ВЫДЕЛЕНИЕ ЗНАНИЙ, ЯЗЫКОВЫХ ФОРМ ИХ ВЫРАЖЕНИЯ И ОЦЕНКА ЭФФЕКТИВНОСТИ ФОРМИРОВАНИЯ МНОЖЕСТВА ТЕМАТИЧЕСКИХ ТЕКСТОВ

Д.В. Михайлов[1], А.П. Козлов[1], Г.М. Емельянов[1]

[1] Новгородский государственный университет имени Ярослава Мудрого, Великий Новгород, Россия

Аннотация

Статья посвящена взаимосвязанным проблемам выделения единиц знаний из множества (корпуса) тематических текстов и отбора текстов в корпус анализом релевантности исходной фразе. Данные проблемы актуальны для построения систем обработки, анализа, оценивания и понимания информации. Конечной практической целью является поиск наиболее рационального варианта передачи смысла средствами заданного естественного языка для последующей фиксации фрагментов знаний в тезаурусе и онтологии предметной области. При этом релевантность текста по описываемому фрагменту знания (включая формы выражения в языке) определяется суммарной численной оценкой силы связи встречающихся в его фразах сочетаний слов исходной фразы. В настоящей работе рассматриваются известные варианты такой оценки и особенности их использования для выделения составляющих образа исходной фразы в виде слов и их сочетаний в текстах при формировании тематического текстового корпуса. По сравнению с поиском совокупностей указанных составляющих на синтаксически размеченном текстовом корпусе, предложенный в работе метод отбора текстов позволяет в среднем в 15 раз сократить выход фраз, не релевантных исходной ни по описываемому фрагменту знания, ни по языковым формам его выражения.

Ключевые слова: распознавание образов, интеллектуальный анализ данных, теория информации, тест открытой формы, языковое представление экспертных знаний, контекстнозависимое аннотирование, поисковое ранжирование документов.

Цитирование: **Михайлов, Д.В.** Выделение знаний, языковых форм их выражения и оценка эффективности формирования множества тематических текстов / Д.В. Михайлов, А.П. Козлов, Г.М. Емельянов // Компьютерная оптика. – 2016. – Т. 40, № 4. – С. 572-582. – DOI: 10.18287/2412-6179-2016-40-4-572-582.

Введение

Эффективность методов и алгоритмов распознавания образов и интеллектуального анализа данных во многом определяется спецификой решаемой задачи [1]. Немаловажную роль при этом играет разработка способов и средств описания самих задач. Как было отмечено в [2], естественным источником знаний при описании задач будут публикации отечественных и зарубежных научных школ по соответствующей проблематике. Актуальная проблема при этом – поиск наиболее рационального варианта передачи смысла в единице знаний, определяемой множеством семантически эквивалентных (СЭ) фраз предметно-ограниченного естественного языка (ЕЯ). Причём помимо отбора фраз из готового текстового корпуса, важнейшей составляющей здесь является формирование самого корпуса с включением в него публикаций, максимально релевантных рассматриваемым экспертом ситуациям действительности и языковым формам их описания. Такая задача возникает, в частности, при построении учебных курсов с использованием открытых тестов. При этом:

- отбор текстов в корпус, как правило, субъективен;
- выбор критерия отбора текстов – задача нетривиальная. Здесь учитывается и уровень сложности текста, и его значимость в решаемой задаче (например, с точки зрения тематической рубрикации [3] для составления теста по тем или иным фрагментам экспертного знания);
- значимость текста в решаемой задаче может определять выбор меры его близости исходной фразе.

Следует отметить, однако, что значимость текста здесь, как правило, безотносительна к образу, представляемому исходной фразой и выделяемому в анализируемых текстах. Сама же исходная фраза лишь в единичных случаях соответствует эталону для сопоставления. Хорошей иллюстрацией данного тезиса могут послужить случаи минимальной встречаемости слова-термина из исходной фразы в текстах корпуса. Как показано в [2], по значению статистической меры TF-IDF слово при этом не может быть безошибочно отнесено к терминам предметной области, равно как и к общей лексике, обеспечивающей переход от исходной фразы к фразам, наиболее близким ей по смыслу (синонимичным перифразам). В качестве же требований к соотношениям составляющих выделяемого в тексте образа здесь отметим следующее:

- фрагмент анализируемого текста, отвечающий составляющей образа, отождествим с некоторой смысловой связью слов в исходной фразе;
- сила связи слов каждого такого фрагмента всегда больше силы связи любого слова данного фрагмента и слова, не принадлежащего ему;
- слабосвязанные слова исходной фразы не могут отождествляться (по определению) с одним фрагментом. Очевидно, что сочетания общей лексики и терминов исходной фразы, преобладающих в корпусе, в анализируемом тексте можно отнести к составляющим искомого образа только по присутствию других фрагментов с большей силой связи слов.

Отметим также, что в общем случае не выдвигается требование наличия в тексте строго заданной части составляющих образа исходной фразы (ОИФ). По-

этому корректное выделение этого образа во всех отбираемых в корпус текстах предполагает исследование встречаемости и отдельных слов, и их сочетаний с оценкой «силы» связи слов относительно текста и корпуса. В настоящей работе рассматриваются варианты такой оценки и эффективность их использования для выделения составляющих ОИФ при формировании тематического текстового корпуса.

1. Выбор оценки силы связи слов

В задачах информационного поиска, компьютерной лексикографии, выявления плагиата, а также упомянутой выше тематической рубрикации текстов используется понятие «словосочетание» (синонимы: «коллокация», «фразеологический оборот») применительно к цепочке минимум двух связанных по смыслу слов, где выбор одного слова зависит от выбора другого. При этом в качестве инструментов выделения словосочетаний используют частоту *L*-грамм (по К. Шеннону, [4]), частоту и фильтрацию по тэгам, а также математическое ожидание и дисперсию. Для проверки же статистической значимости словосочетания применяют методы проверки статистических гипотез, а именно: *t*-критерий Стьюдента, критерий согласия Пирсона (так называемый критерий χ^2), а также критерий отношения правдоподобия. Отметим, что данные методы для своей реализации требуют синтаксически размеченный текстовый корпус. Это необходимо (в первую очередь) для выделения биграмм, с которыми в этих методах ассоциируются словосочетания. Синтаксическая разметка текстов корпуса не поддаётся полной автоматизации и требует существенных временных затрат. Существующие же корпуса, например, [5], в большинстве случаев не содержат требуемых данных по биграммам из анализируемых текстов.

Другая оценка значимости словосочетаний, выделяемых в тексте из n фраз при формировании текстового корпуса фиксированного ЕЯ, была предложена в работе [6]. Данная оценка сравнима с *G*-тестом [7] для распределения Пуассона и определяется как

$$sig\left(A,B\right)=x-k\log x+\log k!, \qquad (1)$$

где $x=ab/n$, a – число фраз, которые содержат слово A, b – слово B, k – A и B одновременно.

Отметим, что для корректного применения оценки (1) каждое из слов пары (A, B) должно присутствовать минимум в одной фразе анализируемого текста. Менее критичен к указанному требованию дистрибутивно-статистический метод построения тезаурусов, изложенный в работе [8]. Этот метод содержательно наиболее близок рассматриваемой нами задаче выделения в анализируемых текстах образа, представляемого исходной фразой. Основная гипотеза метода заключается в наличии некоторой связи между словами, совместно встречающимися в пределах некоторого текстового интервала, в частности, в пределах одной фразы. При этом каких-либо ограничений на применяемые количественные оценки совместной встречаемости слов не накладывается. Тем не менее, вне зависимости от вида используемая оценка, подобно (1), обычно учитывает частотные характеристики совместной встречаемости слов пары и одиночной встречаемости каждого слова.

Из оценок для «силы» связи слов в дистрибутивно-статистическом методе наиболее наглядна, но в то же время учитывает встречаемость каждого слова в отдельности оценка, содержательно близкая коэффициенту Танимото [9] и определяемая как

$$K_{AB}=k/(a+b-k), \qquad (2)$$

где a, b и k имеют тот же смысл, что и в (1). Содержательно близкая оценка использовалась нами и в [10] для оценки силы связи слов относительно множества СЭ-фраз. При анализе релевантности текста исходной фразе в настоящей работе мы воспользуемся вариантом формулы (2), где к значению в знаменателе прибавляется единица в целях предотвращения деления на ноль в случае, когда слова A и B не встречаются во фразах анализируемого текста отдельно друг от друга. Для сравнения в качестве альтернативы указанной величине возьмём оценку (1), приняв значение $sig(A,B)$ равным нулю при равенстве нулю a, b либо k. В целях наглядности изложения далее в работе мы сохраним обозначения $sig(A,B)$ и K_{AB} применительно к введённым вариантам оценок (1) и (2).

2. Отбор максимально релевантных текстов

Пусть X – упорядоченная по убыванию последовательность ненулевых значений $sig(A,B)$ либо K_{AB} относительно текстового документа d для пар слов (A,B), которым в исходной фразе соответствуют некоторые синтаксические связи. Разобьём X на кластеры $H_1,\ldots,H_r$ с применением алгоритма, содержательно близкого алгоритмам класса FOREL [11]. За центр масс кластера H_i здесь, как и в работах [2, 10], мы возьмём 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: $H_i := X$;
2: $X_p := \varnothing$;
3: $X_s := \varnothing$;
4: **если** good(H_i) = true **или** diam(H_i) = 1 **то**
вернуть H_i, X_p и X_s;
5: **иначе если** $|\mathrm{mc}(H_i)-\mathrm{first}(H_i)| > |\mathrm{mc}(H_i)-\mathrm{last}(H_i)|$ **то**
6: $X_p := \{\mathrm{first}(H_i)\} \bullet X_p$;
7: $H_i := \mathrm{rest}(H_i)$;
8: перейти к шагу 4;
9: **иначе**
если $|\mathrm{mc}(H_i)-\mathrm{first}(H_i)| < |\mathrm{mc}(H_i)-\mathrm{last}(H_i)|$ **то**
10: $X_s := \{\mathrm{last}(H_i)\} \bullet X_s$;
11: $H_i := \mathrm{lrev}(H_i)$;
12: перейти к шагу 4;
13: **иначе**
14: $X_s := \{\mathrm{last}(H_i)\} \bullet X_s$;
15: $X_p := \{\mathrm{first}(H_i)\} \bullet X_p$;
16: $Tmp := \mathrm{lrev}(H_i)$;
17: $H_i := \mathrm{rest}(Tmp)$;

18: перейти к шагу 4;
Конец *{Алгоритм 1}*.

Здесь и далее «•» – операция конкатенации.

Табл. 1. Вспомогательные функции Алгоритма 1

Функция	**Возвращаемое значение**
first(X), last(X)	первый/последний элемент последовательности X
lrev(X), rest(X)	исходная последовательность X без последнего/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

Так же, как и в [2, 10], элементы последовательности X будут отнесены к одному кластеру, если

$$\begin{cases} \left| \mathrm{mc}(X) - \mathrm{first}(X) \right| < \dfrac{\mathrm{mc}(X)}{4} \\ \left| \mathrm{mc}(X) - \mathrm{last}(X) \right| < \dfrac{\mathrm{mc}(X)}{4} \end{cases}. \qquad (3)$$

Алгоритм 1 применяется к последовательностям X_p и X_s на его выходе. Данный процесс продолжается рекурсивно до тех пор, пока на очередном шаге X_p и X_s не окажутся пустыми. В итоге исходная последовательность X разбивается на подпоследовательности-кластеры $H_1,\ldots,H_r$, при этом для $\forall i \neq j\ H_i \cap H_j = \varnothing$, а $H_1 \bullet H_2 \bullet \ldots \bullet H_r = X$.

Обозначим суммарную величину оценки «силы» относительно d для всех найденных в исходной фразе связей (A, B) как $K_\Sigma(d)$, а для связей, отнесённых к кластеру H_1 наибольших значений указанной величины, – как $K_1(d)$. Тогда для максимально релевантных документов будет найдено максимумом связей с наибольшими значениями «силы» при максимальных значениях функции K_Σ. Функцию ранжирования документов по релевантности исходной фразе можно определить из геометрических соображений как

$$W(d) = K_\Sigma(d)\frac{K_1(d)}{K_\Sigma(d)} = K_1(d). \qquad (4)$$

Оценка (4) идейно близка описанному в [12] методу определения неестественного происхождения текстового документа, когда он либо генерируется автоматически, либо есть результат работы системы уникализации контента. Действительно, если в неестественном тексте количество редких, нехарактерных для языка сочетаний слов обычно завышено по сравнению со стандартом, а количество частых пар – занижено, то в отбираемом тексте, релевантном исходной фразе, аналогичная ситуация наблюдается с сочетаниями слов, представляемых кластером H_1.

Сортировкой анализируемых документов по убыванию значений функции (4) с последующим разделением на кластеры *Алгоритмом 1* отбираются документы с наибольшими значениями данной оценки (принадлежащими первому кластеру в составе формируемой последовательности). Следуя нотации работы [2], обозначим далее множество указанных документов как D′, где D – исходное текстовое множество.

3. Выделение единиц знаний из текстов

Следующий шаг – отбор фраз из документов $d \in D'$, наиболее близких исходной фразе по представляемым фрагментам знаний. Данный шаг может быть реализован двумя способами: либо по числу найденных во фразе связей, отвечающих кластеру H_1 из формируемых *Алгоритмом 1*, либо по суммарному значению «силы» указанных связей. И в том, и в другом случае отбираемые фразы $s \in d$ кластеризуются по значению выбранной оценки (обозначим её далее символом N) с помощью *Алгоритма 1*, а в качестве результата возвращается набор фраз, которому отвечает кластер наибольших значений. Идейно данный подход близок исследованиям U. Manber [13] и N. Heintze [14] в области нахождения нечетких дубликатов текстовых документов, где в качестве меры сходства двух документов используется отношение числа общих подстрок фиксированной длины (в нашем случае эта длина была бы равна двум) к размеру документа (в словах). Содержательно здесь мы имеем разновидность контекстно-зависимого аннотирования [15], где одна аннотация строится сразу для нескольких документов. Следуя терминологии поисковых систем, назовём далее поиск фраз, близких исходной, в документах $d \in D'$ построением аннотации.

Пусть $S = \{s: s \in d,\ d \in D'\}$, а X^W и X^{N} – последовательности значений оценок $W(d)$ и $\mathrm{N}(s)$ для документов $d \in D$ и фраз $s \in S$ соответственно. Кластеры, формируемые на X^W и X^{N}, а также на D и S, обозначим далее как $H_1^W \ldots H_{r(D)}^W$ и $H_1^D \ldots H_{r(D)}^D$ и, соответственно, $H_1^{\mathrm{N}} \ldots H_{r(S)}^{\mathrm{N}}$ и $H_1^S \ldots H_{r(S)}^S$. Тогда в совокупности с отбором текстов в корпус построение аннотации здесь может быть формально описано следующей совокупностью шагов.

Алгоритм 2. Построение аннотации.
Вход: D; // исходное текстовое множество
Выход: H_1^S; // результирующее множество фраз
Начало
1: $X^W := \varnothing$;
2: **для всех** $d \in D$
3: вычислить $W(d)$ согласно формуле (4);
4: $X^W := X^W \cup \{W(d)\}$;
5: отсортировать X^W и D
по убыванию значения функции W;
6: сформировать
$H_1^W \bullet H_2^W \bullet \ldots \bullet H_{r(D)}^W$ **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: сформировать S;
10: **для всех** $s \in S$
11: вычислить $\mathrm{N}(s)$;
12: $X^{\mathrm{N}} := X^{\mathrm{N}} \cup \{\mathrm{N}(s)\}$;

13: отсортировать X^{N} и S
по убыванию значения функции N;
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};
Конец *{Алгоритм 2}*.

Как видно из представленного псевдокода, по вычислительной сложности предложенный метод в целом идентичен ранее разработанному и описанному авторами в [2]. Учитывая независимость оценок (1) и (2) от конкретного текстового документа, для снижения вычислительных затрат, помимо изложенных в [2] рекомендаций, здесь целесообразно запоминать найденные ранее значения «силы» связи слов, храня их в отсортированном виде, а также выполнять однократно кластеризацию связей для всех исходных фраз.

Следует отметить, что в отличие от вышеупомянутой задачи нахождения нечётких дубликатов слова в составе выделяемой связи во фразе не всегда расположены по соседству, т.е. в общем случае не образуют в ней подстроку. Кроме того, для сочетаний общей лексики и терминов здесь также нельзя делать предположения относительно расстояния между словами выделяемой биграммы. В то же время слова биграммы в нашем случае точно выделяются по разделителям (знакам препинания), поэтому задействовать поиск подстроки для случаев её произвольного вхождения в текст (алгоритмы Рабина–Карпа [16], Кнута–Морриса–Пратта [17], Бойера–Мура [18]) здесь не требуется, что также дополнительно снижает вычислительные затраты.

4. Экспериментальные исследования

Исходные множества текстов для апробации предложенного метода формирования тематического корпуса подбирались таким образом, чтобы сравнить образы исходной фразы, выделяемые в текстах на основе оценки силы связи встречающихся в их фразах слов исходной фразы и на основе меры TF-IDF этих слов [2]. Сами исходные фразы формулировались независимо друг от друга разными экспертами. Основным критерием здесь была максимально полная и наглядная иллюстрация выявления в текстах корпуса контекстов использования как слов-терминов, так и общей лексики, обеспечивающей синонимические перифразы исходной фразы. При этом с учётом упомянутой выше особенности оценки (1) число фраз в текстовом документе предполагалось не менее пяти. Данное ограничение обусловлено отмеченной в [7] проблемой малых выборок для *G*-теста.

Первый вариант исходного текстового множества включал публикации по философским и методологическим проблемам инженерии знаний, в том числе:
- 1 статью в журнале «Вестник Российского экономического университета им. Г.В. Плеханова (Вестник РЭУ)»;
- 1 статью в журнале «Философия науки»;
- материалы тезисов четырёх докладов на 4-й Всероссийской конференции студентов, аспирантов и молодых учёных «Искусственный интеллект: философия, методология, инновации» (ИИ ФМИ, 2010 г.);
- материалы тезисов двух секционных докладов, а также одного пленарного доклада на 7-й Всероссийской конференции студентов, аспирантов и молодых учёных «Искусственный интеллект: философия, методология, инновации» (2013 г.);
- материалы одного пленарного доклада на 8-й Всероссийской конференции студентов, аспирантов и молодых учёных «Искусственный интеллект: философия, методология, инновации» (2014 г.);
- 1 статью в сборнике трудов 9-й Всероссийской конференции студентов, аспирантов и молодых учёных «Искусственный интеллект: философия, методология, инновации» (2015 г.);
- 1 статью в журнале «Таврический вестник информатики и математики (ТВИМ)».

При этом число слов в документах варьировалось от 618 до 3765, число фраз в них – от 38 до 276.

Второй вариант исходного множества текстов полностью совпадал с текстовым корпусом из экспериментов в [2]. Для сравнения: число слов в документах здесь варьировалось от 218 до 6298, число фраз – от 9 до 587.

В экспериментах по формированию текстового корпуса и выделению составляющих ОИФ участвовали две группы фраз (табл. 2 и 3) соответственно для первого и второго варианта исходного текстового множества.

Программная реализация метода на языке Java и результаты экспериментов представлены на портале НовГУ по адресу: http://www.novsu.ru/file/1195999 в ZIP-архиве (редакция от 20.02.2016). Архив включает каталог с исполняемым jar-файлом, подкаталогами текстовых корпусов и обученной модели классификатора для выделения границ предложений (подробнее – там же, файл *readme.doc*). В этом же каталоге находятся примеры аннотации (*issue.txt*), найденных морфологических характеристик слов исходной фразы (*issue_ch_lemmas.txt*) и их связей (*issue_rels.tst*).

В качестве примера следует привести выделение составляющих образа исходной фразы №8 из табл. 3, для которой разбиением её слов на классы по значению меры TF-IDF в работе [2] удовлетворительного решения указанной задачи найдено не было.

Наилучшими здесь оказались результаты для введённого нами варианта оценки (2), где по максимуму числа «наиболее сильных» связей со значениями K_{AB}, отнесённых к кластеру H_1 из формируемых *Алгоритмом 1*, была отобрана фраза: *«Системы ограничений, возникающие в задачах принятия решений, оптимизации, распознавания образов и анализа часто являются несовместными, подразумевающими те или иные подходы к их коррекции, связанной с обобщением классического понятия решения»*.

Данная фраза отобрана из тезисов доклада М.Ю. Хачая на 16-й Всероссийской конференции «Математические методы распознавания образов» (2013 г.). Этот документ здесь оказался единственным из лучших по критерию (4) и отнесённых к кластеру H_1^D (т.е. множеству D'). «Наиболее сильными»

по значению K_{AB} здесь оказались связи для пар слов *«распознавание – с»* и *«принятие – решение»*. Заметим, что первая из них при этом подтверждает выдвинутый во введении тезис о возможности выделения в анализируемом тексте сочетаний слабосвязанных слов по присутствию в нём фрагментов с большей силой связи. Сказанное относится к сочетаниям со словом *связанный*, посредством которых фрагмент знаний, определяемый исходной фразой №8 из табл. 3, связывается со знаниями других экспертов, а именно, представление о *системе ограничений, возникающих в задачах принятия решений, оптимизации, распознавания образов и анализа*, как основе *выбора правила принятия решений*.

Табл. 2. Исходные фразы, предметная область «Философия и методология инженерии знаний»

№	Исходная фраза
1	*Определение модели представления знаний накладывает ограничения на выбор соответствующего механизма логического вывода.*
2	*Под знанием понимается система суждений с принципиальной и единой организацией, основанная на объективной закономерности.*
3	*С точки зрения искусственного интеллекта знание определяется как формализованная информация, на которую ссылаются или используют в процессе логического вывода.*
4	*Факты обычно указывают на хорошо известные обстоятельства в данной предметной области.*
5	*Эвристика основывается на собственном опыте специалиста в данной предметной области, накопленном в результате многолетней практики.*
6	*Метазнания могут касаться свойств, структуры, способов получения и использования знаний при решении практических задач искусственного интеллекта.*
7	*Однородность представления знаний приводит к упрощению механизма управления логическим выводом и упрощению управления знаниями.*
8	*Отличительными чертами логических моделей являются единственность теоретического обоснования и возможность реализации системы формально точных определений и выводов.*
9	*Язык представления знаний на основе фреймовой модели наиболее эффективен для структурного описания сложных понятий и решения задач, в которых в соответствии с ситуацией желательно применять различные способы вывода.*

Указанный тезис подтверждается и в эксперименте с тем же вариантом оценки (2), но отбором фраз по суммарной силе связей кластера «наиболее сильных», где в аннотацию, помимо вышеупомянутой, вошла фраза *«Современная теория комитетных решений и тесно связанных с ними комитетных методов обучения распознаванию опирается на фундаментальные результаты, полученные Вл. Д. Мазуровым»*. Из «наиболее сильных» по величине K_{AB} пар слов здесь содержится только *«распознавание – с»*. Но при большем K_{AB} (0,4000 против 0,3333 у *«принятие – решение»*) данное сочетание слов позволяет посредством того же слова *«связанный»* соотнести понятие *распознавание* из исходной фразы с понятием *обучение*.

Табл. 3. Исходные фразы, предметная область «Математические методы обучения по прецедентам»

№	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.*

Для оценки (1) результаты оказались несколько хуже, а именно: при тех же выделенных сочетаниях слов *«распознавание – с»* и *«принятие – решение»* (на основе значений $sig(A,B)$) по максимуму их встречаемости было отобрано 10 фраз, из которых значима для соотнесения знаний разных экспертов лишь одна: *«Эмпирический решающий лес повысил эффективность распознавания объектов, не участвовавших ранее в обучении, по сравнению с одним решающим деревом, при использовании одного и того же критерия ветвления»* [Дюличева Ю.Ю., ТВИМ 2003 №2]. Эта же фраза, представляющая рассуждение о *решающем дереве* и *лесе* как способах представления *решающих правил*, была и в числе трёх результирующих в эксперименте с тем же вариантом оценки (1) и отбором фраз по суммарному значению силы связей из кластера «наиболее сильных».

Сравнение ОИФ, выделяемого на основе оценки силы связи слов исходной фразы и на основе TF-IDF этих слов, наиболее наглядно иллюстрируется экспериментами с фразами из табл. 2, в текстах предметной области которых доля общей лексики больше аналогичного показателя для фраз табл. 3. Действительно, возьмём для примера фразу №9 из табл. 2. Слова, представленные в кластерах H_1, $H_{r/2}$ и H_r из формируемых для этой фразы *Алгоритмом 1* по значению меры TF-IDF её слов вместе с документами, послужившими источниками отбора результирующих фраз, приведены в табл. 4.

Как видно из таблицы, большая часть общей лексики, которая могла бы обеспечивать синонимические перифразы исходной фразы, попадает в кластер наименьших значений меры TF-IDF (слова *с, который, на, в, основа, для*). Более того, значение TF-IDF для

слов указанного кластера здесь равно нулю. В итоге из десяти отобранных фраз не нашлось семантически эквивалентных для исходной, равно как и фраз, связывающих упоминаемые в исходной фразе понятия с другими понятиями той же предметной области.

Табл. 4. Кластеры по значению меры TF-IDF для отбора фраз (фраза №9, табл. 2)

Янковская А.Е., ТВИМ 2004 №1, слова, представленные в кластерах	
H_1	*различный*
$H_{r/2}$	*применять, модель, наиболее, ситуация, соответствие*
H_r	*с, решение, понятие, сложный, который, вывод, фреймовый, на, задача, в, и, основа, для, знание*

Предложенный в настоящей работе метод построения контекстно-зависимой аннотации при отборе фраз по максимуму числа «наиболее сильных» связей в числе результирующих здесь даёт фразу: *«Специфика структурно-фреймовой организации состоит в том, чтобы во фрейме (а он представляет собой достаточно сложную концептуальную конструкцию, записанную средствами программной части вычислительной (информационной) системы) все понятия, относящиеся к охватываемой данным фреймом предметной области, имели внутреннюю интерпретацию, т.е. были наделены смыслом на соответствующем языке представления знаний»* [Русанов В.В., Вестник РЭУ 2012 №1]. Этот результат уже значим для соотнесения понятий «*сложная концептуальная конструкция – сложное понятие – внутренняя интерпретация*» и «*структурное описание – язык представления знаний*».

Кроме того, при использовании K_{AB} с отбором фраз по суммарному значению силы «наиболее сильных» связей результирующая фраза «*Фреймовые структуры реализуются на базе языков программирования высокого уровня, позволяющих человеку работать с информационной системой, используя лингвистические средства, близкие к языку межчеловеческого общения*» того же документа позволяет соотнести понятия «*структурное описание*» и «*язык программирования высокого уровня*». Для сравнения в табл. 5 представлены сочетания слов, отвечающие «*наиболее сильным*» связям (в порядке убывания оценки «*силы*»), по присутствию которых осуществлялся выбор фраз в данном примере. Поскольку сочетания с предлогами значимы для перифраз вида «*на основе ⇔ на базе*», то подобные сочетания слов также не были исключены из рассмотрения.

И, наконец, рассмотрим выделение предложенным методом составляющих образа исходной фразы №9 из табл. 3, для которой метод на основе меры TF-IDF в [2] показал один из лучших результатов. Применение введённого нами варианта оценки (2) в качестве наиболее близкой исходной здесь даёт единственную фразу: *«Увеличение сложности структуры решающего дерева и уменьшение его обобщающей способности наблюдаются при все более точной, безошибочной "настройке" решающего дерева на исходную обучающую информацию»* [Дюличева Ю.Ю. Математические методы распознавания образов (ММРО-13)].

Табл. 5. Документы, лучшие по критерию (4), и связи слов исходной фразы №9 из табл. 2

Оценка	*«Наиболее сильные» связи для отбора фраз*
Русанов В.В., Вестник РЭУ 2012 №1	
K_{AB}	*язык – на, язык – сложный, на – основа, представление – с, язык – фреймовый, представление – в, представление – для, представление – понятие, язык – основа*
sig(A, B)	*язык – на, на – основа, язык – сложный, язык – фреймовый, основа – с*
Лекторский В.А., ИИ ФМИ, 2014 г.	
K_{AB}	*язык – задача, представление – способ, основа – модель, модель – для, модель – применять, сложный – понятие, в – знание, описание – применять, решение – различный, на – описание*
Крымская Е.Ю., ИИ ФМИ, 2010 г.	
K_{AB}	*решение – задача, решение – с, задача – в, на – решение, решение – для*
Янковская А.Е., ТВИМ 2004 №1	
sig(A, B)	*на – основа, решение – задача*

По варианту оценки (1) в данном примере результат вполне сравним с полученным в [2]: из фраз, содержащих те или иные составляющие образа исходной фразы (не относящиеся исключительно к общей лексике), не найденной оказалась всего одна: *«Наиболее общая теория алгоритмических композиций разработана в алгебраическом подходе к построению корректных алгоритмов, предложенном академиком РАН Ю.И. Журавлёвым и активно развиваемом его учениками»* [Воронцов К.В., ТВИМ 2004 №1], ср. *«композиция алгоритмов ⇔ алгоритмическая композиция»*. Как видно из табл. 6 и 7, помимо *«композиция – алгоритм»*, в число «наиболее сильных» здесь также не вошли сочетания с предлогом «*к*», значимые для перифраз вида *«ведёт к ⇔ приводит к»*.

Табл. 6. Кластеры по значению меры TF-IDF для отбора фраз (фраза №9, табл. 3)

Воронцов К.В., ТВИМ 2004 №1, слова, представленные в кластерах	
H_1	*алгоритм, обобщать, способность*
$H_{r/2}$	*к, классификатор, увеличение*
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*

Тем не менее, предложенный в настоящей работе метод дал меньший, чем метод на основе TF-IDF, выход фраз, не релевантных исходной ни по описываемому фрагменту знания, ни по языковым формам его выражения. Для сравнения в табл. 8 и 9 для исходных фраз из табл. 2 и 3 приведено общее число отобранных фраз (N), в том числе представляющих выразительные средства языка (N_1), синонимы (N_2) и связи понятий предметной области (N_3).

Табл. 7. Документы, лучшие по критерию (4), и связи слов исходной фразы №9 из табл. 3

Оценка	*«Наиболее сильные» связи для отбора фраз*
Дюличева Ю.Ю., ММРО-13	
K_{AB}	*увеличение – обобщать,* *увеличение – способность,* *обобщать – способность*
Воронцов К.В., ТВИМ 2004 №1	
$sig(A, B)$	*обобщать – способность*

Табл. 8. Отбор релевантных фраз для представленных в табл. 2: сравнение с методом на основе TF-IDF

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
на основе TF-IDF слов исходной фразы									
N	5	8	14	9	1	1	29	5	10
N_1	0	0	0	0	0	0	1	0	0
N_2	0	0	1	0	1	0	1	0	0
N_3	2	1	0	1	0	0	1	0	0
по числу «наиболее сильных» связей по величине K_{AB}									
N	2	4	1	3	2	1	6	1	5
N_1	0	1	0	1	2	1	0	0	0
N_2	0	0	0	2	2	1	0	0	0
N_3	1	2	0	0	0	0	2	0	1
по суммарной «силе» для «наиболее сильных» по K_{AB}									
N	1	12	15	1	1	2	2	1	11
N_1	0	1	1	0	1	0	0	0	1
N_2	0	0	0	0	1	0	0	0	2
N_3	0	1	1	0	0	0	1	0	4
по числу «наиболее сильных» связей по $sig(A, B)$									
N	3	2	32	1	2	1	18	1	3
N_1	0	0	0	0	0	0	1	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	1	2	1	0	0	0	2	0	1
по суммарной «силе» «наиболее сильных» по $sig(A, B)$									
N	5	5	1	7	7	2	3	2	10
N_1	0	1	0	0	1	0	0	0	0
N_2	0	0	0	0	1	0	0	0	0
N_3	1	2	0	0	0	0	1	0	0

Отметим, что в отличие от предложенного метода, поиск фраз, близких исходной по описываемому фрагменту знания, на синтаксически размеченном текстовом корпусе, охватывающем весь заданный ЕЯ, требует предварительного выделения экспертом в исходной фразе слов и их сочетаний, представляющих термины предметной области. Как видно из табл. 10, найденные при этом фразы из понятийных связей практически не отражают синонимии. Кроме того, результативность поиска здесь зависит от представленности соответствующей тематики в текстах корпуса.

Для сравнения в табл. 11 приведены выделенные экспертом слова и их сочетания для фраз из табл. 2 и 3, входящие минимум в одну фразу из документов Национального корпуса русского языка [5].

Табл. 9. Отбор релевантных фраз для представленных в табл. 3: сравнение с методом на основе TF-IDF

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
на основе TF-IDF слов исходной фразы									
N	1	1	1	1	3	2	4	1	40
N_1	1	1	1	1	0	0	0	0	7
N_2	0	1	1	1	3	0	0	0	6
N_3	0	0	0	0	1	1	1	0	8
по числу «наиболее сильных» связей по величине K_{AB}									
N	1	1	15	15	5	11	1	1	1
N_1	1	1	3	2	0	0	0	0	1
N_2	0	1	2	2	1	9	0	0	1
N_3	0	0	7	4	0	4	0	1	0
по суммарной «силе» для «наиболее сильных» по K_{AB}									
N	10	9	2	2	8	6	2	2	1
N_1	0	0	0	0	1	0	0	0	1
N_2	0	0	0	0	1	4	0	0	1
N_3	1	0	1	0	1	2	0	2	0
по числу «наиболее сильных» связей по $sig(A, B)$									
N	1	1	11	11	5	20	9	10	19
N_1	1	1	1	2	0	1	0	0	2
N_2	0	1	1	1	1	1	1	0	1
N_3	0	0	4	4	0	0	5	1	7
по суммарной «силе» «наиболее сильных» по $sig(A, B)$									
N	9	9	1	1	1	1	6	3	8
N_1	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	0	0	0	1	0	0	1	1	2

Табл. 10. Отбор релевантных фраз из текстов Национального корпуса русского языка [5]

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
для исходных фраз из табл. 2									
N	13	67	2	15	29	30	79	224	20
N_1	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	2	5	0	1	1	2	3	2	2
для исходных фраз из табл. 3									
N	56	1	1	1	24	17	21	5	2
N_1	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	0	0	0	0	0	0	0	1	0

Таким образом, наряду с решением своей основной задачи, будучи совместно используемым с отбором фраз на основе TF-IDF слов исходной фразы, предложенный в настоящей работе метод позволяет автоматизировать выделение экспертом требуемых слов и их сочетаний для организации поиска в синтаксически размеченном текстовом корпусе нехудожественных текстов по заданной тематике. Кроме того, сам отбор текстов в тематический корпус на основе оценки (4) позволяет точно задать его тему совокупностью специальных терминов предметной области, совместно встречающихся в текстовых документах. Как видно из табл. 8–10, при этом в среднем в 15 раз снижается выход фраз, не релевантных исходной.

Табл. 11. Слова и их сочетания для отбора релевантных фраз из Национального корпуса русского языка

№	Слова и сочетания слов
по исходным фразам из табл. 2	
1	*модель – представление – знание, механизм – логический – вывод*
2	*система – суждение, объективный – закономерность*
3	*процесс – логический – вывод*
4	*данный – предметный – область*
5	*эвристика, данный – предметный – область*
6	*метазнания, свойство – знание, структура – знание, способ – получение – знание, способ – использование – знание, задача – искусственный – интеллект*
7	*представление – знание, управление – вывод, механизм – логический – вывод, управление – знание*
8	*теоретический – обоснование – модель, логический – модель, система – вывод, система – определение, точный – вывод*
9	*язык – представление – знание, фреймовый – модель, способ – вывод*
по исходным фразам из табл. 3	
1	*переобучение, эмпирический – риск*
2	*эмпирический – риск*
3	*эмпирический – риск*
4	*эмпирический – риск*
5	*ошибка – средний*
6	*частота – ошибка, контрольный – выборка*
7	*оценка – частота, контрольный – выборка*
8	*ошибка – распознавание, правило – принятие – решение*
9	*базовый – классификатор*

5. Некоторые технические детали и допущения

Классическая постановка задачи кластерного анализа [11] предполагает, что каждый элемент последовательности, разбиваемой на кластеры с применением *Алгоритма 1*, представлен в ней ровно один раз. Как и в [2], с целью наглядности изложение предлагаемого в настоящей работе метода неявно содержит предположение о выполнении данного условия.

Извлечение текста из PDF-файла, а также морфологический анализ словоформ выполнялись теми же методами, что и в [2]. Для выделения синтаксических связей были взяты правила, задействованные в работе [3]. При этом в отличие от исходной фразы, для правильности установления связей слов во фразах документов, анализируемых по критерию (4), рассматривались все выявленные значения многозначных слов. В исходной же фразе для многозначного слова отбирался вариант с наибольшей встречаемостью в исходном текстовом множестве.

Отдельная задача – выделение границ предложений в тексте по знакам препинания. Для её решения использовалась обученная модель классификатора, построенного с применением интегрированного пакета Apache OpenNLP [19]. Обучение распознаванию границ предложений согласно рекомендациям разработчиков пакета осуществлялось на основе размеченных данных из Leipzig Corpora [20], более точно – газетных текстов на русском языке (2010 г., всего 1 млн. фраз). Оценка точности выделения границ предложений для разных вариантов обучения классификатора – тема отдельного исследования, в данной работе при выборе исходных данных для обучения классификатора авторы ограничились максимальным по объёму русскоязычным текстовым корпусом.

Заключение

Основной *результат* настоящей работы – *метод формирования тематического корпуса текстов, релевантных по описываемым фрагментам знаний исходной фразе, с выделением составляющих её образа*.

Отметим, что предложенная оценка релевантности текста исходной фразе ограничивает рассмотрение связей слов биграммами. Но как показали представленные в табл. 10 и 11 результаты экспериментов с Национальным корпусом русского языка, в целях более продуктивного применения предложенных решений, например, при построении специализированных тезаурусов [21], выделяемые во фразах биграммы в ряде случаев целесообразно расширять до трёх и более элементов. Оценивать «силу» связи слов в рамках выделяемых L-грамм при этом можно, например, рекурсивно используя формулу (2) и рассматривая ранее найденные $L-1$-граммы в качестве слов. Представляется также перспективным задействовать здесь совместно с L-граммами слова, отвечающие кластеру наибольших значений TF-IDF из сформированных для исходной фразы *Алгоритмом 1*. В этом плане отдельного исследования заслуживает введение в рассмотрение меры TF-IDF для указанных L-грамм и их классификация наравне с отдельными словами по значению данной меры предложенным в [2] методом.

Благодарности

Работа выполнена при поддержке Министерства образования и науки РФ (базовая часть госзадания), а также гранта РФФИ (№16-01-00004).

Литература

1. **Кольцов, П.П.** О количественной оценке эффективности алгоритмов анализа изображений / П.П. Кольцов, А.С. Осипов, А.С. Куцаев, А.А. Кравченко, Н.В. Котович, А.В. Захаров // Компьютерная оптика. – 2015. – Т. 39, № 4. – С. 542-556. – DOI: 10.18287/0134-2452-2015-39-4-542-556.
2. **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. – DOI: 10.18287/0134-2452-2015-39-3-429-438.
3. **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.
4. **Шеннон, К.** Работы по теории информации и кибернетики / К. Шеннон; пер. с англ. – М.: Иностранная литература, 1963. – С. 669-686.

5. Национальный корпус русского языка [Электронный ресурс]. – URL: http://www.ruscorpora.ru/ (дата обращения 26.02.2016).
6. **Biemann, C.** Language-independent Methods for Compiling Monolingual Lexical Data / C. Biemann, S. Bordag, G. Heyer, U. Quasthoff, C. Wolff // 5th International Conference "Computational Linguistics and Intelligent Text Processing" (CICLing 2004). – 2004. – Vol. 2945. – P. 217-228.
7. **McDonald, J.H.** G-test of goodness-of-fit / J.H. McDonald. – Handbook of Biological Statistics. – Third ed. – Baltimore, Maryland: Sparky House Publishing, 2014. – P. 53-58.
8. Дистрибутивно-статистический метод построения тезаурусов: современное состояние и перспективы / В.А. Москович. – М., 1971. – 66 с.
9. **Tanimoto, T.T.** An elementary mathematical theory of classification and prediction / T.T. Tanimoto. – New York: International Business Machines Corporation, 1958. – 10 p.
10. **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.
11. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Издательство института математики, 1999. – 270 с.
12. **Гречников, Е.А.** Поиск неестественных текстов / Е.А. Гречников, Г.Г. Гусев, А.А. Кустарев, А.М. Райгородский // Труды XI Всероссийской научной конференции RCDL'2009. – Петрозаводск: КарНЦ РАН, 2009. – С. 306-308.
13. **Manber, U.** Finding Similar Files in a Large File System / U. Manber // USENIX Winter 1994 Technical Conference Proceedings. – 1994. – P. 1-10.
14. **Heintze, N.** Scalable Document Fingerprinting / N. Heintze // Proceedings of the Second USENIX Workshop on Electronic Commerce. – 1996. – P. 191-200.
15. **Бродский, А.** Алгоритмы контекстно-зависимого аннотирования Яндекса на РОМИП-2008 / А. Бродский, Р. Ковалев, М. Лебедев, Д. Лещинер, П. Сушин, И. Мучник // Труды РОМИП 2007-2008. – СПб., 2008. – С. 160-169.
16. **Karp, R.M.** Efficient randomized pattern-matching algorithms / Richard M. Karp, Michael O. Rabin // IBM Journal of Research and Development. – 1987. – Vol. 31(2). – P. 249-260. – ISSN 0018-8646.
17. **Knuth, D.** Fast pattern matching in strings / Donald E. Knuth, James H. Morris, Vaughan R. Pratt // SIAM Journal on Computing. – 1977. – Vol. 6(2). – P. 323-350. – ISSN 0097-5397.
18. **Boyer, R.S.** A fast string searching algorithm / Robert S. Boyer, J. Strother Moore // Communications of the ACM. – 1977. – Vol. 20(10). – P. 762-772. – ISSN 0001-0782.
19. Apache OpenNLP [Электронный ресурс]. – URL: https://opennlp.apache.org/ (дата обращения 31.03.2016).
20. Leipzig Corpora Collection Download Page [Электронный ресурс]. – URL: http://corpora2.informatik.uni-leipzig.de/download.html (дата обращения 31.03.2016).
21. **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). Samara, September 23-28, 2013: Conference Proceedings. – Vol. 1. – Samara: IPSI RAS, 2013. – P. 30-35.

Сведения об авторах

Михайлов Дмитрий Владимирович, 1974 года рождения, в 1997 году окончил Новгородский государственный университет имени Ярослава Мудрого (НовГУ) по специальности 2204 «Программное обеспечение вычислительной техники и автоматизированных систем». В 2003 году защитил диссертацию на соискание ученой степени кандидата, а в 2013 году – доктора физико-математических наук. В настоящее время работает доцентом кафедры информационных технологий и систем (ИТиС) в федеральном государственном бюджетном образовательном учреждении высшего 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* .

Козлов Александр Павлович, 1989 года рождения, в 2011 году окончил НовГУ по специальности «Программное обеспечение вычислительной техники и автоматизированных систем», аспирант кафедры ИТиС НовГУ. Область научных интересов: интеллектуальный анализ данных, компьютерная лингвистика. E-mail: *caleo@yandex.ru* .

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

ГРНТИ: 28.23.11, 28.23.15, 20.23.19
Поступила в редакцию 14 апреля 2016 г. Окончательный вариант – 1 июля 2016 г.

EXTRACTION OF KNOWLEDGE AND RELEVANT LINGUISTIC MEANS WITH EFFICIENCY ESTIMATION FOR THE FORMATION OF SUBJECT-ORIENTED TEXT SETS

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

Abstract

In this paper we look at two interrelated problems of extracting knowledge units from a set of subject-oriented texts (the so-called corpus) and selecting texts to the corpus by analyzing the relevance to the initial phrase. 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 problems are of importance when constructing systems for processing, analysis, estimation and understanding of information. In this paper the text relevance to the initial phrase in terms of the described fragment of actual knowledge (including forms of its expression in a given natural language) is defined by the total numerical estimate of the coupling strength of words from the initial phrase jointly occurring in phrases of the text under analysis. The paper considers known variants of such estimation procedures and their application for the search of distinct components which reflect the initial phrase in the texts selected to the topical text corpus. These components correspond to words and their combinations. In comparison with the search of such components on a syntactically marked text corpus, the method for text selection offered in this paper enables a 15-times reduction (on average) in the output of phrases which are irrelevant to the initial one in terms of either the described knowledge fragment or its expression forms in a given natural language.

Keywords: pattern recognition, intelligent data analysis, information theory, open-form test assignment, natural-language expression of expert knowledge, contextual annotation, document ranking in information retrieval.

Citation: Mikhaylov DV, Kozlov AP, Emelyanov GM. Extraction of knowledge and relevant linguistic means with efficiency estimation for the formation of subject-oriented text sets. Computer Optics 2016; 40(4): 572-582. DOI: 10.18287/2412-6179-2016-40-4-572-582.

Acknowledgements: The work was partially funded by the Russian Federation Ministry of Education and Science (the basic part of the government order) and the Russian Foundation for Basic Research grant No. 16-01-00004.

References

[1] Koltsov PP, Osipov AS, Kutsaev AS, Kravchenko AA, Kotovich NV, Zakharov AV. On the quantitative performance evaluation of image analysis algorithms. Computer Optics 2015; 39(4): 542-556. DOI: 10.18287/0134-2452-2015-39-4-542-556.

[2] 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): 429-438. DOI: 10.18287/0134-2452-2015-39-3-429-438.

[3] Tsarkov SV. Automatic keyphrase extraction for vocabulary reduction in probabilistic topic models [In Russian]. Natural and Technical Sciences 2012; 6: 456-464.

[4] Shannon CE. Prediction and entropy of printed English. BSTJ 1951; 30(1): 50-64.

[5] Russian National Corpus [In Russian]. Source: ⟨http://www.ruscorpora.ru/⟩.

[6] Biemann C, Bordag S, Heyer G, Quasthoff U, Wolff C. Language-independent Methods for Compiling Monolingual Lexical Data. 5th International Conference "Computational Linguistics and Intelligent Text Processing" (CICLing 2004) 2004; 2945: 217-228.

[7] McDonald JH. G-test of goodness-of-fit. Handbook of Biological Statistics (Third ed.). Baltimore, Maryland: Sparky House Publishing; 2014: 53-58.

[8] Moskovich WA. Distributive-Statistical Method of Thesaurus Construction: The State of the Art and Perspectives [In Russian]. Moscow: The Scientific Council "Cybernetics" of the USSR Academy of Science; 1971.

[9] Tanimoto TT. An elementary mathematical theory of classification and prediction. New York: International Business Machines Corporation; 1958.

[10] 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): 1089-1106.

[11] Zagoruiko NG. Applied methods of data and knowledge analysis [In Russian]. Novosibirsk: Institute of Mathematics SD RAS; 1999.

[12] Grechnikov EA., Gusev GG, Kustarev AA, Raigorodsky AM. Detection of artificial texts [In Russian]. RCDL'2009 Proceedings 2009; 306-308.

[13] Manber U. Finding Similar Files in a Large File System. USENIX Winter 1994 Technical Conference Proceedings 1994; 1-10.

[14] Heintze N. Scalable Document Fingerprinting. Proceedings of the Second USENIX Workshop on Electronic Commerce 1996; 191-200.

[15] Brodskiy A., Kovalev R., Lebedev M., Leshchiner D., Sushin P. Yandex algorithms of contextual annotation at ROMIP 2008 [In Russian]. Russian Information Retrieval Evaluation Seminar (ROMIP) 2008; 160-169.

[16] Karp RM, Rabin MO. Efficient randomized pattern-matching algorithms. IBM Journal of Research and Development 1987; 31(2): 249-260.
[17] Knuth DE, Morris JH, Pratt VR. Fast pattern matching in strings. SIAM Journal on Computing 1977; 6(2): 323-350. DOI: 10.1137/0206024.
[18] Boyer RS, Moore JS. A fast string searching algorithm. Communications of the ACM 1977; 20(10): 762-772.
[19] Apache OpenNLP. Source: ⟨https://opennlp.apache.org/⟩.
[20] Leipzig Corpora Collection Download Page. Source: ⟨http://corpora2.informatik.uni-leipzig.de/download.html⟩.
[21] 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). Samara: IPSI RAS; 2013; 1: 30-35.

Authors’ information

Dmitry Vladimirovich Mikhaylov (b. 1974) graduated from Yaroslav-the-Wise Novgorod State University in 1997, majoring in 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. E-mail: *Dmitry.Mikhaylov@novsu.ru* .

Alexander Pavlovich Kozlov (b.1989) graduated from Yaroslav-the-Wise Novgorod State University in 2011, majoring in 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. E-mail: *caleo@yandex.ru* .

Gennady Martinovich Emel’yanov (b. 1943) graduated from the Leningrad Institute of Electrical Engineering in 1966, majoring in Mathematical and Calculating Instruments and Devices. Obtained his PhD (Kandidat Nauk) and his Doctoral (Doktor Nauk) degrees in Technical Sciences in 1971 and 1990, respectively. Now he is a Professor of the Information Technologies and Systems department at the Yaroslav-the-Wise Novgorod State 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. E-mail: *Gennady.Emelyanov@novsu.ru* .

Received April 14, 2016. The final version – July 1, 2016.