояний

$$\rho_i = \sum_{x_j \in H_i} \left| x_j - \mathrm{mc}(H_i) \right| \qquad (4)$$

по всем получившимся кластерам $H_i$. Как уже отмечалось нами в [4], в практических приложениях алгоритмов этого семейства требуются априорные знания о ширине (диаметре) кластера в целях минимизации затрат по пересчёту значений функции (4). Применительно к кластеризации слов исходной фразы по значению меры TF-IDF относительно документов корпуса качество разбиения слов на классы подразумевает, с одной стороны, как можно большее число кластеров при максимально возможном числе слов в отдельном кластере, а с другой стороны – минимум разности значения наибольшего и наименьшего диаметров кластера. Поэтому численную оценку качества кластеризации слов исходной фразы мы возьмём из геометрических соображений и определим как

$$Q(F) = \frac{\sum_{i=1}^{r} \mathrm{diam}(H_i)}{\mathrm{len}(F)} \left( \mathrm{len}(F) - \max(F) \right) \frac{\min(F)}{\max(F)}, \qquad (5)$$

где $\min(F)$ и $\max(F)$ – минимальное и максимальное значения диаметра кластера из представленных в $F$; $\mathrm{len}(F)$ – длина списка $F$.

Содержательно оценка (5) позволяет выделить те документы текстового корпуса, относительно которых разделение слов исходной фразы на общую лексику и термины выражается в наибольшей степени.

### *3. Отбор фраз из документов корпуса*

Сортировкой документов корпуса $D$ по убыванию значений функции (5) с последующим разделением их на кластеры *Алгоритмом 1* отбираются документы с наибольшими значениями данной оценки (принадлежащими первому кластеру в составе формируемой последовательности $F(D)$). Далее обозначим множество указанных документов как $D'$.

Следующий шаг – отбор фраз из документов в составе $D'$ по максимуму слов, представленных в кластерах $H_1$, $H_{r/2}$, $H_r$ из формируемых *Алгоритмом 1*. При этом учитываются слова-термины из исходной фразы, наиболее уникальные для документа (кластер $H_1$), а также слова-термины, преобладающие в корпусе (кластер $H_r$). Введение в рассмотрение «серединного» кластера $H_{r/2}$ необходимо (в первую очередь) для выделения общей лексики, обеспечивающей синонимические перифразы, а также терминов-синонимов. Две указанные категории лексики, не имея значения TF-IDF из $H_1$ и $H_r$, тем не менее, могут быть представлены в $H_{r/2}$, поскольку имеют значения и TF, и IDF, близкие к средним по исходной фразе.

Как и для качества кластеризации, оценка представленности слов фразы $s$ документа $d \in D'$ в кластерах $\{H_1, H_{r/2}, H_r\} =: Cl$ здесь берётся из геометрических соображений и определяется как

$$\mathrm{N}(s, Cl) = \frac{\sqrt{\sum_{j \in \{1, r/2, r\}} \left| \{ t_i \in s : tfidf(t_i, d, D) \in H_j \} \right|^2}}{\sigma\left( \left| \{ t_i \in s : tfidf(t_i, d, D) \in H_j \} \right| \right) + 1}, \qquad (6)$$

где $\sigma\left( \left| \{ t_i : tfidf(t_i, d, D) \in H_j \} \right| \right)$ есть среднеквадратическое отклонение числа слов фразы документа $d$, представленных в кластере из списка $Cl$, $tfidf(t_i, d, D)$ – произведение $tf(t_i, d)$ и $idf(t_i, D)$. Добавление единицы в знаменателе формулы (6) имеет целью предотвратить деление на ноль в случае нулевого среднеквадратического отклонения.

Отбираемые фразы кластеризуются по значению функции (6) с помощью *Алгоритма 1*, в качестве результата возвращается набор фраз, которому отвечает кластер наибольших значений оценки (6).

Итак, предлагаемый метод поиска фраз, близких исходной, может быть формально описан следующей совокупностью шагов.

1: $X^Q := \varnothing$;
2: **для всех** $d \in D$
3: вычислить $Q(F)$ согласно формуле (5);
4: $X^Q := X^Q \cup \{Q(F)\}$;
5: отсортировать $X^Q$ и $D$
по убыванию значения $Q(F)$
6: сформировать
$H_1^Q \bullet H_2^Q \bullet \ldots \bullet H_{r(D)}^Q$ **и** $H_1^D \bullet H_2^D \bullet \ldots \bullet H_{r(D)}^D$
с применением *Алгоритма 1*;
7: $D' := H_1^D$;
8: $X^{\mathrm{N}} := \varnothing$;
9: $S := \{ s : s \in d, d \in D' \}$;
10: **для всех** $s \in S$
11: вычислить $\mathrm{N}(s, Cl)$ согласно формуле (6);
12: $X^{\mathrm{N}} := X^{\mathrm{N}} \cup \{\mathrm{N}(s, Cl)\}$;
13: отсортировать $X^{\mathrm{N}}$ и $S$
по убыванию значения $\mathrm{N}(s, Cl)$;
14: сформировать
$H_1^{\mathrm{N}} \bullet H_2^{\mathrm{N}} \bullet \ldots \bullet H_{r(S)}^{\mathrm{N}}$ **и** $H_1^S \bullet H_2^S \bullet \ldots \bullet H_{r(S)}^S$
с применением *Алгоритма 1*;
15: вернуть $H_1^S$;

**Замечание.** Кластеры, формируемые на последовательностях $X^Q$ и $X^{\mathrm{N}}$, а также на множествах $D$ и

$S$, здесь обозначены как $H_1^Q \dots H_{r(D)}^Q$ и $H_1^D \dots H_{r(D)}^D$ и соответственно $H_1^{\mathrm{N}} \dots H_{r(S)}^{\mathrm{N}}$ и $H_1^S \dots H_{r(S)}^S$.

Как видно из представленного псевдокода, метод предполагает $(|D|+2)$-кратное применение *Алгоритма 1*, имеющего сложность $O(n^2)$. Каждому вызову этого алгоритма предшествует сортировка разбиваемой последовательности (в данной работе она выполнялась методом вставок), требующая $O(n^2)$ шагов ($n$ – длина разбиваемой последовательности). Для снижения вычислительной сложности решения рассматриваемой задачи целесообразно запоминать результаты кластеризации слов различных исходных фраз (если поиск ведётся по одному и тому же корпусу).

Заметим, что предложенный метод не учитывает синтаксический контекст, привязка слова к нему исключила бы поиск фраз, синонимия которых исходной фразе затрагивает и синтаксис, и лексику (пример – упомянутые ранее *конверсивные замены*).

### *4. Экспериментальные исследования*

Для апробации предложенного метода был сформирован текстовый корпус, включающий:

- 3 статьи в журнале «Таврический вестник информатики и математики (ТВИМ)»;
- 2 статьи в сборниках трудов конференций «Интеллектуализация обработки информации» 2010 и 2012 гг.;
- 1 статью в сборнике трудов 15-й Всероссийской конференции «Математические методы распознавания образов (ММРО-15)» (2011 г.);
- материалы тезисов двух докладов на 13-й Всероссийской конференции «Математические методы распознавания образов» (2007 г.);
- материалы тезисов четырнадцати докладов на 16-й Всероссийской конференции «Математические методы распознавания образов» (2013 г.);
- материалы тезисов двух докладов на конференции «Интеллектуализация обработки информации» 2014 г.;
- материалы одного научного отчёта, подготовленного Д.В. Михайловым в 2003 г.

Тематика отбираемых работ включала:

- математические методы обучения по прецедентам (К.В. Воронцов, М.Ю. Хачай, Е.В. Дюкова, Н.Г. Загоруйко, Ю.Ю. Дюличева, И.Е. Генрихов, А.А. Ивахненко);
- модели и методы распознавания и прогнозирования (В.В. Моттль, О.С. Середин, А.И. Татарчук, П.А. Турков, М.А. Суворов, А.И. Майсурадзе);
- интеллектуальный анализ экспериментальных данных (С.Д. Двоенко, Н.И. Боровых);
- обработку, анализ, классификацию и распознавание изображений (А.Л. Жизняков, К.В. Жукова, И.А. Рейер, Д.М. Мурашов, Н.Г. Федотов, В.Ю. Мартьянов, М.В. Харинов).

Число слов в документах корпуса варьировалось от 218 до 6298.

В экспериментах по формированию единиц экспертных знаний участвовали девять исходных фраз, которые описывали факты предметной области «Математические методы обучения по прецедентам» (табл. 2).

*Табл. 2. Исходные фразы*

| № | Исходная фраза |
|---|---|
| 1 | *Переобучение приводит к заниженности эмпирического риска.* |
| 2 | *Переподгонка приводит к заниженности эмпирического риска.* |
| 3 | *Переподгонка служит причиной заниженности эмпирического риска.* |
| 4 | *Заниженность эмпирического риска является результатом нежелательной переподгонки.* |
| 5 | *Переусложнение модели приводит к заниженности средней ошибки на тренировочной выборке.* |
| 6 | *Переподгонка приводит к увеличению частоты ошибок дерева принятия решений на контрольной выборке.* |
| 7 | *Переподгонка приводит к заниженности оценки частоты ошибок алгоритма на контрольной выборке.* |
| 8 | *Заниженность оценки ошибки распознавания связана с выбором правила принятия решений.* |
| 9 | *Рост числа базовых классификаторов ведёт к практически неограниченному увеличению обобщающей способности композиции алгоритмов.* |

Программная реализация метода на языке Java и результаты экспериментов представлены на портале НовГУ по адресу: http://www.novsu.ru/file/1146133.

В качестве примера можно привести поиск в текстах корпуса фраз, максимально близких фразе №9 из табл. 2 по описываемому фрагменту знания и формам его выражения в русском языке.

Из документов корпуса по критерию (5) лучшими оказались две статьи К.В. Воронцова: в журнале «Таврический вестник информатики и математики» (№1, 2004 г.) и в сборнике трудов 15-й Всероссийской конференции «Математические методы распознавания образов». Эти два документа и послужили источником отбора фраз по максимуму оценки (6).

*Табл. 3. Кластеры для отбора фраз (фраза №9, табл. 2)*

| **Воронцов К.В., ТВИМ 2004 №1, слова, представленные в кластерах** | |
|---|---|
| $H_1$ | *алгоритм, обобщать, способность* |
| $H_{r/2}$ | *классификатор, увеличение, число* |
| $H_r$ | *вести* |
| **Воронцов К.В., ММРО-15, слова, представленные в кластерах** | |
| $H_1$ | *алгоритм* |
| $H_{r/2}$ | *рост, композиция* |
| $H_r$ | *неограниченный, базовый, увеличение* |

Пример фразы, отобранной по максимуму оценки (6) для кластеров из табл. 3 и выражающей связь за-

данного фрагмента знаний из табл. 2 со знаниями других экспертов: «*Обобщающая способность определяется как вероятность ошибки найденного алгоритма, либо как частота его ошибок на неизвестной контрольной выборке, также случайной, независимой и одинаково распределённой*». Из представленных в табл. 3 слов фраза содержит слова *обобщать*, *способность* и *алгоритм*, а определение *обобщающей способности алгоритма*, упоминаемой в исходной фразе №9 из табл. 2, связывается здесь с понятиями *вероятность ошибки* и *частота ошибок на контрольной выборке*. Сказанное немаловажно для правильного подбора фраз, семантически эквивалентных фразам №6 и 7 в той же таблице.

Другая фраза, также отобранная по максимуму критерия (6) и содержащая те же слова *алгоритм*, *обобщать* и *способность*, является примером уже языковых выразительных средств конструирования экспертом синонимичных перифраз: «*Результатом обучения является не только сам алгоритм, но и достаточно точная оценка его обобщающей способности*», ср. «*ведёт к ⇔ является результатом*».

*Табл. 4. Кластеры для отбора фраз (фраза №4, табл. 2)*

| **Воронцов К.В., ТВИМ 2004 №1, слова, представленные в кластерах** | |
|---|---|
| $H_1$ | *риск, эмпирический* |
| $H_{r/2}$ | *заниженность, являться, переподгонка* |
| $H_r$ | *нежелательный* |
| **диапазоны значений TF-IDF** | |
| $H_1$ | 0,0020…0,0026 |
| $H_{r/2}$ | $1{,}4386\cdot10^{-4}$…$2{,}1839\cdot10^{-4}$ |
| $H_r$ | 0,0000…0,0000 |
| **Воронцов К.В., ММРО-15, слова, представленные в кластерах** | |
| $H_1$ | *риск* |
| $H_{r/2}$ | *результат* |
| $H_r$ | *нежелательный, заниженность, переподгонка* |
| **диапазоны значений TF-IDF** | |
| $H_1$ | 0,0021…0,0021 |
| $H_{r/2}$ | $4{,}3890\cdot10^{-4}$…$4{,}3890\cdot10^{-4}$ |
| $H_r$ | 0,0000…0,0000 |
| **Дюличева Ю.Ю., ТВИМ 2002 №1, слова, представленные в кластерах** | |
| $H_1$ | *переподгонка* |
| $H_{r/2}$ | *являться* |
| $H_r$ | *нежелательный, заниженность, риск* |
| **диапазоны значений TF-IDF** | |
| $H_1$ | 0,0040…0,0040 |
| $H_{r/2}$ | $1{,}7015\cdot10^{-4}$…$1{,}7015\cdot10^{-4}$ |
| $H_r$ | 0,0000…0,0000 |

Следующий пример для фразы №4 из табл. 2 иллюстрирует поиск синонима для слова-термина.

*Табл. 5. Значения TF (по первому документу в табл. 4) и IDF слов фразы №4 из табл. 2*

| **слово** | **TF** | **IDF** |
|---|---|---|
| нежелательный | 0,0000 | 1,3979 |
| заниженность | $1{,}5623\cdot10^{-4}$ | 1,3979 |
| переподгонка | $1{,}5623\cdot10^{-4}$ | 0,9208 |
| являться | 0,0031 | 0,0555 |
| результат | 0,0022 | 0,1938 |
| эмпирический | 0,0033 | 0,6198 |
| риск | 0,0028 | 0,9208 |

*Табл. 6. Кластеры для отбора фраз (фраза №8, табл. 2)*

| **Воронцов К.В., ТВИМ 2004 №1, слова, представленные в кластерах** | |
|---|---|
| $H_1$ | *оценка, ошибка* |
| $H_{r/2}$ | *заниженность* |
| $H_r$ | *с, принятие* |
| **диапазоны значений TF-IDF** | |
| $H_1$ | 0,0019…0,0029 |
| $H_{r/2}$ | $2{,}1839\cdot10^{-4}$…$2{,}1839\cdot10^{-4}$ |
| $H_r$ | 0,0000…0,0000 |
| **Дюличева Ю.Ю., ТВИМ 2002 №1, слова, представленные в кластерах** | |
| $H_1$ | *ошибка* |
| $H_{r/2}$ | *решение, распознавание, принятие* |
| $H_r$ | *заниженность, с, связанный* |
| **диапазоны значений TF-IDF** | |
| $H_1$ | 0,0068…0,0068 |
| $H_{r/2}$ | $3{,}0603\cdot10^{-4}$…$3{,}7303\cdot10^{-4}$ |
| $H_r$ | 0,0000…0,0000 |
| **Дюличева Ю.Ю., ТВИМ 2003 №2, слова, представленные в кластерах** | |
| $H_1$ | *решение, распознавание, принятие* |
| $H_{r/2}$ | *правило* |
| $H_r$ | *заниженность, с* |
| **диапазоны значений TF-IDF** | |
| $H_1$ | 0,0017…0,0018 |
| $H_{r/2}$ | $4{,}2541\cdot10^{-4}$…$4{,}2541\cdot10^{-4}$ |
| $H_r$ | 0,0000…0,0000 |

Заметим, что «*переподгонка*» имеет синоним «*переобучение*» в текстах корпуса, и относительно первого из документов в табл. 4 значение TF-IDF для него вошло в кластер $H_{r/2}$, для сравнения см. табл. 5. Указанный синоним присутствует во фразе, отобранной по максимуму критерия (6) для кластеров из табл. 4: «*Причиной является всё то же переобучение, которое приводит к заниженности эмпирического риска*» (из представленных в табл. 4 фраза содержит слова *эмпирический*, *риск*, *являться*, *заниженность*). Эта же фраза содержит вариант конверсивной замены для исходной фразы, ср. «*причина ⇔ результат*».

Следует отметить, что если слово-термин имеет минимальную встречаемость в текстах корпуса, то для большинства документов кластера наибольших значений оценки (5) (множество $D'$) слово будет отнесено к кластеру $H_r$. Как следствие – достаточно невысокая совместная встречаемость в одной фразе с другими словами, представленными в кластерах из списка *Cl*. При этом фразы, близкие исходной с точки зрения эксперта по описываемому фрагменту знания либо формам его выражения, найдены не будут.

Примером может послужить слово *заниженность* в эксперименте по поиску фраз, максимально близких фразе №8 из табл. 2.

Из документов корпуса по критерию (5) лучшими оказались статьи в Таврическом вестнике информатики и математики: К.В. Воронцова: (№1, 2004 г.) и Ю.Ю. Дюличевой (№1, 2002 г. и №2, 2003 г.).

*Табл. 7. Значения TF (по первому документу в табл. 6) и IDF слов фразы №8 из табл. 2*

| **слово** | **TF** | **IDF** |
|---|---|---|
| заниженность | $1{,}5623\cdot10^{-4}$ | 1,3979 |
| с | 0,0102 | 0,0000 |
| распознавание | 0,0022 | 0,1427 |
| оценка | 0,0100 | 0,2840 |
| ошибка | 0,0042 | 0,4437 |
| выбор | $9{,}3735\cdot10^{-4}$ | 0,4437 |
| принятие | 0,0000 | 0,6990 |
| связанный | $1{,}5623\cdot10^{-4}$ | 0,6990 |
| решение | 0,0013 | 0,2840 |
| правило | 0,0013 | 0,4437 |

При этом по максимуму критерия (6) была выбрана фраза: *«Сравнивая прогнозируемый коэффициент ошибки $t$ с ошибками ветви $T(t)$ и наибольшей из ветвей с корнем в дочерней вершине вершины $t$, принимается решение о том, оставлять без изменений $T(t)$, редуцировать или наращивать в вершине $t$»* [Дюличева Ю.Ю., ТВИМ 2002 №1].

Следует отметить, что в процессе отбора в данном эксперименте не нашлось фраз, где помимо максимизации критерия (6) выполнялось бы требование наличия слова *заниженность*. Для сравнения в табл. 7 приведены значения TF и IDF слов исходной фразы относительно первого из документов в табл. 6. Этот документ единственный, где слово *заниженность* имеет и TF, и IDF ненулевыми.

Одним из вариантов качественного улучшения поиска для рассматриваемого случая могло бы стать использование суммарного значения TF-IDF слов исходной фразы, встречающихся во фразе *s* документа $d \in D'$, в качестве альтернативы оценке (6).

Но, как показывает эксперимент, с той же фразой №8 из табл. 2 это приводит лишь к росту числа отбираемых фраз, а встречающиеся там слова исходной фразы (см. табл. 8) лишь в 2 % случаев имеют значения TF-IDF, представленные в кластере $H_{r/2}$ («серединном»). В 83 % случаев это слова со значениями TF-IDF, большими представленных в указанном кластере. Содержательно это означает отсутствие перифраз исходной фразы в совокупности с отсутствием фраз, связывающих упоминаемые в исходной фразе понятия с другими понятиями предметной области. Для сравнения в табл. 9 для исходных фраз из табл. 2 приведено общее число отобранных фраз из документов корпуса (*N*), в том числе представляющих языковые выразительные средства ($N_1$), синонимы ($N_2$) и связи понятий предметной области ($N_3$).

*Табл. 8. Эксперимент с отбором фраз по сумме TF-IDF встречающихся слов фразы №8 из табл. 2*

| **Дюличева Ю.Ю., ТВИМ 2002 №1, число отобранных фраз, содержащих слово, *N*** | | |
|---|---|---|
| ***N*** | **слово исходной фразы** | **TF-IDF** |
| 30 | *ошибка* | 0,0068 |
| 9 | *оценка* | 0,0016 |
| 1 | *выбор* | $1{,}9426\cdot10^{-4}$ |
| 1 | *правило* | $7{,}7705\cdot10^{-4}$ |
| 1 | *решение* | $3{,}7303\cdot10^{-4}$ |
| 6 | *с* | 0,0000 |

*Табл. 9. Отбор фраз: сравнение двух критериев*

| **№** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|
| *по максимуму критерия (6)* | | | | | | | | | |
| *N* | 1 | 1 | 1 | 1 | 3 | 1 | 11 | 1 | 31 |
| $N_1$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| $N_2$ | 0 | 1 | 1 | 1 | 3 | 0 | 1 | 0 | 2 |
| $N_3$ | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 8 |
| *по сумме TF-IDF встречающихся слов исходной фразы* | | | | | | | | | |
| *N* | 2 | 1 | 11 | 1 | 5 | 2 | 1 | 30 | 9 |
| $N_1$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| $N_2$ | 0 | 1 | 1 | 1 | 3 | 0 | 0 | 0 | 1 |
| $N_3$ | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 7 |

Ещё один альтернативный вариант решения рассматриваемой в работе задачи основан на использовании знаний об известных семантических отношениях и формах их выражения в текстах. Здесь следует отметить систему «Серелекс» [13], а также тезаурус WordNet [14]. Но как показали приведённые в табл. 10 и 11 результаты экспериментов с лексикой исходных фраз из табл. 2, таких знаний, как правило, недостаточно. В данном примере слова представлены начальными формами (леммами) и их английскими эквивалентами. Последние использовались и в эксперименте с WordNet (доступной на момент подготовки статьи версией, Принстонский университет, США).

Ключевой проблемой указанных решений является пополнение задействуемых баз знаний. Так, системой “Серелекс” здесь использовалась коллекция документов, включающая заголовки статей Википедии ($2{,}026\cdot10^9$ словоформ, 3368147 лемм) и текстовый корпус ukWaC [15] ($0{,}889\cdot10^9$ словоформ, 5469313 лемм). При этом предметная классификация лексики не предусматривалась в принципе, что затрудняет использование лексико-синтаксических шаблонов, задействованных системой, для выделения требуемых фрагментов в текстах заданного множества. Сказанное справедливо и для синонимиче-

ских рядов (групп синонимов, синсетов [14]) как базовых единиц тезауруса WordNet: внутри каждой группы степень синонимии слов в действительности может варьироваться в зависимости от их предметной ориентации. Так, в синонимическом ряду для слова *model* варианты *model* и *theoretical account* не являются синонимами относительно предметной области “Математические методы обучения по прецедентам”.

*Табл. 10. Начальные формы слов, английские эквиваленты и связи, найденные системой «Серелекс»*

| **Лемма** | **Найдено отношений** | **в том числе** | | **Английский эквивалент** | **Найдено отношений** | **в том числе** | |
|---|---|---|---|---|---|---|---|
| | | **связи понятий ПО для фраз из табл. 2** | **синонимы** | | | **связи понятий ПО для фраз из табл. 2** | **синонимы** |
| *нежелательный* | 0 | 0 | 0 | *undesirable* | 107 | 1 | 0 |
| *переподгонка* | 0 | 0 | 0 | *overfitting* | 2 | 0 | 0 |
| *переобучение* | 0 | 0 | 0 | *overtraining* | 9 | 1 | 0 |
| *заниженность* | 0 | 0 | 0 | *underestimate* | 8 | 2 | 0 |
| *эмпирический* | 0 | 0 | 0 | *empiric* | 4 | 0 | 0 |
| *риск* | 25 | 1 | 0 | *risk* | 2197 | 4 | 0 |
| *переусложнение* | 0 | 0 | 0 | *(excessively) complication* | 837 | 2 | 0 |
| *модель* | 263 | 0 | 0 | *model* | 5193 | 10 | 0 |
| *средний* | 45 | 0 | 0 | *mean (value)* | 3524 | 7 | 1 |
| *ошибка* | 43 | 2 | 0 | *error* | 1155 | 9 | 1 |
| *тренировочный* | 0 | 0 | 0 | *training* | 3994 | 4 | 1 |
| *контрольный* | 0 | 0 | 0 | *control* | 4749 | 9 | 1 |
| *выборка* | 4 | 0 | 0 | *sample* | 1133 | 2 | 0 |
| *увеличение* | 53 | 1 | 0 | *growth* | 1868 | 2 | 1 |
| *частота ошибок* | 0 | 0 | 0 | *error rate* | 43 | 2 | 0 |
| *дерево принятия решений* | 0 | 0 | 0 | *decision tree* | 47 | 9 | 0 |
| *оценка (частоты)* | 51 | 1 | 0 | *(rate) estimation* | 418 | 10 | 1 |
| *алгоритм* | 51 | 0 | 0 | *algorithm* | 983 | 9 | 0 |
| *распознавание* | 10 | 0 | 0 | *recognition* | 1233 | 11 | 0 |
| *рост (числа)* | 100 | 1 | 0 | *increasing (the number)* | 902 | 2 | 1 |
| *базовый* | 0 | 0 | 0 | *base* | 2293 | 8 | 1 |
| *классификатор* | 8 | 0 | 0 | *classifier* | 62 | 8 | 0 |
| *практически* | 0 | 0 | 0 | *practically* | 0 | 0 | 0 |
| *неограниченный* | 0 | 0 | 0 | *unlimited* | 0 | 0 | 0 |
| *обобщать* | 0 | 0 | 0 | *(to) generalize* | 0 | 0 | 0 |
| *способность* | 124 | 1 | 1 | *capability* | 1533 | 10 | 0 |
| *композиция* | 90 | 0 | 0 | *composition* | 1782 | 9 | 0 |
| *приводить (к)* | 0 | 0 | 0 | *(to) result (in)* | 2557 | 0 | 3 |
| *вести (к)* | 0 | 0 | 0 | *(to) lead (to)* | 1791 | 0 | 2 |
| *служить* | 13 | 0 | 0 | *(to) be* | 0 | 0 | 0 |
| *являться* | 0 | 0 | 0 | *(to) be* | 0 | 0 | 0 |
| *причина* | 145 | 0 | 1 | *reason* | 2728 | 0 | 3 |
| *результат* | 52 | 0 | 0 | *result* | 2557 | 0 | 3 |
| *связанный (с)* | 0 | 0 | 0 | *relate(d) (to, with)* | 0 | 0 | 0 |

*Табл. 11. Синонимические ряды (группы синонимов) по WordNet для английских эквивалентов слов из табл. 10*

| **Слово (английский эквивалент)** | **Найдено групп синонимов** | **Число синонимов по предметной области фраз из табл. 2** | **Слово (английский эквивалент)** | **Найдено групп синонимов** | **Число синонимов по предметной области фраз из табл. 2** |
|---|---|---|---|---|---|
| *undesirable* | 3 | 1 | *recognition* | 8 | 2 |
| *underestim\|\|ate;~ation* | 4 | 4 | *increasing* | 4 | 0 |
| *empiric* | 2 | 2 | *number* | 18 | 1 |
| *risk* | 6 | 2 | *base* | 30 | 1 |
| *complication* | 5 | 4 | *classifier* | 2 | 0 |
| *excessively* | 1 | 4 | *practically* | 3 | 1 |
| *model* | 14 | 1 | *unlimited* | 3 | 1 |
| *mean* | 16 | 2 | *generalize* | 4 | 3 |
| *training* | 13 | 0 | *capability* | 3 | 1 |
| *control* | 20 | 0 | *composition* | 9 | 1 |
| *sample* | 4 | 1 | *result* | 3 | 2 |
| *growth* | 7 | 3 | *lead* | 31 | 1 |
| *estimation* | 4 | 3 | *be* | 14 | 1 |
| *rate* | 7 | 1 | *reason* | 9 | 1 |
| *algorithm* | 1 | 2 | *relate* | 5 | 1 |

Таким образом, наряду с решением своей основной задачи, предложенный в настоящей работе метод позволяет автоматизировать формирование исходных данных для построения базы знаний смысловых отношений на основе форм их выражения в текстах.

### *5. Некоторые технические детали и допущения*

Классическая постановка задачи кластерного анализа [8] предполагает, что каждый элемент последовательности, разбиваемой на кластеры с применением Алгоритма 1, представлен в ней один раз. В целях наглядности изложение предлагаемого метода неявно содержит предположение о выполнении данного условия, в частности, что каждое слово исходной фразы имеет уникальное значение меры TF-IDF.

В программной реализации метода для подсчёта статистики слова приводились к начальной форме с помощью функции *getNormalForms* в составе библиотеки русской морфологии [10]. Для многозначных слов отбирался вариант с наименьшим значением меры IDF (наиболее распространённый в корпусе).

Извлечение текста из PDF-файла выполнялось с помощью метода *getText* класса *PDFTextStripper* в составе библиотеки *Apache PDFBox* [11]. По причине отсутствия распознавания формул в числе решаемых задач метода *getText* все формулы из анализируемых документов переводились экспертом вручную в формат, близкий используемому в LaTeX.

Отдельная задача – выделение границ предложений в тексте по знакам препинания. Поскольку в анализируемых текстах употребление инициалов и иных сокращений, которые привели бы к ложному выделению границ фраз, минимально, для решения данной задачи авторы ограничились регулярным выражением [\\.!\\?]. Здесь представляется перспективной реализация обучаемых моделей, в частности на основе метода максимальной энтропии [12].

### *Заключение*

Основной *результат* настоящей работы – *метод поиска в текстовом корпусе описаний близких фрагментов знаний и языковых форм их выражения*.

Помимо подготовки открытых тестов, важнейшая сфера приложения данного метода – построение специализированных тезаурусов, идейно близких «Чёрному квадрату» [3], развиваемому исследовательским коллективом ВЦ РАН. По сравнению с известными подходами, предложенный метод позволяет решить задачу выделения классов понятий предметной области и отношений между ними на основе меньших обучающих выборок и без ориентации на определённые типы связей слов исходных фраз.

Отметим, что результаты работы предложенного метода существенно зависят от подбора исходного корпуса экспертом. Наиболее значимыми критериями отбора документов в корпус могут служить [9]:

- качество выделения тем – совокупностей специальных терминов предметной области, совместно встречающихся в документах;
- характер распределения терминов в теме;
- характер распределения тем в документе.

Выработка численной оценки, которая учитывала бы все три указанных критерия качества, заслуживает отдельного исследования.

Отдельного исследования здесь также заслуживает одновременная встречаемость всех слов, представленных в кластере из затрагиваемых оценкой (6). В этом случае мера предсказуемости появления требуемых слов во фразе будет аналогична известным в $L$ -граммном (по К. Шеннону, [16]) анализе.

### *Благодарности*

Работа поддержана РФФИ (проект №13-01-00055) и Минобрнауки РФ (базовая часть госзадания).

### *Литература*

1. **Сойфер, В.А.** Анализ и распознавание наномасштабных изображений: традиционные подходы и новые постановки задач / В.А. Сойфер, А.В. Куприянов // Компьютерная оптика. – 2011. – Т. 35, № 2. – С. 136-144. – ISSN 0134-2452.
2. **Царьков, С.В.** Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов / С.В. Царьков // Естественные и технические науки. – 2012. – № 6. – С. 456-464. – ISSN 1684-2626.
3. **Gurevich, I.** The challenges, the problems and the tasks of the descriptive approach to image analysis / I. Gurevich, Yu. Trusova, V. Yashina // 11th International Conference «Pattern Recognition and Image Analysis: New Information Technologies» (PRIA-11-2013). – 2013. – Vol. 1. – P. 30-35.
4. **Емельянов, Г.М.** Формирование единиц представления предметных знаний в задаче их оценки на основе открытых тестов / Г.М. Емельянов, Д.В. Михайлов, А.П. Козлов // Машинное обучение и анализ данных. – 2014. – Т. 1, № 8. – С. 1089-1106. – ISSN 2223-3792.
5. **Мельчук, И.А.** Опыт теории лингвистических моделей «Смысл⇔Текст»: Семантика, синтаксис / И.А. Мельчук. – М.: Школа «Языки русской культуры», 1999. – 345 с.
6. **Huang, E.** Paraphrase Detection Using Recursive Autoencoder / E. Huang [Электронный ресурс]. – 2011. – URL: http://nlp.stanford.edu/courses/cs224n/2011/reports/ehhuang.pdf (дата обращения 22.05.2015).
7. **Jones, K.S.** A statistical interpretation of term specificity and its application in retrieval / K.S. Jones // Journal of Documentation. – 2004. – Vol. 60(5). – P. 493-502.
8. **Загоруйко, Н.Г.** Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Издательство института математики, 1999. – 270 с.
9. **Воронцов, К.В.** Многокритериальные и многомодальные вероятностные тематические модели коллекций текстовых документов / К.В. Воронцов, А.А. Потапенко, А.И. Фрей, М.А. Апишев, Н.В. Дойков, А.В. Шапулин, Н.А. Чиркова // 10-я Междунар. конф. ИОИ-2014: Тезисы докладов. – 2014. – С. 198.
10. russianmorphology: Russian Morphology for lucene [Электронный ресурс]. – URL: http://code.google.com/p/russianmorphology/ (дата обращения 19.04.2015).
11. Apache PDFBox [Электронный ресурс]. – URL: https://pdfbox.apache.org (дата обращения 19.04.2015).
12. **Турдаков, Д.** Texterra: инфраструктура для анализа текстов [Электронный ресурс] / Д. Турдаков, Н. Астраханцев, Я. Недумов, А. Сысоев, И. Андрианов, В. Майоров, Д. Федоренко, А. Коршунов, С. Кузнецов. – 2014. – URL: http://www.ispras.ru/ru/proceedings/docs/2014/26/1/isp_26_2014_1_421.pdf (дата обращения 19.04.2015).

13. Serelex [Электронный ресурс]. – URL: http://serelex.cental.be (дата обращения 19.04.2015).
14. WordNet [Электронный ресурс]. – URL: https://wordnet.princeton.edu/ (дата обращения 25.05.2015).
15. **Baroni, M.** The wacky wide web: A collection of very large linguistically processed web-crawled corpora / M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta [Электронный ресурс]. – 2008. – URL: http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf (дата обращения 19.04.2015).
16. **Шеннон, К.** Работы по теории информации и кибернетики / К. Шеннон; пер. с англ. – М.: Иностранная литература, 1963. – С. 669–686. (Shannon C.E. Prediction and entropy of printed English. BSTJ 1951; 30(1): 50-64).

### *References*

[1] Soifer VA, Kupriyanov AV. Analysis and recognition of the nanoscale images: conventional approach and novel problem statement [In Russian]. Computer Optics 2011; 35(2): 136-44.
[2] Tsarkov SV. Automatic keyphrase extraction for vocabulary reduction in probabilistic topic models [In Russian]. Natural and Technical Sciences 2012; 6: 456-64.
[3] Gurevich I, Trusova Yu, Yashina V. The challenges, the problems and the tasks of the descriptive approach to image analysis. 11th International Conference «Pattern Recognition and Image Analysis: New Information Technologies» (PRIA-11-2013) 2013; 1: 30-5.
[4] Emelyanov GM, Mikhaylov DV, Kozlov AP. Formation of the representation of topical knowledge units in the problem of their estimation on the basis of open tests [In Russian]. Machine Learning and Data Analysis 2014; 1(8): 1089-106.
[5] Mel'chuk IA. An Attempt at a Theory of «Meaning⇔Text» Linguistic Models: Semantics, Syntax [In Russian]. Moscow: Languages of Slavonic Culture; 1999.
[6] Huang E. Paraphrase Detection Using Recursive Autoencoder. Source: [http://nlp.stanford.edu/courses/cs224n/2011/reports/ehhuang.pdf].
[7] Jones KS. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 2004; 60(5): 493-502.
[8] Zagoruiko NG. Applied Methods of Data and Knowledge Analysis [In Russian]. Novosibirsk: Institute of Mathematics SD RAS; 1999.
[9] Vorontsov K, Potapenko A, Frei O, Apishev M, Doikov N, Shapulin A, Chirkova N. Multi-criteria and multimodal probabilistic topic models of text collections. International Conference «Intelligent Information Processing» IIP-10 2014; 199.
[10] russianmorphology: Russian Morphology for lucene. Source: [http://code.google.com/p/russianmorphology/].
[11] Apache PDFBox. Source: [https://pdfbox.apache.org].
[12] Turdakov D, Astrakhantsev N, Nedumov Ya, Sysoev A, Andrianov I, Mayorov V, Fedorenko D, Korshunov A, Kuznetsov S. Texterra: A Framework for Text Analysis. Source: [http://www.ispras.ru/ru/proceedings/docs/2014/26/1/isp_26_2014_1_421.pdf].
[13] Serelex. Source: [http://serelex.cental.be].
[14] WordNet. Source: [https://wordnet.princeton.edu/].
[15] Baroni M, Bernardini S, Ferraresi A, Zanchetta E. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. Source: [http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf].
[16] Shannon CE. Prediction and entropy of printed English. BSTJ 1951; 30(1): 50-64.

## AN APPROACH BASED ON TF-IDF METRICS TO EXTRACT THE KNOWLEDGE AND RELEVANT LINGUISTIC MEANS ON SUBJECT-ORIENTED TEXT SETS

*D.V. Mikhaylov [1], A.P. Kozlov [1], G.M. Emelyanov [1]*
*[1] Yaroslav-the-Wise Novgorod State University, Novgorod, Russia*

### *Abstract*

In this paper we look at a problem of extracting knowledge units from the sets of subject-oriented texts. Each such text set is considered as a corpus. The main practical goal here is finding the most rational variant to express the knowledge fragment in a given natural language for further reflection in the thesaurus and ontology of a subject area. The problem is of importance when constructing systems for processing, analysis, estimation and understanding of information represented, in particular, by images. In this paper, by applying the TF-IDF metrics to classify words of the initial phrase in relation to given text corpora we address the task of selecting phrases closest to the initial one in terms of the described fragment of actual knowledge or forms of its expression in a given natural language.

*Keywords:* pattern recognition, intelligent data analysis, information theory, open-form test assignment, natural-language expression of expert knowledge.

*Citation:* Mikhaylov DV, Kozlov AP, Emelyanov GM. An approach based on TF-IDF metrics to extract the knowledge and relevant linguistic means on subject-oriented text sets. Computer Optics 2015; 39(3): 429-38.

### *Сведения об авторах*

**Михайлов Дмитрий Владимирович,** 1974 года рождения, в 1997 году окончил Новгородский государственный университет имени Ярослава Мудрого по специальности 2204 «Программное обеспечение вычислительной техники и автоматизированных систем». В 2003 году защитил диссертацию на соискание учёной степени кандидата, а в 2013 году – доктора физико-математических наук. В настоящее время работает доцентом кафедры информационных тех-

нологий и систем в федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Новгородский государственный университет имени Ярослава Мудрого». Опубликовал более 80 научных работ (из них более 20 статей в рецензируемых журналах из списка ВАК). Область научных интересов: интеллектуальный анализ данных, компьютерная лингвистика.

E-mail: *Dmitry.Mikhaylov@novsu.ru* .

**Dmitry Vladimirovich Mikhaylov** (b. 1974) graduated from Yaroslav-the-Wise Novgorod State University in 1997, specializing in the «Software of Computers and Automated Systems». Obtained his PhD (Kandidat Nauk) and his Doctoral (Doktor Nauk) degrees in Physics and Mathematics in 2003 and 2013, respectively. Currently he works as the Docent of the Information Technologies and Systems department at the same university. Author of more than 80 scientific papers. Research interests are intelligent data analysis and computational linguistics.

**Козлов Александр Павлович,** 1989 года рождения, окончил Новгородский государственный университет имени Ярослава Мудрого (НовГУ) в 2011 г. по специальности «Программное обеспечение вычислительной техники и автоматизированных систем», аспирант кафедры информационных технологий и систем НовГУ. Область научных интересов: интеллектуальный анализ данных, компьютерная лингвистика.

E-mail: *caleo@yandex.ru* .

**Alexander Pavlovich Kozlov** (b.1989) graduated from Yaroslav-the-Wise Novgorod State University in 2011, specializing in the «Software of Computers and Automated Systems». Now he is post-graduate student of the same university. Research interests are intelligent data analysis and computational linguistics.

**Емельянов Геннадий Мартинович,** 1943 года рождения, окончил Ленинградский электротехнический институт им. В.И. Ульянова (Ленина) в 1966 году по специальности «Математические и счётно-решающие приборы и устройства». В 1971 году защитил диссертацию на соискание учёной степени кандидата технических наук. Доктор технических наук (1990 год). В настоящее время – профессор кафедры информационных технологий и систем. Его научные интересы включают построение проблемно-ориентированных вычислительных систем обработки и анализа изображений. Автор более 150 научных работ.

E-mail: *Gennady.Emelyanov@novsu.ru* .

**Gennady Martinovich Emel'yanov** (b. 1943) graduated from the Leningrad Institute of Electrical Engineering in 1966. Obtained his PhD (Kandidat Nauk) and his Doctoral (Doktor Nauk) degrees in 1971 and 1990, respectively. Now he is a Professor of the Information Technologies and Systems department at the same university. Scientific interests include the construction of problem-oriented computing systems of image processing and analysis. He is the author of more than 150 publications.

*Поступила в редакцию 22 апреля 2015 г.*
*Окончательный вариант – 2 июня 2015 г.*