В. А. Минин, И. М. Зацман, В. А. Хаванскov, С. К. Шубников, Индикаторы тематических взаимосвязей отраслей науки и информационно-компьютерных технологий в начале XXI века, *Информ. и её примен.*, 2015, том 9, выпуск 2, 111–123

DOI: 10.14357/19922264150212

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2015. Т. 9. Вып. 2. С. 111–123

ИНДИКАТОРЫ ТЕМАТИЧЕСКИХ ВЗАИМОСВЯЗЕЙ ОТРАСЛЕЙ НАУКИ И ИНФОРМАЦИОННО-КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ В НАЧАЛЕ XXI ВЕКА*

В. А. Минин[1], И. М. Зацман[2], В. А. Хаванскov[3], С. К. Шубников[4]

Аннотация: Представлены результаты экспериментальных вычислений индикаторов тематических взаимосвязей науки и информационно-компьютерных технологий. Вычисленные значения индикаторов получены с помощью макета аналитической информационной системы (АИС), который был создан в рамках проекта Российского гуманитарного научного фонда (РГНФ) «Информационная система мониторинга и оценивания инновационно-технологического потенциала направлений фундаментальных научных исследований». Макет позволяет вычислять значения индикаторов взаимосвязей отраслей науки и направлений научных исследований (ННИ) с заданным видом технологий. В экспериментальных вычислениях в качестве исходной информации использовались официальные данные Роспатента об изобретениях по классу G06 Международной патентной классификации (МПК) (Обработка данных; Вычисления; Счет), опубликованные в 2000–2012 гг. Исходные данные для проведения расчетов не являются числовой информацией, а представляют собой полные тексты описаний изобретений на естественном языке (ЕЯ). Поэтому до начала экспериментальных вычислений индикаторов выполнялось извлечение из полных текстов изобретений информации о научных публикациях, цитируемых в описаниях изобретений, и определялось число публикаций по каждой отрасли науки и ННИ. Полученная числовая информация является исходной для вычислений значений индикаторов и дает возможность экспертам определять степень интенсивности переноса знаний из отраслей науки и ННИ в сферу технологий и оценивать их с помощью количественных индикаторов.

Ключевые слова: взаимосвязи науки и технологий; информационно-компьютерные технологии; обработка текста изобретения; регулярные выражения; рубрицирование; расчет значений индикаторов

DOI: 10.14357/19922264150212

1 Введение

Данная работа содержит описание итоговых результатов по проекту РГНФ (грант № 12-02-12019в) «Информационная система мониторинга и оценивания инновационно-технологического потенциала направлений фундаментальных научных исследований».

Методологические проблемы мониторинга, включая вопросы определения значений индикаторов тематических взаимосвязей науки и технологий, изложены в работах [1–6]. В [6] приводится описание индикаторов, предназначенных для количественного оценивания интенсивности процессов передачи знаний от науки к технологиям. В качестве степени интенсивности процессов передачи знаний берется число научных публикаций, цитируемых экспертами в отчетах о патентном поиске или авторами изобретений в их описаниях. Определение степени интенсивности являлось целью проекта. В проекте использовались описания изобретений, опубликованных Федеральной службой по интеллектуальной собственности (Роспатент) за период с 2000 по 2012 гг. и относящихся к классу G06 (Обработка данных; Вычисление; Счет) МПК.

В текстовых описаниях отобранных изобретений были найдены сделанные авторами ссылки на научные публикации, библиографические данные которых позволяют отнести их к определенным отраслям науки и ННИ. Таким образом, с помощью

*Работа выполнена в Институте проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук при частичной поддержке РГНФ (грант № 12-02-12019в).

[1]Российский фонд фундаментальных исследований, minin@rfbr.ru

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, iz_ipi@a170.ipi.ac.ru

[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, havanskov@a170.ipi.ac.ru

[4]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, sergeysh50@yandex.ru

ссылок цитирования устанавливалась связь между технологической сферой (в виде совокупности индексов МПК) и цитируемыми результатами научных исследований (для кодирования результатов использовались Государственный рубрикатор научно-технической информации (ГРНТИ) и рубрикатор Российского фонда фундаментальных исследований (РФФИ)).

Использование информационных ресурсов Роспатента было обусловлено тем, что они представлены в электронном виде и доступны для автоматизированной обработки. В соответствии со Страсбургским соглашением от 24 марта 1971 г. о МПК компетентные органы стран-участниц Союза по МПК при классифицировании патентных документов должны указывать «полные индексы МПК, присвоенные изобретению, описанному в документе» (ст. 4, п. 3). Это означает, что публикуемые в Роспатенте сведения о выданных патентах на изобретения содержат индексы МПК, которые можно использовать для описания тематики исследуемых групп технологий. Одновременно имеются полнотекстовые описания изобретений, представляющие собой неструктурированные тексты, в которых при изложении сути изобретения авторы нередко ссылаются на публикации в научных изданиях. Таким образом, библиографические ссылки на публикации в описаниях изобретений, привязанные к одной или нескольким рубрикам ННИ, можно сопоставить с индексами МПК, описывающими технологическую сферу, к которой относится изобретение, и таким образом косвенно оценить интенсивность процессов передачи знаний от науки к технологиям.

В процессе анализа изобретений объем отобранных полнотекстовых описаний может достигать нескольких сотен тысяч. Например, в работе [7] описывается процесс обработки массива из 656 695 патентов на изобретения, выданных Патентным ведомством США. Для установления взаимосвязей между индексами МПК и кодами рубрик ННИ из описаний изобретений в процессе обработки были выделены 1 147 160 ссылок на цитируемые публикации (ссылки на патенты не рассматривались). Затем из них для дальнейшей обработки были отобраны только те ссылки на журнальные статьи, для которых удалось идентифицировать название журнала и соотнести его с нормативным списком названий журналов, в котором каждому названию присвоена одна или несколько рубрик ННИ. При таких ограничениях для анализа было отобрано 106 636 ссылок, т. е. менее 10% от выделенных ссылок на цитируемые публикации.

При реализации подобной методологии для анализа отечественных описаний изобретений аналогичные зарубежные решения трудно адаптировать в силу ряда причин, подробно рассмотренных в работах [8, 9], а именно:

- отсутствие в «Административном регламенте исполнения Роспатентом приема заявок на изобретение, их рассмотрения и экспертизы» требований именно к структурированному представлению ссылок на цитируемые публикации (см. п. 10.11(12) Регламента [10]);
- отсутствие в опубликованных электронных версиях полнотекстовых описаний изобретений групп меток, выделяющих ссылки на цитируемые публикации согласно рекомендациям стандарта Всемирной организации интеллектуальной собственности (ВОИС) ST.14 [11];
- отсутствие или неполнота списка нормализованных и сокращенных названий журналов, используемых в ссылках на цитируемые публикации, в системах подготовки электронных патентных заявок.

Таким образом, при исследовании тематических взаимосвязей технологий и ННИ возникает задача анализа десятков и сотен тысяч полнотекстовых описаний изобретений и поиска в их текстах на ЕЯ ссылок на публикации с последующей их структуризацией и привязкой ссылок к рубрикам ННИ. Как следствие, возникает задача автоматизации данного процесса. При этом необходимо учитывать, что по своему содержанию само библиографическое описание является структурированным информационным объектом, состоящим из нескольких полей, который может размещаться в любом месте неструктурированного текста описания изобретения, а разные поля библиографической информации могут быть, в общем случае, на разных языках.

Целью данной работы является описание результатов завершенного проекта РГНФ и тех его задач, результаты решения которых в основном и определяют точность вычисленных значений индикаторов тематических взаимосвязей науки и технологий, а именно:

- поиск и выделение ссылок на цитируемые научные публикации в тексте изобретений на ЕЯ;
- рубрицирование выделенных ссылок по заданным классификаторам ННИ.

2 Поиск ссылок и их рубрикация

Всего для экспериментальных расчетов с использованием макета АИС [9] было отобрано 6665 изобретений, опубликованных Роспатентом

в 2000–2012 гг. и относящихся к классу G06 МПК. В полнотекстовых описаниях изобретений было выделено и верифицировано 8847 ссылок на цитируемые научные публикации. Архитектура АИС, ее описание и технология проведения расчетов подробно представлены в работах [9, 12]. Полнота и точность вычисленных значений индикаторов зависят от точности выделения ссылок в тексте описания изобретения и полноты перечня найденных ссылок (считается как доля от общего их числа, которые присутствуют в описании изобретения).

Построение шаблона поиска ссылки (в шаблонах используется язык регулярных выражений [13]; подробнее о построении шаблонов см. [14–17]) опирается, в основном, на требование «Административного регламента исполнения Роспатентом приема заявок на изобретение, их рассмотрения и экспертизы», содержащееся в п. 10.11(12) [10]. Это требование звучит следующим образом: «Библиографические данные источников информации указываются таким образом, чтобы источник информации мог быть по ним обнаружен». Как показывает анализ описаний изобретений, данное положение Административного регламента трактуется авторами достаточно широко, от классического представления ссылок на цитируемые публикации в научных статьях (опирающегося на стандарт представления библиографических данных [18]) до произвольной формы упоминания цитируемой публикации.

Например, в патенте 2144211 авторы упоминают источник следующим образом: «. . .как он [способ — прим. авт.] описан в главе 6 и главе 12 книги под названием «Адаптивная обработка сигналов» Бернарда Уидроу и Сэмьюэла Стернса, опубликованной издательством «Прентис Холл», копирайт 1985 г.».

Таким образом, создать шаблоны поиска ссылок на цитируемые публикации, которые покрывали бы все возможные варианты представления ссылок на цитируемые публикации, практически невозможно. Поэтому создаваемые шаблоны ориентированы, как правило, на поиск ссылок на публикации, приближенных к требованиям стандарта на представление библиографических данных публикаций, или нестандартных, но чаще используемых видов представления ссылок.

Ввиду того, что и сам стандарт допускает различные варианты представления библиографических данных для разных видов публикаций (книги, статьи, доклады на конференциях и пр.), возникла необходимость в разработке целой коллекции шаблонов $\{R\}$, каждый из которых нацелен на поиск заданного вида публикации в некоторой стандартной форме представления библиографических данных или нестандартной, но частотной форме. В общем виде структура этих шаблонов опирается на обобщенную структуру ссылки цитирования публикации, которая может быть представлена в следующем виде: [автор$\{S_1\}$] [название публикации] [$\{S_2\}$название источника] $\{S_3\}$атрибуты публикации.

Наличие квадратных скобок говорит о необязательности присутствия данного элемента структуры в реальной ссылке на цитируемую публикацию.

В процессе поиска ссылок на публикацию в описании изобретений к его тексту применяется коллекция $\{R\}$ шаблонов поиска ссылок. Использование классов регулярных выражений обеспечивает возможность получения непересекающихся фрагментов текста, содержащих признаки ссылки на публикацию, для отдельно взятого шаблона поиска из коллекции $\{R\}$. Но в то же время разные шаблоны коллекции могут формировать пересекающиеся фрагменты текста. Следовательно, после применения к тексту всей коллекции $\{R\}$ необходимо выполнить процедуру интеграции выделенных каждым шаблоном фрагментов текста (см. [17, рис. 1]).

Каждый шаблон из коллекции $\{R\}$ имеет видовую направленность, но заранее вид публикации неизвестен, вследствие чего при его применении возможны следующие варианты выделения текста:

(а) выделенный текст полностью соответствует ссылке на цитируемую публикацию в тексте описания изобретения;

(б) выделенный текст содержит часть ссылки на цитируемую публикацию в тексте описания;

(в) выделенный текст превышает ссылку на цитируемую публикацию в тексте описания и содержит часть содержательного текста описания;

(г) выделенный текст не содержит ссылки на цитируемую публикацию, но соответствует шаблону поиска.

Для того чтобы определить качество шаблона для поиска ссылок соответствующего вида, был сформирован тестовый массив. Каждый элемент этого массива содержит неструктурированный текст описания изобретения, внутри которого находится одна ссылка на цитируемую публикацию определенного вида. Таблица 1 включает примеры ссылок на публикации в этих текстах.

Всего в тестовом массиве представлено 277 ссылок на публикации разных видов. Для каждой ссылки на публикацию указан вид публикации (статья, книга и пр.) и тип структуры ссылки. Распределение ссылок на публикации по видам и типам структуры даны в табл. 2 и 3 соответственно.

Таблица 1 Примеры из тестового массива

Номер патента	Текст ссылки на публикацию	Вид публикации
2144274	J. B. Postel. Simple mail transfer protokol. August 1982, Information Sciences Institute, University of Southern California, RFC 821	Стандарт (RFC, ГОСТ, ОСТ и пр.)
2144785	Журнал «Science», 1998. No. 5. Vol. 280. P. 1723	Статья в журнале или сборнике
2145466	Рекомендация V.110 (1988) «Голубой книги» Международного консультативного комитета по телефонии и телеграфии (CCITT)	Стандарт (RFC, ГОСТ, ОСТ и пр.)
2146840	Баранов С. И., Скляров В. А. Цифровые устройства на программируемых БИС с матричной структурой. — М.: Радио и связь, 1986. С. 43	Книга
2148274	Горбань А. Н. Обучение нейронных сетей. М.: СП Параграф, 1990	Книга
2149450	Вопросы проектирования радиоэлектронной аппаратуры. Опыт, результаты, проблемы. — Таллин: ЭстНИИНТИ, 1989. С. 87–90. Рис. 6, *б*	Книга
2149455	Омельченко В. В. Теоретические основы классификации нечетких ситуаций при испытаниях сложных технических комплексов. — М.: МО РФ, 1998. С. 328–351	Книга
2150140	Шустер Г. Детерминированный хаос. — М.: Мир, 1988. С. 33, 38	Книга

Таблица 2 Распределение ссылок по типам публикаций

Тип публикации	Количество
Книга	100
Статья в журнале или сборнике	71
Стандарт (ИСО, ГОСТ, ОСТ и пр.)	31
Веб-публикация	29
Материалы конференции	25
Отчет	16
Статья в энциклопедии	5
Всего	277

Таблица 3 Распределение по типам структур ссылок на публикации

Тип структуры текста ссылки	Количество
Авторы(ФИО)/Название/Источник/Атрибуты	67
Название/Источник/Атрибуты	64
Авторы(ИОФ)/Название/Источник/Атрибуты	59
Произвольный текст	29
Источник	18
Источник/Атрибуты	16
Название/Авторы(ФИО)/Источник/Атрибуты	9
Название/Авторы(ИОФ)/Источник/Атрибуты	4
Источник/Название/Атрибуты	5
Авторы(ИОФ)/Источник/Атрибуты	3
Авторы(ИОФ)/Название/Атрибуты	2
Авторы(ФИО)/Название/Атрибуты	1
Всего	277

Для получения сравнительных количественных характеристик качества шаблонов поиска в рамках макета АИС была разработана отдельная подсистема, в которой реализована следующая методика анализа качества поиска ссылок на публикации: для каждого шаблона поиска ссылок создается задание на его тестирование (рис. 1).

В задании указывается имя тестируемого шаблона и формируется тестовый массив текстов изобретений на основании исходного[1], содержащих ссылки (табл. 1 содержит их примеры). В тестовом массиве помимо выделенной ссылки на публикацию в описании изобретения указывается стартовая позиция искомой ссылки на публикацию в тексте описания изобретения и длина текста ссылки в символах. Далее запускается программа поиска, которая для каждого исследуемого описания изобретения составляет список фрагментов текста, который данный шаблон поиска выделил как ссылки с указанием стартовой позиции фрагмента в тексте и его длины.

В результате формируется массив фрагментов, в котором путем сравнения стартовых позиций и длин имеющейся искомой ссылки и выделенного в процессе поиска фрагмента текста получаются количественные характеристики тестируемого шаблона поиска ссылки на публикацию, которые определяются следующим образом.

[1] Исходный тестовый массив в течение времени может пополняться новыми образцами ссылок на цитируемые публикации. Поэтому при создании тестового задания для чистоты эксперимента и последующего корректного анализа создается копия исходного массива, существующего на момент создания тестового задания.

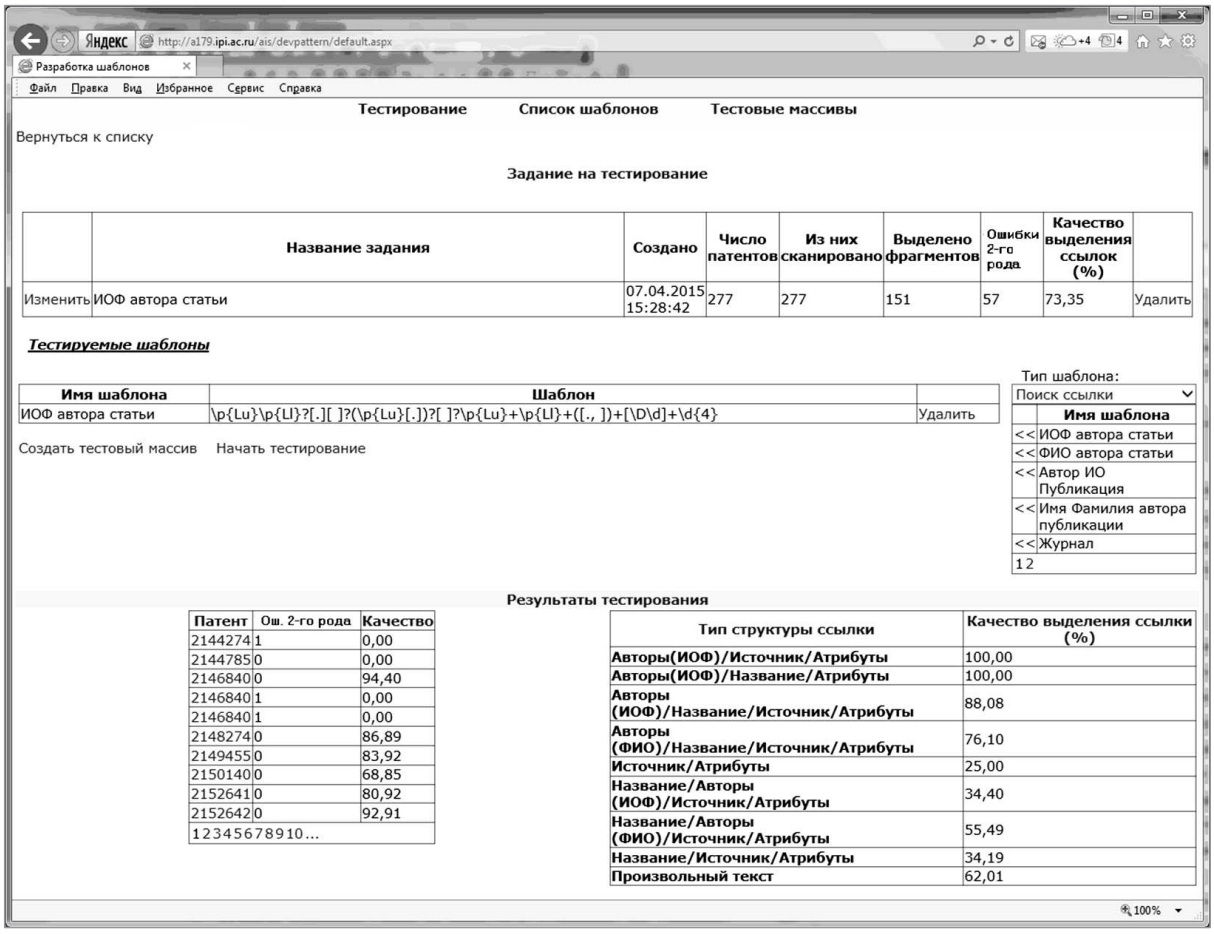

Рис. 1 Пример задания на тестирование шаблона поиска ссылок

Если выделенный фрагмент текста полностью находится за пределами искомой ссылки на публикацию, то считается, что это ложное выделение (так называемая ошибка 2-го рода). В противном случае вычисляется процент покрытия выделенным фрагментом текста искомого текста ссылки на публикацию; иными словами, определяется точность выделения ссылки цитирования.

Рисунок 1 иллюстрирует полученные результаты тестирования. Кроме того, дается список найденных фрагментов текста, который дает возможность качественной оценки работы шаблона для каждого случая его срабатывания (рис. 2).

На основе этой оценки предлагаются рекомендации по доработке (уточнению) существующего шаблона и/или созданию нового шаблона поиска. Проверяя полноту и точность каждого шаблона с использованием тестового массива, можно получить ряд его характеристик:

1. Точность выделения текста ссылки на цитируемую публикацию.
2. Полноту поиска ссылок как отношение числа найденных фрагментов текста с ненулевой точностью выделения к общему числу искомых ссылок на публикации, существующих в тестовом массиве.
3. Коэффициент ложных выделений как отношение числа найденных фрагментов текста с нулевой точностью выделения к общему числу ссылок на публикацию в тестовом массиве.

Если первые две характеристики относятся к точности и полноте поиска с помощью анализируемого шаблона, то третья характеристика представляет собой частотность ошибок второго рода, порождаемых этим шаблоном.

Возможно получение и еще одной характеристики, определяющей приоритет использования анализируемого шаблона для поиска ссылок на публикации. Для этого в тестовом массиве указаны одновременно тип структуры ссылки и вид цитируемой публикации (книга, статья, доклад и пр.). Имея сводную таблицу результатов тестирования

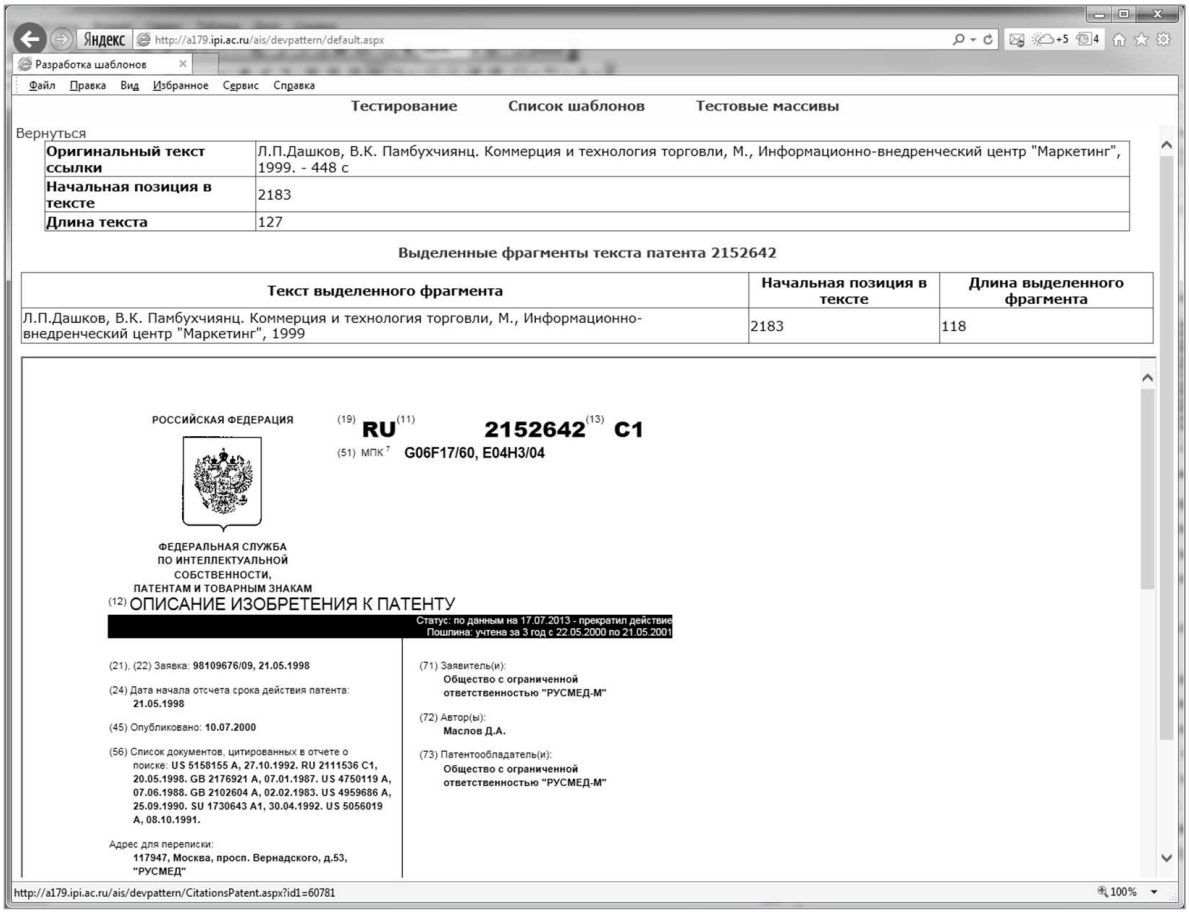

Рис. 2 Данные о выделенном фрагменте текста из описания изобретения, на которое был выдан патент 2152642

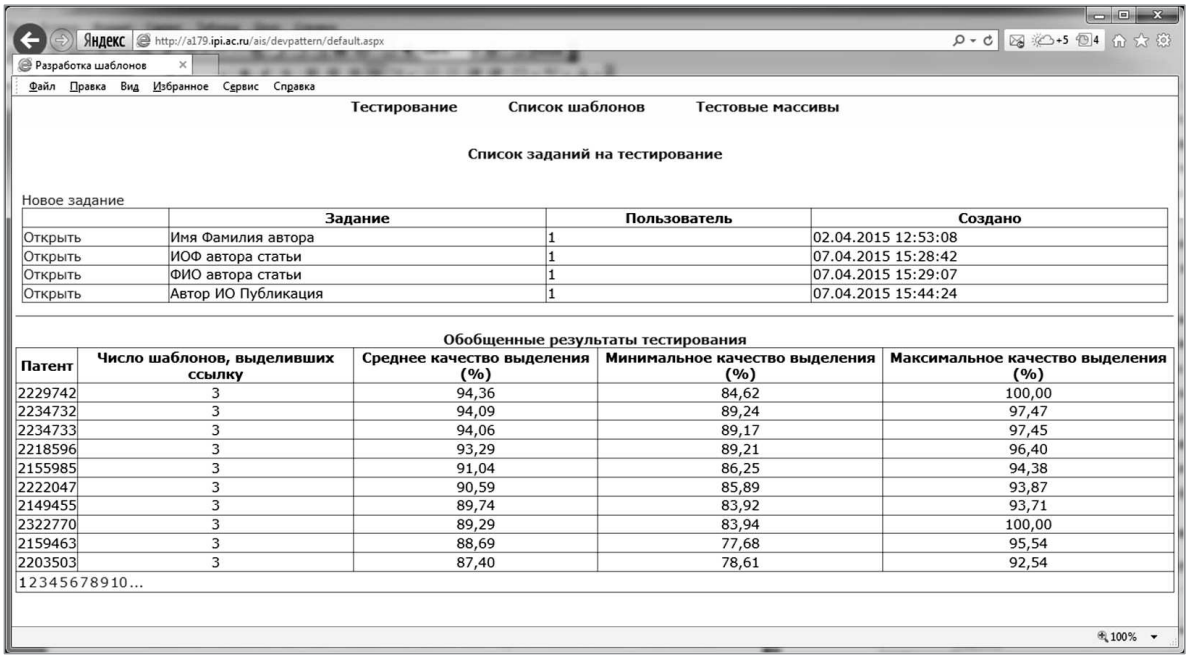

Рис. 3 Результаты выполнения тестовых заданий на тестирование шаблонов

для всех разработанных шаблонов (рис. 3), можно выбрать те шаблоны, которые дают наилучшие результаты поиска ссылок. Это позволяет сократить число используемых шаблонов поиска, выделив базовые, и, как следствие, увеличить производительность АИС на основании результатов тестирования.

Используемые в АИС методы определения ННИ, которые взаимосвязаны с заданными технологическими областями, предполагают оперирование классификаторами научной и научно-технической информации (ГРНТИ, РФФИ и пр.). Найденные и структурированные ссылки на цитируемые публикации в процессе работы АИС соотносятся с теми или иными рубриками используемых классификаторов. В процессе вычисления значений индикаторов с помощью макета АИС предполагалось, что рубрики цитируемых публикаций совпадают с рубриками изданий, в которых они были опубликованы. Это предположение, с одной стороны, существенно упростило макетирование. С другой стороны, оно снизило точность вычисленных значений индикаторов. Однако, так как основной задачей макета являлась демонстрация реализуемости методологии вычисления индикаторов взаимосвязей науки и технологий, то для этапа макетирования это предположение было допустимым.

Естественно, что в промышленном варианте АИС рубрики публикаций и изданий в общем случае совпадать не будут и должны использоваться именно рубрики публикаций.

Отметим, что существующие информационно-библиотечные системы предоставляют возможность использования именно рубрик публикаций. Например, такая возможность есть в электронных каталогах Государственной публичной научно-технической библиотеки или Всероссийского института научной и технической информации Российской академии наук [19, 20].

3 Индикаторы взаимосвязей науки и технологий

В проведенном эксперименте с использованием макета АИС вычислялись значения следующих двух индикаторов:

(1) матрица корреляций между индексами МПК (для класса G06 и его подклассов) и рубриками ННИ ГРНТИ;

(2) распределение времени отклика на статью (от момента ее публикации до момента публикации патента на изобретение, где она цитируется).

Подробное описание матрицы корреляций представлено в работе [9]. Частотности связей между индексами МПК и рубриками ГРНТИ, вычисленные для всего массива изобретений по информационно-компьютерным технологиям, запатентованных в РФ в период 2000–2012 гг., показаны в ячейках матрицы (табл. 4). Для компактности представления все рубрики ННИ представлены только самым верхним уровнем ГРНТИ, технологии — подклассами класса G06 МПК.

Таблица 5 отображает распределения цитируемых публикаций и изобретений для всего класса G06. Таблица 6 содержит названия индексов ГРНТИ.

В строках матрицы (см. табл. 4) приведены первые девять ННИ в классификации ГРНТИ в порядке убывания числа связей с индексами МПК по всему классу G06 (см. последний столбец таблицы). Эти частотности связей между индексами МПК и рубриками ГРНТИ были вычислены впервые в отечественной научно-технической сфере. Они определены в процессе обработки всех статей, цитируемых в описаниях изобретений по классу G06, независимо от того, кем цитируется статья: экспертами, что обозначается в описании изобретения меткой 56, и/или авторами изобретений.

Таблица 4 Частотности связей между индексами МПК и рубриками ГРНТИ, %

Код рубрики ГРНТИ	Подкласс								Класс
	G06E	G06F	G06G	G06K	G06M	G06N	G06Q	G06T	G06
50.00.00	0	9,58	0,27	4,85	0	0,62	1,01	2,05	18,38
28.00.00	0	8,40	0,11	4,10	0	0,59	1,05	1,35	15,60
47.00.00	0,10	5,16	0,41	3,86	0	0,17	0	0,50	10,20
45.00.00	0	4,88	0,26	3,49	0	0,19	0	0,35	9,17
20.00.00	0	4,41	0,02	3,25	0	0,12	0	0,05	7,85
30.00.00	0	4,19	0,00	3,23	0	0,11	0	0,04	7,57
29.00.00	0	3,79	0,02	3,17	0	0,01	0	0	6,99
84.00.00	0	3,50	0	3,11	0	0	0,01	0	6,62
27.00.00	0	2,08	0	0,71	0	0,52	1,01	1,30	5,62
Остальные рубрики	0,00	8,69	0,23	1,39	0,01	1,16	0,15	0,37	12,00

Таблица 5 Распределение изобретений и статей по подклассам класса G06

Индекс МПК	Название подкласса	Число изобретений	Число статей	Число статей на одно изобретение
G06C	Механические цифровые вычислительные машины	7	0	0
G06D	Гидравлические и пневматические цифровые вычислительные устройства	1	0	0
G06E	Оптические вычислительные устройства	52	8	0,15
G06F	Обработка цифровых данных с помощью электрических устройств	3415	107	0,03
G06G	Аналоговые вычислительные машины. . .	228	14	0,06
G06J	Гибридные вычислительные устройства	3	0	0
G06K	Распознавание, представление и воспроизведение данных; манипулирование носителями информации; носители информации	681	64	0,09
G06M	Счетчики; способы и устройства для подсчета предметов, не отнесенные к другим подклассам	12	0	0
G06N	Компьютерные системы, основанные на специфических вычислительных моделях	107	8	0,07
G06Q	Системы обработки данных или способы, специально предназначенные для административных, коммерческих $\langle \ldots \rangle$ целей	417	5	0,01
G06T	Обработка или генерация данных изображения. . .	320	43	0,13
Всего по классу G06		5243	249	0,05

Таблица 6 Названия рубрик ГРНТИ из матрицы корреляций

Код ГРНТИ	Название рубрики
20.00.00	Информатика
27.00.00	Математика
28.00.00	Кибернетика
29.00.00	Физика
30.00.00	Механика
45.00.00	Электротехника
47.00.00	Электроника. Радиотехника
50.00.00	Автоматика. Вычислительная техника
84.00.00	Стандартизация

В рамках проведенного эксперимента вычислялись значения еще одного индикатора — распределения времени отклика на статьи (рис. 4). В процессе вычисления этого индикатора для каждой пары «индекс МПК – рубрика ГРНТИ» было определено время отклика как разность между годом публикации патента на изобретение и годом публикации статьи, на которую есть ссылка в описании этого изобретения. Отдельно отмечались статьи, цитируемые экспертами в отчетах о патентном поиске. Затем было построено распределение времени отклика с учетом авторства ссылок на статьи (эксперты включили ссылку в отчет о патентном поиске или авторы изобретения в его полное описание).

Экспериментальные данные позволяют утверждать, что в патентах на изобретения, опубликованные в период 2000–2012 гг., эксперты в отчетах о поиске и авторы изобретений по информационным технологиям наиболее часто цитировали статьи, опубликованные за 10, 20 и 30 лет до публикации патентов на эти изобретения.

4 Заключение

Разработанный макет АИС и технология его применения впервые в отечественной научно-технической сфере дали возможность выявить количественные взаимосвязи отраслей науки и ННИ с заданным видом технологий. С помощью макета были вычислены значения индикатора тематических взаимосвязей информационных технологий, относящихся к классу G06 МПК, с рубриками ГРНТИ. Вычисленные значения показывают, что наиболее часто в изобретениях по информационно-компьютерным технологиям цитируются статьи по автоматике, вычислительной технике, кибернетике, электронике, радиотехнике, электротехнике

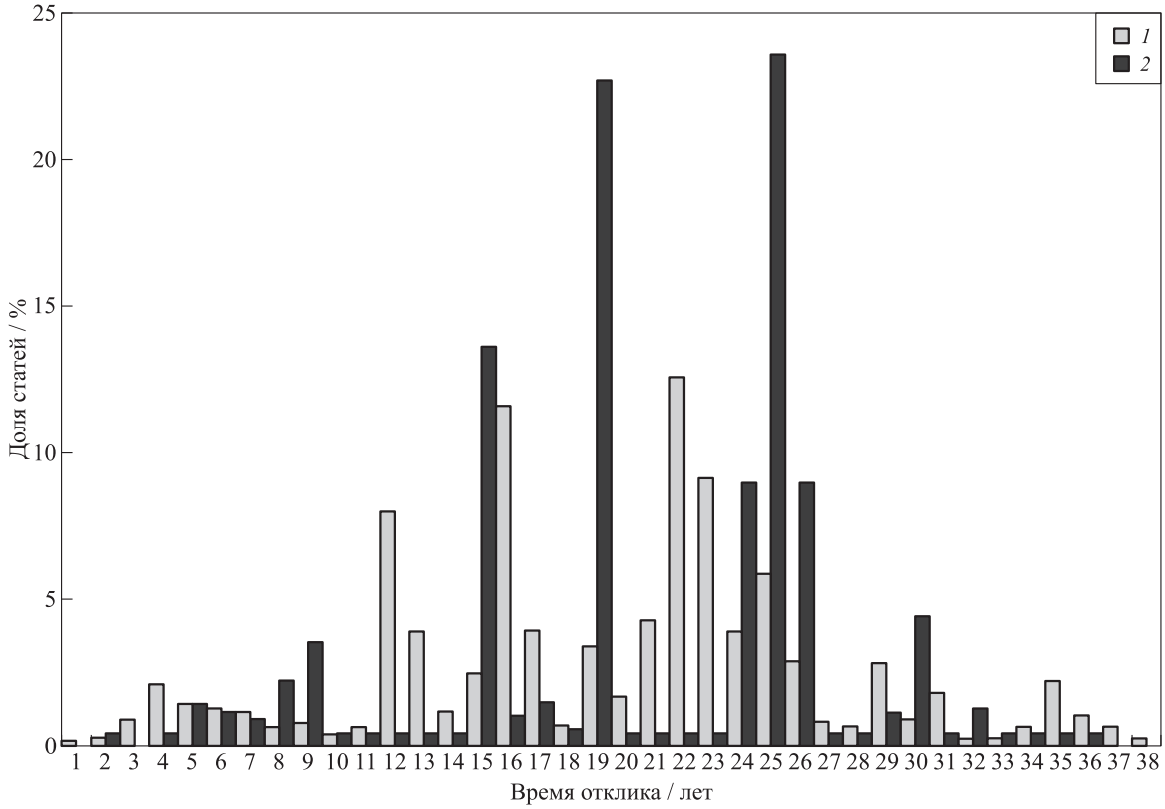

Рис. 4 Распределение времени между публикацией статьи и патента по классу G06: *1* — по полным описаниям изобретений; *2* — по спискам документов, цитируемых в отчетах об информационном поиске

и информатике (см. табл. 4). Таким образом, использование рубрик ГРНТИ дает в первую очередь прикладной разрез тематических взаимосвязей. Для получения более полной картины взаимосвязей отраслей науки с технологиями необходимо также использовать рубрики фундаментальных наук, например классификатор РФФИ.

Экспериментальные расчеты, проведенные с помощью макета АИС, позволяют сделать вывод о реализуемости методологии определения взаимосвязей отраслей науки и технологий с использованием отечественных патентных информационных ресурсов. В процессе проведения эксперимента была проведена оценка качества шаблонов поиска, используемых функциональными подсистемами макета АИС. Были получены численные оценки точности и полноты поиска ссылок на цитируемые публикации в полнотекстовых описаниях изобретений.

Выявлены наиболее сложные технологические операции поиска ссылок, рубрицирования публикаций в автоматическом режиме и определены подходы к повышению точности и полноты выделения ссылок и рубрицирования. В макете АИС предусмотрена возможность для экспертного уточнения результатов автоматической рубрикации, выполняемой сейчас с использованием рубрик изданий, а не публикаций. Это позволяет уже сегодня использовать макет АИС для вычисления значений индикаторов взаимосвязей отраслей науки и ННИ с любым заданным видом технологий на ретроспективе в 12–15 лет и использовать ретроспективные данные для прогноза изменения значений этих индикаторов в краткосрочной перспективе.

Литература

1. *Schmoch U.* Tracing the knowledge transfer from science to technology as reflected in patent indicators // Scientometrics, 1993. Vol. 26. No. 1. P. 193–211.
2. *Зацман И. М., Веревкин Г. Ф.* Информационный мониторинг сферы науки в задачах программно-целевого управления // Системы и средства информатики, 2006. Вып. 16. С. 164–189.
3. *Зацман И. М., Кожунова О. С.* Семантический словарь системы информационного мониторинга в сфере науки: задачи и функции // Системы и средства информатики, 2007. Вып. 17. С. 124–141.
4. *Архипова М. Ю., Зацман И М., Шульга С. Ю.* Индикаторы патентной активности в сфере информаци-

онно-коммуникационных технологий и методика их вычисления // Экономика, статистика и информатика. Вестник УМО, 2010. № 4. С. 93–104.

5. *Зацман И. М., Дурново А. А.* Моделирование процессов формирования экспертных знаний для мониторинга программно-целевой деятельности // Информатика и её применения, 2011. Т. 5. Вып. 4. С. 84–98.
6. *Минин В. А., Зацман И. М., Кружков М. Г., Норекян Т. П.* Методологические основы создания информационных систем для вычисления индикаторов тематических взаимосвязей науки и технологий // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 70–81.
7. *Verbeek A., Debackere K., Luwel M., Andries P., Zimmermann E., Deleus D.* Linking science to technology: Using bibliographic references in patents to build linkage schemes // Scientometrics, 2002. Vol. 54. No. 3. P. 399–420.
8. *Зацман И. М., Шубников С. К.* Принципы обработки информационных ресурсов для оценки инновационного потенциала направлений научных исследований // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. IX Всеросс. науч. конф. RCDL'2007. — Переславль: Ун-т города Переславля, 2007. С. 35–44.
9. *Минин В. А., Зацман И. М., Хаванскоv В. А., Шубников С. К.* Индикаторы тематических взаимосвязей науки и технологий: от текста к числам // Информатика и её применения, 2014. Т. 8. Вып. 3. С. 114–125.
10. Административный регламент исполнения Роспатентом приема заявок на изобретение, их рассмотрения и экспертизы. — М.: ФИПС, 2008. http://www1.fips.ru/wps/wcm/connect/content_ru/ru/documents/russian_laws/order_minobr/administrative_regulations/test_8.
11. Стандарт ВОИС ST.14. Рекомендации по включению ссылок, цитируемых в патентных документах. http://www.rupto.ru/rupto/nfile/52b8dfc1-1049-11e1-a520-9c8e9921fb2c/03_14_01.pdf.
12. *Минин В. А., Зацман И. М., Хаванскоv В. А., Шубников С. К.* Архитектурные решения для систем вычисления индикаторов тематических взаимосвязей науки и технологий // Системы и средства информатики, 2013. Т. 23. № 2. С. 260–283.
13. Регулярные выражения в .NET Framework // MSDN. Библиотека. http://msdn.microsoft.com/ru-ru/library/hs600312.aspx.
14. *Васильев А., Козлов Д., Самусев С., Шамина О.* Извлечение метаинформации и библиографических ссылок из текстов русскоязычных научных статей // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. IX Всеросс. науч. конф. RCDL'2007. — Переславль: Ун-т города Переславля, 2007. С. 175–184.
15. *Васильев А., Козлов Д., Самусев С., Шамина О.* Создание электронной библиотеки русскоязычных научных статей // Сб. работ стипендиатов гранта «Интернет-математика 2007». — Екатеринбург: Уральский ун-т, 2007. С. 37–45.
16. *Зацман И. М., Хаванскоv В. А., Шубников С. К.* Метод извлечения библиографической информации из полнотекстовых описаний изобретений // Информатика и её применения, 2013. Т. 7. Вып. 4. С. 52–65.
17. *Хаванскоv В. А., Шубников С. К.* Поиск и рубрицирование ссылок на цитируемые публикации в электронных библиотеках полнотекстовых описаний изобретений // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. XVI Всеросс. науч. конф. RCDL-2014. — Дубна: ОИЯИ, 2014. С. 165–173.
18. ГОСТ 7.1-2003. Библиографическая запись. Библиографическое описание. Общие требования и правила составления. http://lib.usfeu.ru/index.php/gost-7-1-2003.
19. *Сбойчаков К. О.* Распределение ключевых слов по рубрикам ГРНТИ в базе данных Электронного каталога ГПНТБ России // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: Тр. XI Междунар. конф. «Крым 2004». — М., 2004. http://www.gpntb.ru/win/inter-events/crimea2004/292.pdf.
20. *Гиляревский Р. С., Шапкин А. В., Белоозеров В. Н.* Рубрикатор как инструмент информационной навигации. — С.-Петербург: Профессия, 2008. 352 с.

Поступила в редакцию 21.04.15

INDICATORS FOR THEMATIC LINKAGES BETWEEN SCIENCE AND INFORMATION AND COMPUTER TECHNOLOGIES AT THE BEGINNING OF THE XXI CENTURY

V. A. Minin[1], I. M. Zatsman[2], V. A. Havanskov[2], and S. K. Shubnikov[2]

[1]Russian Foundation for Basic Research, 32A Leninsky Prosp., Moscow 119991, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Outcomes of experimental evaluation of thematic linkages between science and information and computer technologies (ICT) are presented. The indicator values for the linkages are calculated by the testbed of an analytical information system that was created within the project of the Russian Foundation for Humanities "Information system for monitoring and evaluating innovative and technological potential of the fields of basic research." Texts of inventions on the class G06 (Data processing; Calculations; Account) of the International patent classification were used. These texts, which were published in 2000–2012 by Rospatent, are full-text descriptions of inventions in a natural language. Prior to experimental calculation of indicator values for the linkages, automated extraction of information on the cited scientific publications was retrieved from full-text descriptions. A number of publications was determined for each field of basic research. Obtained numerical information was used for quantitative evaluation of thematic science–ICT linkages and gave the possibility to define an intensity of knowledge transfer from science to ICT sphere and estimate the linkages by quantitative indicators.

Keywords: science–technology linkages; information and communication technologies; processing of invention text; regular expressions; classifying; evaluation of indicator values

DOI: 10.14357/19922264150212

Acknowledgments

The research was performed at the Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences and was partially supported by the Russian Foundation of Humanities (grant No. 12-02-12019в).

References

1. Schmoch, U. 1993. Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics* 26(1):193–211.
2. Zatsman, I. M., and G. F. Verevkin. 2006. Informatsionnyy monitoring sfery nauki v zadachakh programmno-tselevogo upravleniya [Information monitoring in sphere of science and problems of program-oriented management]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 16:164–189.
3. Zatsman, I. M., and O. S. Kozhunova. 2007. Semanticheskiy slovar' sistemy informatsionnogo monitoringa v sfere nauki: Zadachi i funktsii [The semantic dictionary of system for information monitoring in science sphere: Tasks and functions]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 17:124–141.
4. Arhipova, M. Yu., I. M. Zatsman, and S. Yu. Shul'ga. 2010. Indikatory patentnoy aktivnosti v sfere informatsionno-kommunikatsionnykh tekhnologiy i metodika ikh vychisleniya [Indicators of patent activity in the sphere of information and communication technologies and technique of their calculation]. *Ekonomika, Statistika i Informatika. Vestnik UMO* [Economy, Statistics, and Informatics. Herald of the UMO] 4:93–104.
5. Zatsman, I. M., and A. A. Durnovo. 2011. Modelirovanie protsessov formirovaniya ekspertnykh znaniy dlya monitoringa programmno-tselevoy deyatel'nosti [Modeling of creation processes of expert knowledge for monitoring program-oriented activities]. *Informatika i ee Primeneniya — Inform. Appl.*] 5(4):84–98.
6. Minin, V. A., I. M. Zatsman, M. G. Kruzhkov, and T. P. Norekyan. 2013. Metodologicheskie osnovy sozdaniya informatsionnykh sistem dlya vychisleniya indikatorov tematicheskikh vzaimosvyazey nauki i tekhnologiy [Methodological basis for the creation of information systems for the calculation of indicators of thematic linkages between science and technology]. *Informatika i ee Primeneniya — Inform. Appl.* 7(1):70–81.
7. Verbeek, A., K. Debackere, M. Luwel, P. Andries, E. Zimmermann, and D. Deleus. 2002. Linking science to technology: Using bibliographic references in patents to build linkage schemes. *Scientometrics* 54(3):399–420.
8. Zatsman, I. M., and S. K. Shubnikov. 2007. Printsipy obrabotki informatsionnykh resursov dlya otsenki innova-

tsionnogo potentsiala napravleniy nauchnykh issledovaniy [Principles of processing of information resources for assessment of innovative potential of fields of scientific research]. *Elektronnye biblioteki: Perspektivnye metody i tekhnologii, elektronnye kollektsii: Tr. 9-y Vseross. nauch. konf. RCDL'2007* [Digital Libraries: Perspective Methods and Technologies, Electronic Collections: 9th All-Russia Scientific Conference RCDL'2007 Proceedings]. Pereslavl': Pereslavl University. 35–44.

9. Minin, V. A., I. M. Zatsman, V. A. Havanskov, and S. K. Shubnikov. 2014. Indikatory tematicheskikh vzaimosvyazey nauki i tekhnologiy: Ot teksta k chislam [Indicators of thematic science–technology linkages: From text to numbers]. *Informatika i ee Primeneniya — Inform. Appl.* 8(3):114–125.
10. Administrativnyy reglament ispolneniya Rospatentom priema zayavok na izobretenie, ikh rassmotreniya i ekspertizy [Administrative regulations of execution by Rospatent of demands acceptance for the invention, their considerations and examination]. Available at: http://www1.fips.ru/wps/wcm/connect/content_ru/ru/documents/russian_laws/order_minobr/administrative_regulations/test_8/ (accessed April 17, 2015).
11. Standart VOIS ST.14 "Rekomendatsii po vklyucheniyu ssylok, tsitiruemykh v patentnykh dokumentakh" [WIPO Standard ST.14 "Recommendations for the Inclusion of References Cited in Patent Documents"]. Available at: http://www.rupto.ru/rupto/nfile/52b8dfc1-1049-11e1-a520-9c8e9921fb2c/03_14_01.pdf (accessed April 17, 2015).
12. Minin, V. A., I. M. Zatsman, V. A. Havanskov, and S. K. Shubnikov. 2013. Arkhitekturnye resheniya dlya sistem vychisleniya indikatorov tematicheskikh vzaimosvyazey nauki i tekhnologiy [Information system conceptual decisions for assessment of linkages between science and technologies]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 23(2):260–283.
13. Regulyarnye vyrazheniya v .NET Framework [Regular expressions within .NET Framework]. Available at: http://msdn.microsoft.com/ru-ru/library/hs600312.aspx (accessed April 17, 2015).
14. Vasil'ev, A., D. Kozlov, S. Samusev, and O. Shamina. 2007. Izvlechenie metainformatsii i bibliograficheskikh ssylok iz tekstov russkoyazychnykh nauchnykh statey [Extraction of metainformation and bibliographic references from texts of Russian language scientific articles]. *Elektronnye biblioteki: Perspektivnye metody i tekhnologii, elektronnye kollektsii: Tr. 9-y Vseross. nauch. konf. RCDL'2007* [Digital Libraries: Perspective Methods and Technologies, Electronic Collections: 9th All-Russia Scientific Conference RCDL'2007 Proceedings]. Pereslavl': Pereslavl University. 175–184.
15. Vasil'ev, A., D. Kozlov, S. Samusev, and O. Shamina. 2007. Sozdanie elektronnoy biblioteki russkoyazychnykh nauchnykh statey [Creation of digital library of Russian language scientific articles]. *Sb. rabot stipendiatov granta "Internet-matematika 2007"* [Collection of works of scholars of a grant "Internet mathematics 2007"]. Ekaterinburg: Ural University. 37–45.
16. Zatsman, I. M., V. A. Havanskov, and S. K. Shubnikov. 2013. Metod izvlecheniya bibliograficheskoy informatsii iz polnotekstovykh opisaniy izobreteniy [Method of bibliographic information extraction from full-text descriptions of inventions]. *Informatika i ee Primeneniya — Inform. Appl.* 7(4):52–65.
17. Havanskov, V. A., and S. K. Shubnikov. 2013. Poisk i rubritsirovanie ssylok na tsitiruemye publikatsii v elektronnykh bibliotekakh polnotekstovykh opisaniy izobreteniy [Search and classifying of cited publications in digital libraries of full-text descriptions of inventions]. *Elektronnye biblioteki: Perspektivnye metody i tekhnologii, elektronnye kollektsii: Tr. 16-y Vseross. nauch. konf. RCDL'2014* [Digital Libraries: Perspective Methods and Technologies, Electronic Collections: 16th All-Russia Scientific Conference RCDL'2014 Proceedings]. Dubna: Joint Institute for Nuclear Research. 165–173.
18. GOST 7.1-2003. Bibliograficheskaya zapis'. Bibliograficheskoe opisanie. Obshchie trebovaniya i pravila sostavleniya [Bibliographic record. Bibliographic description. General requirements and drawing-up rules]. Available at: http://lib.usfeu.ru/index.php/gost-7-1-2003 (accessed April 17, 2015).
19. Sboychakov, K. O. 2004. Raspredelenie klyuchevykh slov po rubrikam GRNTI v baze dannykh Elektronnogo kataloga GPNTB Rossii [Distribution of keywords on SCSTI headings in a database of the Electronic catalog of State Public Scientific Technical Library of Russia]. *Biblioteki i informatsionnye resursy v sovremennom mire nauki, kul'tury, obrazovaniya i biznesa: 11-ya Mezhdunar. konf. "Krym 2004"* [Libraries and Information Resources in the Modern World of Science, Culture, Education, and Business: 11th Conference (International) "Crimea 2004"]. Moscow. Available at: http://www.gpntb.ru/win/inter-events/crimea2004/292.pdf (accessed May 27, 2015).
20. Gilyarevskiy, R. S., A. V. Shapkin, and V. N. Beloozerov. 2008. *Rubrikator kak instrument informatsionnoy navigatsii* [Subject authority as instrument of information navigation]. St. Petersburg: Professiya. 352 p.

Received April 21, 2015

Contributors

Minin Vladimir A. (b. 1941) — Doctor of Science in physics and mathematics, adviser, Russian Foundation for Basic Research, 32A Leninsky Prosp., Moscow 119991, Russian Federation; minin@rfbr.ru

Zatsman Igor M. (b. 1952) — Doctor of Sciences in technology, Head of Department, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; iz_ipi@a170.ipi.ac.ru

Havanskov Valerij A. (b. 1950) — scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; havanskov@a170.ipi.ac.ru

Shubnikov Sergej K. (b. 1955) — senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sergeysh50@yandex.ru